Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 50 (3D) 126-129 (2018)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Temporal Analysis of Egocentric Email Network by using Graph Database

K. NUSRATULLAH, A. SHAH, N. KHAN

Department of Computer Science & Engineering, Yanbu University College, KSA

Received 10th June 2018 and Revised 15th September 2018

Abstract: Necessity is the mother of invention. It is a very famous proverb and hundred percent true in the case of computer technology. Computer scientists are extensively involve to invent technology that fulfills the requirement of exploring today's big data. Today technology is not only needed to store and manage data, it's more needed to analyze and discover knowledge from data. Data analysis is being considered as a major evaluation criteria to dig useful information. It supports organizations to lead smarter moves and more efficient operations on the basis of knowledge evaluation. Egocentric analysis of email networks on temporal basis is the striking field that results thought-provoking analysis for a personal network. Such networks revolve around the single individual and its relationship and depict interesting changes in an individual's network as the time passed by. Technologies like relational databases, NOSQL, Hadoop, EgoLines are working as the analytical tools to capture the provocative timely changes of egocentric networks. Among them an emerging technology called Graph Databases is getting popular as an efficient tool to depict better analysis in the said fields. It is being observed that graph databases are like the next generation of relational databases with proficient support for "relationships, flexible and fine-grained data model that allows modeling and managing rich domains in an easy and intuitive way. The intension of this work is to present a review study of GDBs in order to identify the success graph of this emerging technology in the field of data analysis. Also, it is an effort to investigate the uttermost functionality of the GDBS in order to analyze the egocentric email networks on temporal basis.

1. INTRODUCTION

With the widespread use of technology the requirements of today's user are changed. Previously people used to seek the smart solution for traditional applications like retail recommendation engines or fraud detection. Now they are more interseted to dig information of connected-data via intelligent, high performance and scalable data repositories.The requirement doesn't end at these characteristcs.The visualization tools are also in demand to illustrate the information (Smith, 2017). Also,the existence of shared networks have enhanced the importance of connected information. Google, Facebook, Orkut, Gmail etc are well-known examples of large shared networks of dynamic data. Shared Networks is one of a huge repository to obtain valuable information in the various field. For the field of advertising a shared network may be useful for community managers to target the the right sets of people in the network. Sometimes the shared networks carry the hidden evidences for criminal organizations. Likewise, shared networks are considered as a source to identify the reasons of spreading illnesses in the form of epidemics (Lluis, 2014). Different shared media exist to collect and share the diverse pieces of information. (Smith, 2014).

Egocentric email networks is a shared network that shows the network of an individual. The individual is the focal node with his/her contacts and their relationships. In contrast to whole-network analysis, egocentric analysis focuses on the local subnetwork around a particular node, the ego, and its surrounding neighbors, the alters. The ego is the central actor of interest in a particular domain (e.g., an individual, a device, or a synapse). The boundary of an egonetwork is essential to define in terms of levels. If the boundary limit is defined as 1-level its mean that ego-network includes only alters directly connected to the ego, while a 2-level ego-network includes all alters within a path distance of two, and all connections between them. In practice, only 1-level and 2-level ego-networks are typically considered (Prell. 2011).

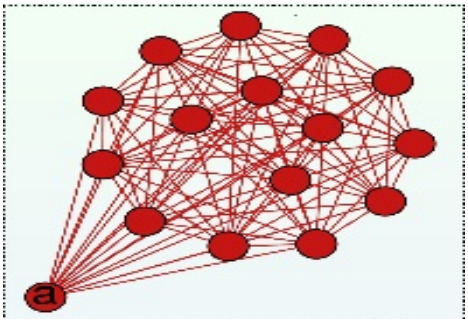

Fig 1: 1-level Egocentric Network

Kulliyyah of Information & Communication Technology, International Islamic University, Malaysia.
Department Of Computer Science & Engineering, Yanbu University College, KSA

To extract information from such shared networks "S Network Analysis (SNA)" is an effective and popular technique. SNA is an important and valuable technique for knowledge extraction from massive and unstructured data (Mincera, *et al,* 2012).A number of tools are available that can efficiently help to analyze shared networks. Gephi, NodeXL, Pajek etc are some of the names. Other tools can be identified in june 2015's report of KDnuggets news. Together with these GDBs is also an evolving technology that could play a noteworthy role in the field of SNA.

