The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

Discovering Suspicious Patterns Using a Graph Based Approach

Sirisha Velampalli
Assistant Professor
CR Rao AIMSCS
Prof. CR.Rao Road, Hyderabad
sirisha.crraoaimscs@gmail.com

Lenin Mookiah
Tennessee Technological University
Cookeville, TN USA
lenin.world@gmail.com

William Eberle
Professor
Tennessee Technological University
Cookeville, TN USA
weberle@tntech.edu

Abstract

Recently, there has been much attention on tools and techniques for visualizing and acquiring new knowledge and insights. In the VAST 2018 competition, one of the challenges is to discover the fraudulent group of employees at Kasios, a furniture manufacturing company. In this work, we use a graph-based approach that analyzes the data for suspicious employee activities at Kasios. Graph based approaches enable one to handle rich contextual data and provide a deeper understanding of data due to the ability to discover patterns in databases that are not easily found using traditional query or statistical tools. We focus on graph based knowledge discovery in structural data to mine for interesting patterns and anomalies. Our approach first reports the normative patterns in the data, and then discovers any anomalous patterns associated with the previously discovered patterns. For visualizing the suspicious patterns, we also use the enterprise graph database Neo4j. Neo4j Browser provides a way to visualize graph structures.

Introduction

Data analytics techniques enables one to discover valid patterns and inferences, leading to more informed decisions, conclusions, and correlations in the data. They can be used by scientists and researchers to verify or disprove scientific models, theories and hypotheses. Many analytical tasks can be performed efficiently using graphs. Graph based approaches can mine patterns from diverse domains such as chemical data analysis, communication networks, traffic networks, and protein interaction networks. Many traditional data mining algorithms such as classification, clustering, association analysis, and outlier detection have been extended into graph mining.

For the purposes of validating graph-based approaches for performing data analytics, we experiment with employees datasets provided by the Visual Analytics Science and Technology (VAST) community, using the goals of one of the challenges to discover fraudelent group of employees.

The Visual Analytics Science and Technology (VAST) 2018 challenge (cit), deals with unusual activity at a wildlife preserve. Boonsong Lekagul, a large nature preserve, is located southwest of Mistford city. It is identified that in this preserve, there are a number of nesting pairs of Rose-Crested Blue Pipit, a popular local, singing bird with attractive plumage. In this challenge, contestants are asked to identify the possible reasons for the decrease in the number of nesting pairs of those birds in the nature preserve. The VAST 2018 competition poses 3 mini-challenges. In this work, we focussed on Mini-Challenge 3. In the case of Mini-Challenge 3, we are tasked with (1) identifying employees who make suspicious purchases, according to data from a company insider, (2) characterizing the group's organizational structure, and visualize the communication patterns within the group, and (3) discovering any other instances of suspicious activities in the company using the structure of the group provided by the insider.

One approach to solving this problem is to structure each individual employee into separate groups (i.e., graphs) based on the kind of activity they perform, such as calls, emails, purchases, and meetings. We decided to use a graph-based approach because graphs provide a meaningful structural representation to data that can be used for analyzing or discovering interesting patterns in complex relationships.

Related Work

There has been a wide variety of data analytical graph-based approaches. Mookiah et al. use a graph based anomaly detection approach to uncover the mystery of people who have disappeared using various datasets, including GPS in vehicles, credit card transactions, loyalty card usage, and a map (Mookiah, Eberle, and Holder 2014). Vaqar et al. (Vaqar and Basir 2009) propose a weight-of-evidence-based classification algorithm to detect different road traffic conditions. They analyze traffic patterns by representing each vehicle as individual nodes in a graph. Velampalli et al. use the GBAD approach to detect anomalous patterns by adding background knowledge to evaluation metrics (Velampalli and Eberle). Background knowledge is added in the form of rule coverage which reports the percentage of the final graph covered by the instances of the substructure. Velampalli et al. also use conceptual graphs and a MapReduce programming model to extract skill-sets from a dataset of resumes (Velampalli

