Measurement and Social Network Analysis With Parallel Frequent Pattern Mining

YERMES Mohammed EL Amine
University of Mustapha Stambouli, Mascara, ALGERIA
amine.yermes@univ-mascara.dz

Mohammed REBBAH
University of Mustapha Stambouli, Mascara, ALGERIA LRSBG Laboratory
rebbahmed@univ-mascara.dz

Abstract: Social network analysis ('SNA') measures are a vital tool for understanding the behavior of networks and graphs. However, the huge amount of data generated by networks require new techniques to optimize the calculation. Many efficient algorithms based on graph calculation have been used in the last decades; most of them do not scale the large amount of data.

In this paper, we propose a new approach based on parallel frequent pattern mining as an essential data-mining task, with a goal of calculating degree centralities. To take advantage of the computing power provided by High Performance Computing clusters (HPC), the use of hybrid distribution of data (horizontal and vertical) seems necessary to reduce the computing time. The implementation show how is benefic to combine parallel and distributed programming techniques such as MPI with data mining tools.

Keywords: ***Parallel Frequent Pattern Mining, Social Network Analysis, Degree Centralities, High Performance Computing***

I. Introduction

Social networks, dynamic structures made of individuals or organizations, have always played a major role in our societies in the last years. Facebook alone attracts 1.3 billion users with 640 million minutes spent each month on the site. Consequently, discovering trending topics or influential users is of interest for many researchers interested in areas such as marketing [1]. However, to take advantage of all these platforms we have to offer new tools to help us to understand this phenomenon. Social Network Analysis or SNA has become one of the hottest topic of new information technologies.

Traditional algorithms based on graph calculations have been studied but cannot fit on the huge data sets generated by social networks and their new characteristics such as distribution, traditional algorithms are often unsuitable. Data mining, an important step in this process of knowledge discovery, consists of methods that discover interesting, non-trivial, and useful patterns hidden in the data [2]. Frequent pattern have been well studied for discovering regularities between items in relational data.

The focus of our work is the calculation of degree centralities of users and groups in social networks using parallel frequent pattern mining. To really take advantage of HPC platforms we had to review Apriori algorithm to reduce execution time and get better results. The rest of the paper is organized as follows. Next section gives the background and related work. Section III we propose our model. Evaluation is presented in Section IV. Conclusions are given in Section V.

II. Related Work

Social Network Analysis (SNA) is a sociological approach for analyzing patterns of relationships and interactions between social actors in order to discover underlying social structure such as: central nodes that act as hubs, leaders or gatekeepers; highly connected groups; and patterns of interactions between groups [3]. Data mining techniques have been found to be capable of handling the three dominant disputes with social network data namely; size, noise and dynamism. The voluminous nature of social network datasets require automated information processing for analyzing it within a reasonable time. Interestingly, data mining techniques also require huge data sets to mine remarkable patterns from data; social network sites appear to be perfect sites to mine with data mining tools [4]. A number of research issues and challenges facing the realization of using data mining techniques in social network analysis could be identified as follows: Linkage-based and Structural Analysis, in this analysis we construct the linkage behavior in order to determine communities, important nodes, and links. Another issue is the adding content-based; it has been observed that content-based analysis with linkage-based analysis provides more effective results in a wide variety of applications.

Distributed computing plays an important role in the Data Mining process for several reasons. First, due to the important size of data generated by social networks Data Mining techniques often requires huge amounts of resources in storage space and computation time. To make systems scalable, it is important to develop mechanisms that distribute the workload among several sites in a flexible way. Second, data is often inherently distributed into several databases, making a centralized processing of this data very inefficient and prone to security risks. Distributed Data Mining explores techniques of how to apply Data Mining in a non-centralized way.

Previous research on application of frequent pattern mining in social networks datasets are briefly summarized in the following paragraphs. Lahiri and Berger-Wolf [5] introduce mining periodic behavior in social networks, Different teams around the world also work on: (i) personality prediction for micro blog users [6], (ii) churn prediction and its influence on the network [7,8], (iii) community evolution prediction [9,10].