A temporal analysis of egocentric email networks shows the dynamic extractions of networks like addition and deltion of links,formation of new clusters,a certain communication pattern between specific nodes,difference in prominence of ego or alters,the intensity of relationship among ego or alters etc.The connections-first approach of GDBs create a compatibility with these extractions of egocentric email network.

This focal point of this paper is to find out the compatibilies of GDBs in order to analyze the egocentric email networks.The paper is further organized in to two more sections.The second section explains some features of GDBs that makes it compatible to find the dynamics of egocentric email networks.The third section defines the recent applications of GDBs.

2. CHARACTERISTICS OF GDBS

In this era of computing , the world is connected with rich domains all around us.One of a natural structure of connecting object is graph. Graph DBMS, also called graph-oriented DBMS or graph database, represent data in graph structures as nodes and edges, edges are relationships between nodes and allow easy processing of data in graphical form. Several tools for SNA have made the graphical analysis conveinient.They provide the good charateristics to connect, extend and visualize the information.Next section is intendid to highlight these features in term of GDBs.

2.1. A native graph approach is more flexible

The native graph approach of GDBs is more flexible and continually in flux. New nodes, properties and edges are constantly added and removed as situations change.Temporal analysis of egocentric emails essentially requires this change to capture. Conventional schemas where columns or fields are predefined, do not provide such flexibility. Graph databases are usually schema-less and allow a set of nodes with dynamic properties to be arbitrary linked to other nodes through edges.

2.2. Relationships at first priority with flexible schema

Relationships are the fundamental of egocentric email network analysis and in GDBs the relationships are at first priority with flexible schema. As we said earlier that the data became large in volume and more interconnected so some time ad hoc relationships are required to process the business logic. At this point the greatest weakness of relational databases is that their schema is not flexible. Also the dynamic networks are constantly changing and evolving. Whereas the schema of a relational database can't efficiently keep up with these dynamic and uncertain variables. To pay off, the result will be more null able columns with more code that multiplies data's complexity and diversity. Finally the performance suffered with large join tables. In difference, graph databases store data relationships as relationships. The flexibility of a graph model allows to add new nodes and relationships without compromising existing network or expensively migrating data. All of the original data and its original relationships remain intact. Graph databases are incredibly efficient when it comes to query speeds, even for deep and complex queries.

2.3. Scalability and Performance

The evolving nature of dynamic email networks require the scalability and performance. The graph processor engine in some graph databases like Neo4j is capable of supporting sub-second graph queries on large data sets to allow for real-time decision making and excel at managing highly connected data and complex queries. Some high-performance graph database allow for a compact representation of the data, basing its implementation on bitmaps. Performance of graph databases is better; graph databases directly store the relationships between records as pointers rather to create a foreign key in another table. This approach decreases the search operations. For example if you would like to search for all of the telephone numbers for alters in area code "422", the engine would first perform a conventional search to find the users in "422", but then retrieve the email addresses by following the links found in those records. While the relational database will perform two search operations, first it would find all the users in "422", extract a list of users who belongs to area code 422 then it will perform another search for any records in the address table that will match the telephone number with those