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

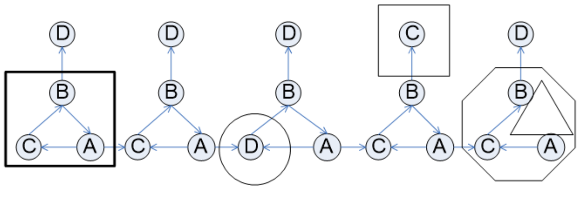

Figure 1: Example Graph Showing Different Types of Anomalies

and Eberle 2016). Michalak (Michalak and Korczak 2011) propose a machine learning method for mining transaction graphs for detecting suspicious banking transactions. Their method involves the use of fuzzy numbers to represent parameters of transactions involved in a money laundering event. Jedrzejek et al. use a graph based approach to detect a large class of financial crimes. Their proposed model uses a minimal model of descriptions and a sufficient ontology to detect criminal activities (Jedrzejek, Bak, and Falkowski 2009). Depren et al. (Depren et al. 2005) proposed intrusion detection system for anomaly and misuse detection. They used self-organizing map (SOM) structure to model normal behaviour. They defined attack as some deviation from the normal behavior. They used J.48 decision tree algorithms for misuse detection module to classify various types of attacks.

Graph-Based Anomaly Detection (GBAD)

In this work, we use the Graph Based Anomaly Detection (Eberle and Holder 2007) system to discover both normative and anomalous patterns. The GBAD system can detect three structural anomalies: modifications, insertions, and deletions. Figure 1 demonstrates each of the different types of structural changes. GBAD implements three algorithms GBAD-MDL, GBAD-P and GBAD-MPS to detect these three types of anomalies. Through these algorithms, GBAD first discovers normative patterns based on the SUBDUE (Ketkar, Holder, and Cook 2005) graph-based knowledge discovery system. Each of the three algorithms use a Minimum Description Length (MDL) (Rissanen 1985) heuristic to discover normative pattern, where the normative, or best, pattern is the one that minimizes the following objective function:

$$M(S,G) = DL(G|S) + DL(S) \tag{1}$$

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. The *GBAD-MDL* algorithm first finds the best substructure in the graph using the MDL approach and subsequently searches for all substructures that are similar to the best substructure within some threshold. *GBAD-P* algorithm also use MDL to find best substructure in the graph first, but instead of examining similar substructures, it searches for all extensions to the normative substructure, extracting substructures with extensions that have lower probability. Finally, the *GBAD-MPS* algorithm also use the MDL approach to discover the best substructure and subsequently examines all of the substructures that are missing edges and vertices, again within some threshold of change.

For more information on the GBAD system, the reader should refer to (Eberle and Holder 2007) .

Visualizing Graphs-Neo4J

We developed *GBAD2Neo4J* [1] to convert GBAD graph output files to Neo4J cypher, and then subsequently insert them into a Neo4j graph database. Neo4j is one of the popular Graph Databases and Cypher Query Language (CQL). It is highly scalable and schema free (NoSQL). Neo4j Graph Database follows the Property Graph Model to store and manage its data. Following are the key features of Property Graph Model:

- The model represents data in Nodes, Relationships and Properties.
- Properties are key-value pairs.
- Nodes are represented using circle and Relationships are represented using arrow keys.
- Relationships have directions: Unidirectional and Bidirectional.
- Each Relationship contains Start Node and To Node or End Node.
- Both Nodes and Relationships contain properties.
- Relationships connects nodes.

Using the Neo4j browser, we can visualize and explore the graph structure for relevant insights. Figure 4 and Figure 5 are examples of some of the graph structures that were discovered.