III. Proposed Model

Modeling a social network with graphs having similar properties to real networks, allows to perform simulations of failures, attacks, propagation and other events that may occur on real networks. In 1959, Erdős and Rényi [11] published a seminal article in which they introduced the concept of a random graph. A random graph is simple to define. One takes some number N of nodes or "vertices" and places connections or "edges" between them, such that each pair of vertices i, j has a connecting edge with independent probability p. The Erdos-Renyi model, failed to replicate the clustering, triadic closure, and hubs seen in real-world networks. As a result, there were two notable models created in an attempt to fix some of the problems Erdos-Renyi had: The Watts-Strogatz model [12], which generated random-graphs

with small-world properties, and the Barabasi-Albert model [13], which generated scale-free networks through preferential attachment. As it turns out, most large networks in the present day have scale-free structure, or a network structure in which most nodes have few connections, some nodes have a medium number, and a few nodes are very well connected.

For the application of a data mining process on social networks, switching to a binary data table for automatic processing is necessary. The matrices are the most appropriate mathematical model for our design, where the columns and lines represents the nodes of the graph (individuals of the social network) (see Fig 1).

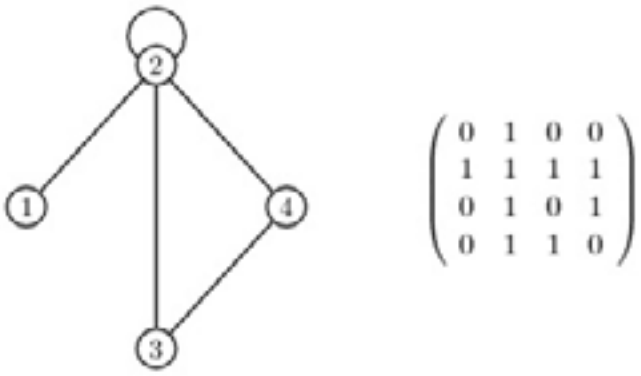

Fig.1. Example of a graph and its matrix

3.1 Association Rules

Association rules analysis is a technique to uncover how items are associated to each other. The problem of association rules extraction can be formalized as follows:

Let I = {i1, i2, …, in} be a set of literals call items. Let D be a set of all transactions where each transaction T is a set of items such that T ⊆ I. Let X, Y be a set of items such that X, Y ⊆ I. An association rule is an implication in the form X ⇒ Y, where

$$X \subset I, Y \subset I, X \cap Y = \emptyset \text{ [14]}.$$

Essentially, association mining is about discovering a set of rules that are shared among a large percentage of the data [15]. Association rules mining tend to produce a large number of rules. The goal is to find the rules that are useful to users. There are two ways of measuring usefulness, being objectively and subjectively. Objective measures involve statistical analysis of the data, such as support and confidence [14].

Support: The rule X ⇒ Y holds with support s if s% of transactions in D contain X ∪ Y.

Items that have an s greater than a user-specified support are declared to have minimum support.

Confidence: The rule X => Y holds with confidence c if c% of the transactions in D that contain X also contain Y. Rules that have a c greater than a user-specified confidence are said to have minimum confidence.

3.2 Distributed Apriori

The mining of huge amount of data stored in databases or data warehouses from social networks required asubstantial processing power. Distributed systems such as HPC with a no common memory needs an intelligent distribution of data to take advantage of the power of calculation provided by HPC. In our approach, we use a hybrid distribution of data. The principle is to realize a hybrid distribution by an algorithm that takes partitioning of the data set so that each node has a vertical partition with another horizontal to keep the active role of the calculation unit and reduce the waiting time (see Fig 2). In the following, a calculation unit is a processor or core with no common memory.