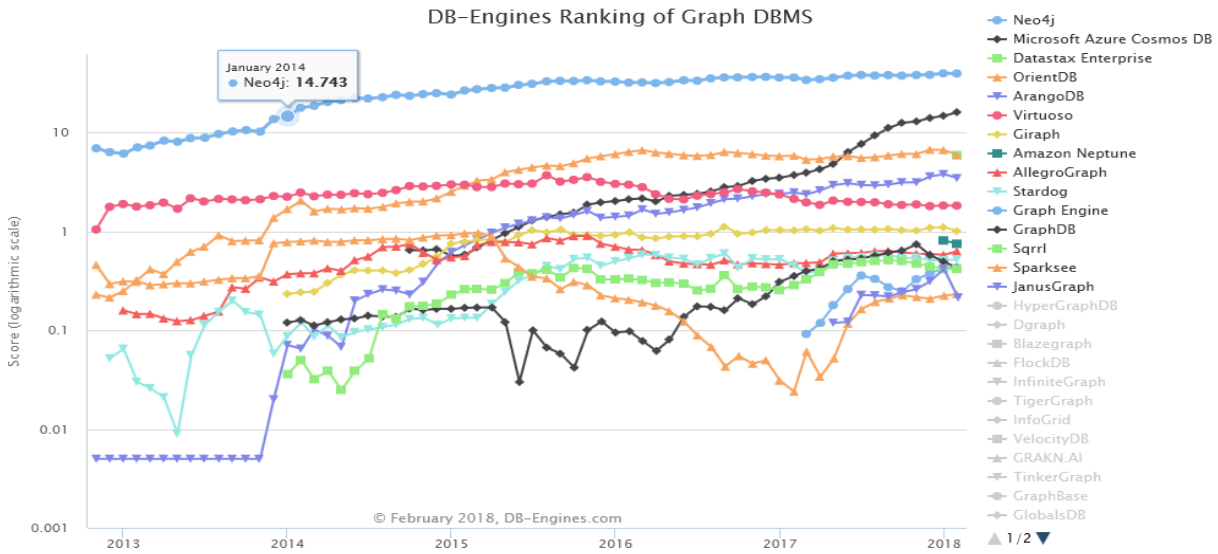

Fig 2:DB Engines Ranking (Gelbmann, 2018)

users, and link the matching records together. For these types of common operations, a graph database is significantly faster (Neo,Tec.,2016).

2.4. Visualization

Graph is called the complex and rich data structure because it adapts changes with time.While analysing such periodic changes in egocentric email network,it is an essential requirement to visualise the changes in an efficient manner so that the extracted network changes could be easy to understand.GDBs like Neo4j has an ability to integrate with graph visualization toolkit that can provide the full support for dynamic graphs. Vendors like Cambridge Intelligence released one of such toolkit that is a Javascript based solution named "KeyLines 2.0" (Grant, 2015).The data obtained from queries can be transferred via JSON and the graph are displayed after formatting.KeyLines has an ability to display the attractive snapshots of network according to selected range of time window.

3. STATE OF THE ART

The real world is richly interconnected and graph database is a technology that is easily fit in this scenario.A dynamic technology like graph databases are extremely useful to design real time applications like, logistics route optimization, retail suggestion engines, fraud detection and egocentric network monitoring etc. This section is talking about the real time implementation of graph databases by different vendors. ***Fig 2*** is showing a graph depict that graph databases are on the rise and gaining in popularity faster than any other database category, growing 300 percent since January of last year. (Gelbmann, 2018).

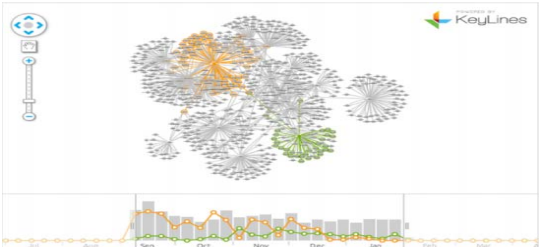

Fig 3:Time based visualization of network (Parodi, 2018)

Graph Databases are the ideal enabler for efficient and manageable fraud detection solutions (Tribolet , 2016). The technology simplifies and speed up access to data that is complex and contains many connections. Many have taken a shift toward graph database technology. Companies like Linkurious have played an important role in designing real time applications using graph database. The Panama Papers investigation is a brilliant example of what can be accomplished using the graph database technology. In this biggest data leak in history, they uncovered the names of various criminals, high-level politicians and stars who were involved in money laundering and sometimes tax evasion, and they were able to find these suspicious connections via a visual and interactive interface. A group of about 370 journalists used Linkurious to investigate and search the Panama Papers data, which included 11.5 million documents (Sarum 2016). Twitter has released FlockDB as open source, graph database. Neo Technology, the creator of Neo4j, the most popular graph database is adapted by more than 30 Global 2000 companies including enterprise brands like Wal-Mart, eBay, Lufthansa, and Deutsche Telekom.Teradata just released a new type of SQL called SQL-GR, intended to make the graph analytics easy for enterprise users.