Data Set

The insider at Kasios has provided various company data (because they want to protect the birds), including call records, emails, purchases, and meetings. The data includes the source of each transaction, the recipient (destination), and the time of the transaction. The data are provided in comma-separated format with four columns:

- Source (contains the company ID# for the person who called, sent an email, purchased something, or invited people to a meeting)
- Etype (contains a number designating what kind of connection is made)
  a) 0 is for calls b) 1 is for emails c) 2 is for purchases d) 3 is for meetings
- Destination (contains company ID# for the person who is receiving a call, receiving an email, selling something to a buyer, or being invited to a meeting).
- Time stamp – in seconds starting on May 11, 2015 at 14:00.

There are four data files that cover the whole company: calls.csv has information on 10.6 million calls (251 MB uncompressed) emails.csv has information on 14.6 million

[1] https://github.com/leninworld/GBAD2Neo4J

Table 1: Data Snippet

Src_ID	Dest_ID	Time_Stamp
1066153	2015580	1095
1122557	2015580	1095
1066943	1891741	1065348
1106679	1761893	1065348

Table 2: Sample File Instance G(V, E)

```
XP # 1
v 1 ”1690582”
v 2 ”1847246”
e 1 2 ”Start meeting with”
v 3 ”2017-09-02”
e 2 3 ”Date”
v 4 ”September”
e 3 4 ”Month”
v 5 ”Saturday”
e 3 5 ”Day”
v 6 ”13:18:53”
e 2 6 ”time”
v 7 ”afternoon”
```

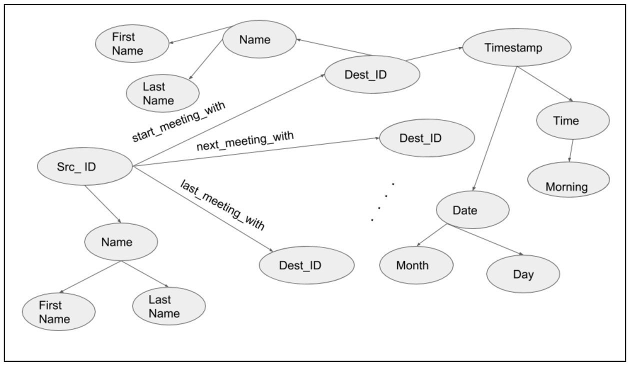

Figure 2: Proposed Graph Topology

Table 3: Suspicious Employees Provided by Insider

Alex Hall	May Burton	Augusta Sharp
Lizbeth Jindra	Glen Grant	Kerstin Belveal
Patrick Lane	Dylan Ballard	Rosalia Larroque
Richard Fox	Meryl Pastuch	Lindsy Henion
Sara Ballard	Tobi Gatlin	Julie Tierno
Jose Ringwald	Melita Scarpaci	Refugio Orrantia
	Ramiro Gault	Jenice Savaria

emails (345 MB uncompressed) purchases.csv has information on 762 thousand purchases (18.8 MB uncompressed) meetings.csv has information on 127 thousand meetings (3.26 MB uncompressed)

There are four data files that contain information about individuals that the Insider has indicated as suspicious: Suspicious_calls.csv (1.76 KB uncompressed) Suspicious_emails.csv (1.55 KB uncompressed) Suspicious_purchases.csv (27 B uncompressed) Suspicious_meetings.csv (130 B uncompressed)

An example segment of the data is shown in Table 1. We created a script using python that pre-processes and transforms all the VAST data files for Mini-Challenge 3 into a graph input file for the GBAD tool.

Graph Topology

We experimented with several candidate graph topologies, but found that the one shown in Figure 2 provides the best results. Each employee in the company is assigned a unique ID, and we structured each individual employee into separate graphs based on the kind of activity they performed, i.e., calls, emails, purchases, meetings. We then used GBAD to discover normative as well as suspicious patterns. One instance of the input file obtained after pre-processing the data is shown in Table 2.

Ground Truth

The insider, a fellow pipit (bird) lover, has provided a list of transactions by potentially suspicious employees that could indicate illicit company activity. Based on the information provided by the insider, we are able to discover employees that are involved in suspicious transactions. In Table 3, we can see the list of suspicious employees names provided by insider.