Items Transaction	A_1	…	A_k	…	A_n
T_1					
…					
T_k					
…					
T_n					

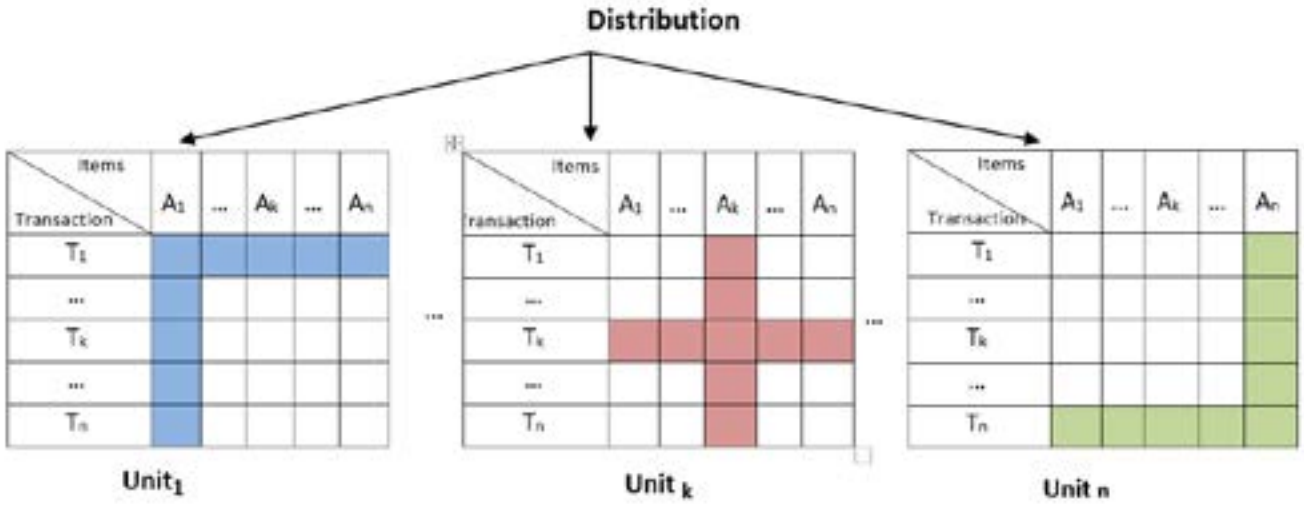

Fig.2. Hybrid distribution of the base on *n* units.

As number of algorithms have been proposed for efficient frequent pattern mining on a distributed system, which include Count Distribution, Data Distribution and Candidate Distribution algorithms by Agrawal and Shafer [16], Parallel Eclat by Zaki el. al. [17], Parallel FP-growth algorithm by Li el. al. [18], Single Pass Counting, Fixed Passes Combined counting and Dynamic Passes Combined-counting algorithms by Lin et. al. [19] and Distributed Eclat algorithm by Moenset. al. [20]. In our work we implement HV-Distrib algorithm [21], proposed in previous paper. HV-Distrib is more suitable for platforms like HPC.

3.3 Contribution

As mentioned above the main goal of our work is the calculation of degree centralities using frequent pattern mining. Degree centralities of groups and users are the most studied measures in social networks analysis. In graph theory the calculation of these measures are formulated as follow:

Degree centrality of a user: The degree centrality indicates how well a node is connected in terms of direct connections, The degree centrality Cd(i; g) of node i in network g is given by

$$C_D(i) = \frac{d(i)}{n-1}$$

Where the degree di(g) of a node i in g is the number of i's neighbors in g, n is the number of all nodes [22]. Using the frequent pattern mining, the centrality degree of a user is equal to the support of the attribute in the database representing the user.

Degree centrality of a group: The degree centrality of groups refer to the number of nodes connected to one node (member of the group) at less, normalized by the number of nodes outside the group.

$$GC_D(C) = \frac{N(C)}{|n|-|C|}$$

Where N(C) is the number of nodes connected to the group, C number of nodes in the group [22]. Using Apriori algorithm we can deduct that degree centrality of group of users represent the support of itemsets containing this users.

IV. Evaluation and performance

4.1 Development tools

We have developed our application on HPC Emir at the University of MASCARA; HPC Emir has a theoretical computing performance equal to 12 teraflops and a storage capacity of 11 TB running on Redhat RHEL 6.4. HPC Emir is composed of 32 computation Nodes, each one have 20 cores and 64 Gbytes of RAM. Nodes are connected with QR Infiniband network 40 Go/s. Parallel programming is ensured by MPI with JAVA.

4.2 Data and Results

For our tests, we explore binary datasets (0/1) in order to test the functioning of the application and the validation of its results. The data represent graphs from real social networks. A main feature of these datasets is the large size of these datasets, and other properties mentioned in the following table:

TABLE I. GRAPHS CHARACTERISTICS

Data sets Name	Size (Ko)	# nodes	# links
BlogText	201815	10 738	9 381
Random Graph	60	90	80

The table below represents the different results of different data sets taking several values of confidence and support.