Linkurious has analyzed the ENRON dataset by using GDB approach. They analyzed the email activity

of Tim Belden, the head of trading at Enron, was one of the first executives to be prosecuted and to admit wrongdoings at Enron. While investigating his email network Linkurious has identified two more key players who were culprits in the whole story. The functionality of graph dabase is evaluated as the ego centric email analyzer on temporal basis. The Hillary Clinton email archives being released by the US Department of State.The said dataset is used for experiment.The sender,receiver and the time stamp of send email was visible in the network (Gussman, 2018).

Another use case of graph database is Master Data It is a comprehensive method of enabling an enterprise to link all of its critical data to one file, called a master file that provides a common point of reference to share among personnel and departments. It is a good example of integrating ,querying and reducing the vastly different sources of data. The visual representations of data objects play a considerable role in simplifying queries of different data types.

Additionally ,GDBs are also being used to design numerous data organization applications like gene sequencing, mobile shared application, portfolio analytics,web browsing and content organization etc.

4. CONCLUSION

The world of computing is moving with a high speed. Vendors are trying their level best to supply on demands. Technologies are being introduced as per demand of people. The complexities, inter connectedness and dynamics nature are some of the requirements of today's big data. Such complicated, dynamic interconnected systems tend to be less static and predictable, and are ideal candidates for graph databases. GDB is an emerging technology and it is in a phase where people can easily see its effectiveness as a result of several real time applications. Sparksee, Neo4j or OrientDB have teams that are doing huge efforts to make GDBs more understandable and solve the issues explained above. More number of applications like business users, Shared network analyst,fraud detector are becoming more comfortable with graph analytics. GDBs are much capable to analyze the close relationship like how many steps are required to get from one point to another or how many "degrees of separation" are there between two people? Besides that the flexible dynamic structure of GDBs supports the fast retrieval of data, they make it easy to import data without creating complex schemas. These charateristics of GDBs provide ease to analyze the ego centric network anaylsis on temporal basis. Graph databases are easy to query and navigate using the Cypher query language and available visualisation tools, even for native users. Nevertheless it's interesting to note that Graph DBMS are grabbing an ever-larger slice of developer's attention.

In the light of these facts it could be said that GDBs is an efficient technology to store information about the relationships between things especially where the relationship between two items in the database is as important as the items themselves. Hence with these benefits GDBs has made a good reputation in market and organizations can take a chance to implement this technology.

REFERENCES:

Data Mining, Analytics, Big Data, and Data Science KDnuggets June (2015).

From Relational to Neo4j,Blog, (2016) Neo Technology, Inc.

Gelbmann, M., (2018). Graph DBMSs are gaining in popularity faster than any other database category,DB-Engines, Knowledge Base of Relational and NoSQL Database Magement Systems.Copyright © 2018 solid IT gmbh. https://db-engines.com/en/ranking_trend/graph+dbms.

Grant, D., (2015). How to visualize time-based graphs with Neo4j. https://cambridge-intelligence.com/visualize-neo4j-time-graph/

Gussman, A.,(2018). GraphGist: Shared Networks in the Clinton Email Corpus https://neo4j.com/graphgist/shared-networks-in-the-clinton-email-corpus

Lluis, J.,(2014). Why is SNA important? Technical University of Catalunya.

Mincera, M., E.,N. Szynkiewicz, (2012). Application of SNA to the Investigation of Interpersonal Connections, Journal telecommunication and information technology.

Neoj powers the biggest financial leaks in history – the tax haven scandals exposed in 'the panama papers', sarum pr, (2016).

Prell. C. (2011). SNA: History, Theory and Methodology. SAGE. https://us.sagepub.com/ en-us/nam/socil-network-analysis/book231856.

Parodi, M., (2018). An Introduction to KeyLines and Network Visualization. Cambridge Intelligent.

Smith, C., (2014). big data: Each S Network is Using A very different Data Lens Understand And Target Users.

Smith,T., (2017). How Neo4j Is Making Graph Technology More Accessible. https://dzone.com/articles/how-neo4j-is-making-graph-technology-more-accessib.

Tribolet M., (2016). Investigating Enron's email corpus: The trail of Tim2222 Belden, Linkurios visualize graph data easily.