Experiments

Using the graph topology shown in Figure 2 as input to GBAD, we are able to discover several suspicious employees.

Hardware specifications for all our experiments are as follows:

- Processor Intel(R) Core(TM) i3-5005U CPU @2.00GHz 2.00 GHz, 2 Core(s), 4 Logical Processor(s).
- RAM 4.00GB.
- Operating system: xubuntu 16.04.

Run-times needed to discover patterns using GBAD are shown in Figure 3. Run-time needed to discover patterns using GBAD-MPS is comparatively less than GBAD-MDL and GBAD-P.

Results and Discussion

The GBAD algorithms discover anomalous substructures in the graph-representation of the provided data. Insider has provided some employees as suspicious as shown in Table 3. GBAD is able to detect instances of those employees. These patterns are then visually displayed using Neo4j. One such pattern is shown in Figure 4, where we can see a group of employees at various meetings - Meryl Pastuch, Sherrell Biebel, and Rosalia Larroque - are marked by GBAD as anomalous (i.e., suspicious). In Figure 5, one can note another group of employees - Alex Hall, Glen Grant, and Julie Tierno - that are involved in suspicious calls.

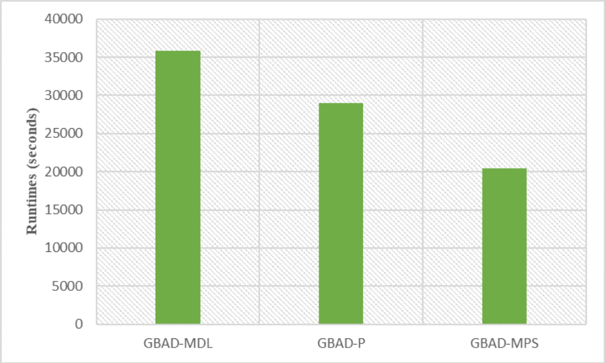

Figure 3: Runtimes Needed to Discover Patterns using GBAD

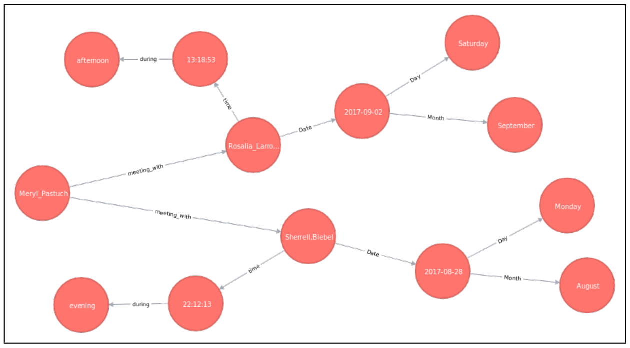

Figure 4: Suspicious Meeting Pattern: Employees Meryl Pastuch, Sherrell Biebel, Rosalia Larroque are Involved

Table 4: Employee Sheilah Stachniw

S.No	Year	No. of Calls
1.	2015	02
2.	2016	07
3.	2017	11

Table 5: Employee Laure Pelkey

S.No	Year	No. of Purchases
1.	2015	03
2.	2016	03
3.	2017	05

Likewise in Figures 6, 7 and 8 we provide suspicious purchase patterns.

In Table 4, we can see number of suspicious calls made by employee Sheilah Stachniw. In Table 5, we can see number of purchases made by employee Laure Pelkey.

In Table 6, we can see number of meetings made by employee Giovanni Overbaugh.

In Table 7, we can see number of emails made by employee Sheilah Stachniw.

Taking into account all of the anomalous patterns detected by GBAD, we suspect the 33 people (out of 642,632) listed in Table 8 are suspicious employees.

Based on the ground truth and obtained results from GBAD, Confusion matrix is shown in Table 9. From confusion matrix, values of True Positive Rate and False Positive Rate are 100% and 65% respectively.