TABLE II. TABLE 2: RESULT OF TESTS

Data sets	Support	Confidence	#Partitions	#Frequent Itemsets	#Rules	Time execution
BlogText	10	11	1 (Sequential)	211	173	00 : 10 : 58
	10	11	3	211	173	00 :07 :54
	10	11	5	211	173	00 :06 :06
	10	11	10	211	173	00 :04 :17
	10	11	15	211	173	00 :03 :56
Random Graph	75	85	1 (Sequential)	537	12	00 :05 :57
	75	85	3	537	12	00 :03 :34
	75	85	5	537	12	00 :01 :30
	75	85	10	537	12	00 :00 :50
	75	85	15	537	12	00 :00 :30

The results obtained in the phase of the test confirm the interest of the emergence of Data mining in HPC platforms. And to better demonstrate this interest we have compared our results with others obtained with vertical and horizontal distributions implemented on one partition (see Fig 3). The results of experimentations presented by Oded Green and David A. Bader [23] show that using parallel frequent pattern mining to calculate degree centralities is very beneficial in order to reduce execution time (see Fig 4). The algorithm used by Oded Green and David A. Bader is based on graph calculation.

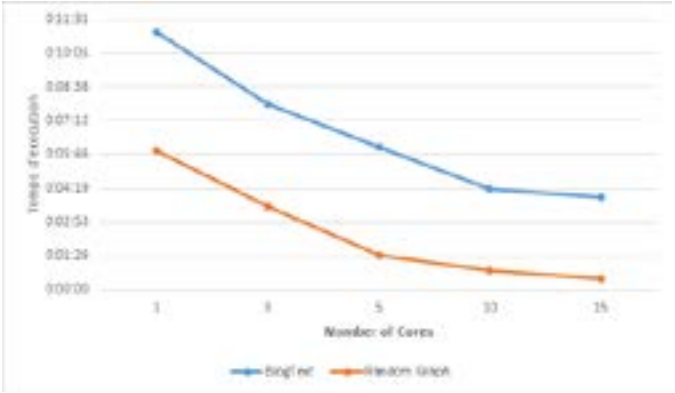

Fig.3. Reduced execution time compared to Cores numbers.

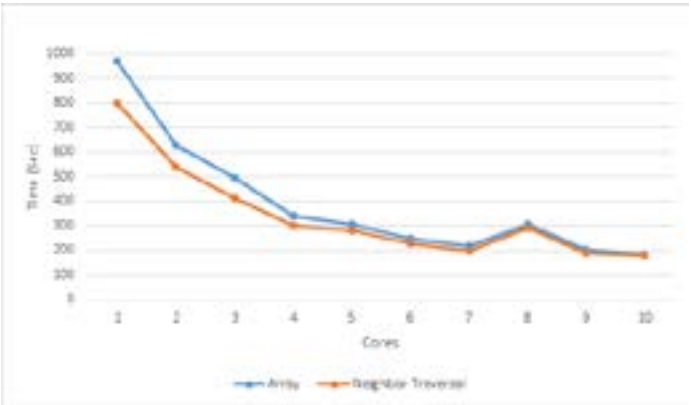

Fig.4. Execution time of the fine grain parallelism using the SNAP package.

V. CONCLUSIONS

This article presents two contributions. Firstly, a new design and implementation of distributed Apriori algorithm for High Performance Computing platforms based on hybrid distribution of data. This new vision of Apriori reduce the time of execution and can exploit a huge amount of data due to the capacity of HPC servers and MPI as a powerful parallel programming paradigm. Secondly, it is possible to generate degree centralities calculation from frequent itemsets. This new method give more efficient results than traditional algorithms based on graph calculation.

For future work, it would be interesting to include association rules extraction in more graph calculation and measures, communities detection can benefit as well from social network mining.

References

[3] Cha, M.; Haddadi, H.; Benevenuto, F.; Gummadi, P.K. Measuring User Influence in Twitter: The Million Follower Fallacy. ICWSM **2010**, 10, 10–17.