- True Positive Rate=TP/Actual Yes=20/20=**100%**

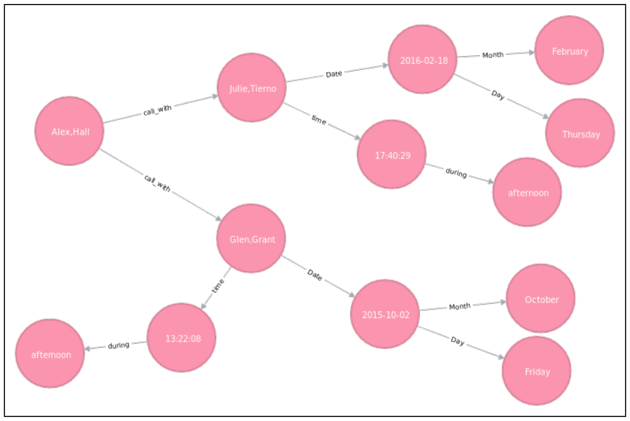

Figure 5: Suspicious Calls Pattern: Alex Hall, Glen Grant, Julie Teirno are involved

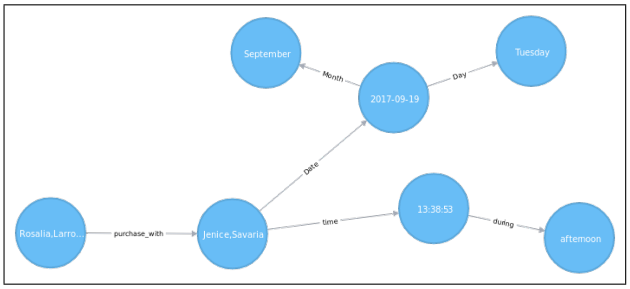

Figure 6: Suspicious Purchases: Rosalia Larroque, Jenice Savaria

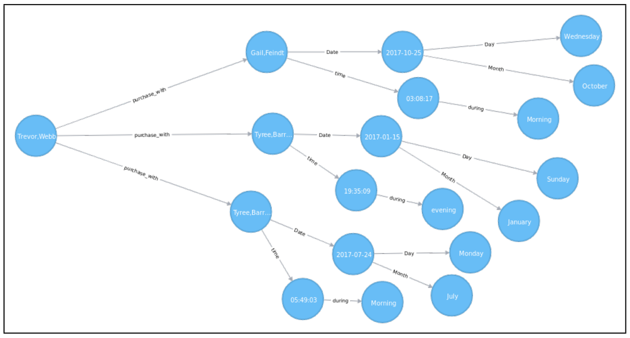

Figure 7: Suspicious Purchases:Trevor, Webb Tyree, Barreneche Gail,Feindt

Table 6: Employee Giovanni Overbaugh

S.No	Year	No. of Meetings
1.	2015	00
2.	2016	11
3.	2017	27

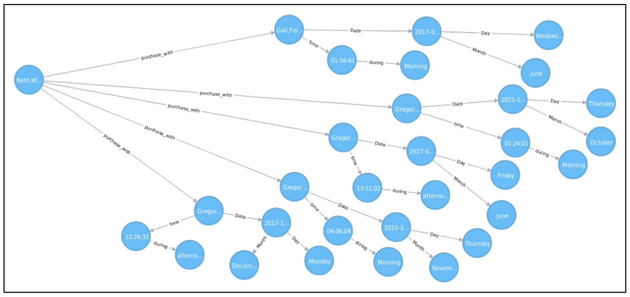

Figure 8: Suspicious Purchases- Beth Wilensky, Gregory Russell, Gail Feindt

Table 7: Employee Sheilah Stachniw

S.No	Year	No. of emails
1.	2015	04
2.	2016	11
3.	2017	11

- False Positive Rate=FP/Actual No=13/20=**65%**

Based on the above observations in the graphs, we hypothesize the following as reasons for the decrease in the number of nesting pairs of the Rose-Crested Blue Pipit:

- Production suddenly increased from 2016 to 2017, indicating that perhaps the company has restarted production of a banned solvent (like perhaps Methylosmolene) which can harm birds.
- EuroKasios is a furniture manufacturing factory. There is a gradual increase in the production of furniture from 2015 to 2017. For furniture, lots of wood is needed, which involves cutting many trees in the forest, leading to decreased nesting habitats, which can cause the birds to migrate elsewhere.
- Many suspicious meetings are held during evenings, i.e., beyond normal office hours.

Conclusion

In this work, we used a graph-based approach to discover suspicious employees and their activities that help to understand the reasons for the decrease in the number of nesting pairs of Rose-Crested Blue Pipit. Using tools like GBAD and Neo4j allowed us to discover and visualize suspicious employees and their patterns of communication. In this work, we analyzed only calls, emails, purchases and meetings datasets of employees, but in the future we want to analyze varied domains and mine interesting patterns from heterogeneous datasets. Also as part of our future work, we will investigate the inclusion of a ranking function and distance metrics to our graph based approach, whereby we can better detect most suspicious patterns. We also want to decrease the run-times by using filtering techniques, as well as add some interactive visualization techniques that can provide quick insights and analysis.

Table 8: Suspicious Employees

Sherrell,Biebel	Alex Hall	Meryl Pastuch
Julie,Tierno	Glen Grant	Sherrell Biebel
Kerstin Belveal	Julie Tierno	Rosalia Larroque
Sherrell,Biebel	Dylan Ballard	Richard Fox
Meryl Pastuch	Augusta Sharp	Madeline Nindorf
Lindsy Henion	Meryl Pastuch	Calvin Davidson
Ricky Miles	Lindsy Henion	Sherlyn Wombacher
Marian Ahmadi	Yer Dolph	Bethanie Folmer
Loriann Gerard	Jade Meucci	Dortha Bratt
Kerstin Belveal Craig	Maria Hupman	Violet Little
Carr Chang Tulip	Adele Farmer	Tajuana Lampron

Table 9: Values of TN, FP, FN and TP

Confusion Matrix	
TN=0	FP=13
FN=0	TP=20

References

http://vacommunity.org/vast+challenge+2018.

Depren, O.; Topallar, M.; Anarim, E.; and Ciliz, M. K. 2005. An intelligent intrusion detection system (ids) for anomaly and misuse detection in computer networks. *Expert systems with Applications* 29(4):713–722.

Eberle, W., and Holder, L. 2007. Anomaly detection in data represented as graphs. *Intelligent Data Analysis* 11(6):663–689.

Jedrzejek, C.; Bak, J.; and Falkowski, M. 2009. Graph mining for detection of a large class of financial crimes. In *17th International Conference on Conceptual Structures, Moscow, Russia*, volume 46.

Ketkar, N. S.; Holder, L. B.; and Cook, D. J. 2005. Subdue: Compression-based frequent pattern discovery in graph data. In *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, 71–76. ACM.

Michalak, K., and Korczak, J. 2011. Graph mining approach to suspicious transaction detection. In *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, 69–75. IEEE.

Mookiah, L.; Eberle, W.; and Holder, L. 2014. Detecting suspicious behavior using a graph-based approach. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, 357–358. IEEE.

Rissanen, J. 1985. *Minimum description length principle*. Wiley Online Library.

Vaqar, S. A., and Basir, O. 2009. Traffic pattern detection in a partially deployed vehicular ad hoc network of vehicles. *IEEE Wireless Communications* 16(6).

Velampalli, S., and Eberle, W. Novel graph based anomaly detection using background knowledge. In *Florida AI Research Society (FLAIRS), May 2017.*

Velampalli, S., and Eberle, W. 2016. Novel application of mapreduce and conceptual graphs. In *Computational Science and Computational Intelligence (CSCI), 2016 International Conference on*, 1107–1112. IEEE.