[4] M. Stonebraker, R. Agrawal, U. Dayal, E. J. Neuhold, and A. Reuter. DBMSresearch at a crossroads: The vienna update. In Proc. of the 19th VLDBConference, pages 688{692, Dublin, Ireland, 1993.

[5] Wasserman, S., & Faust, K ; Social Network Analysis: Methods and Applications. New YorkCity, New York, U.S.A.: Cambridge University Press1994.

[6] Cortizo, J., Carrero, F., Gomez, J., Monsalve, B., Puertas, E.:Introduction to Mining SM. In: Proceedings of the 1st InternationalWorkshop on Mining SM, 1 – 3, 2009.

[7] Lahiri, M. and Berger-Wolf, T. 2010. Periodic subgraph mining in dynamic networks. Journal of Knowledgeand Information Systems 24, 3, 467-497.

[8] Zu, Q.; Hu, B.; Gu, N.; Seng, S. Human Centered Computing. In Proceedings of the 1st Human Centered

[9] Computing Conference International Conference, (HCC 2014), Phnom Penh, Cambodia, 27–29 November2014.

[10] Au, W.H.; Chan, K.C.; Yao, X. A novel evolutionary data mining algorithm with applications to churnprediction. IEEE Trans. Evolut. Comput. 2003, 7, 532–545.

[11] Ruta, D.; Kazienko, P.; Bródka, P. Network-Aware Customer Value in Telecommunication Social Networks.

[12] In Proceedings of the 2009 International Conference on Artificial Intelligence, (ICAI'09), Las Vegas, NE,USA, 13–16 July 2009; pp. 261–267.

[13] P. Erdös, A. Rényi, On random graphs i. Publ. Math. Debrecen, 6:290–297, 1959.

[14] Watts, D. J.*;* Strogatz, S. H. *(1998).* "Collective dynamics of 'small-world' networks". *Nature. 393 (6684): 440–442.*

[15] Réka Albert et Albert-László Barabási, « Statistical mechanics of complex networks », Reviews of Modern Physics, vol. 74, no 1, ٢٠٠٢, p. –٤٧ ٩٧ (ISSN 0034-6861)

[16] Agrawal, R., Imielienski, T., and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In: *Proc. Conf. on Management of Data*, 207–216. New York: ACM Press

[17] Zaki M.J., Scalable Algorithms for Association Mining. IEEE Transaction on Knowledge and Data Engineering, Vol. 12, No. 3, pages 372-390, May – June 2000.

[18] Agrawal, R., and Shafer, J., 1996, Parallel Mining of Association Rules, IEEE Transactions on Knowledge and Data Engineering, 8(6), pp. 962-969.

[19] Zaki, M. J., Parthasarathy, S., Ogihara, M., and Li, W., 1997, Parallel Algorithms for Discovery of Association Rules, Data Mining and Knowledge Discovery, 1(4), pp. 343-373.

[20] Li, H., Wang, Y., Zhang, D., Zhang, M., and Chang, E. Y., 2008, Pfp: Parallel FP-growth for Query Recommendation, Proc. ACM Conference on Recommender Systems, Lausanne, pp. 107-114.

[21] Lin, M. Y., Lee, P. Y., and Hsueh, S. C., 2012, Apriori-based Frequent Itemset Mining Algorithms on MapReduce, Proc. 6th International Conference on Ubiquitous Information Management and Communication, Kuala Lumpur, Article no. 76.

[22] Moens, S., Aksehirli, E., and Goethals, B., 2013, Frequent Itemset Mining for Big Data, Proc. International Conference on Big Data, Santa Clara, California, pp. 111-118

[23] REBBAH; M, YEMRES; M A; Hybrid Distribution for Association Rules Extraction on Grid Computing. Proceedings of 2015 International Conference on Image Processing, Production and Computer Science(ICIPCS'2015) Istanbul (Turkey), June 3-4, 2015 pp. 14-22

[24] M. MALEK, Introduction à l'analyse des réseaux sociaux, Rapport EISTI-LARIS, 2009, p. 8

[25] Oded Green, David A. Bader; Faster Betweenness Centrality Based on Data Structure Experimentation. International Conference on Computational Science, ICCS 2013, p. 400.