A Review of Big Graph Mining Research

I Atastina[1], B Sitohang[1], G A P Saptawati[1], and V S Moertini[2]

[1]School of Electrical Engineering and Informatics, Bandung Institute of Technology

[2] Informatics Department, Parahyangan Chatolic University

*imeldas3@students.itb.ac.id

Abstract. Big Graph Mining" is a continuously developing research that was started in 2009 until now. After 7 years, there are many researches that put this topic as the main concern. However, there is no mapping or summary concerning the important issues and solutions to explain this topic. This paper contains a summary of researches that have been conducted since 2009. The result is grouped based on the algorithms, built system and also preprocess techniques that have been developed. Based on survey, there are 11 algorithms and 6 distributed systems to analyse the Big Graph have been improved. While improved pre-process algorithm only covers: sampling and compression technique. These improving algorithms are usually aimed to frequent sub graphs discovery, whereas slightly those of is aimed to cluster Big Graph, and there is no algorithm to classify Big Graph. As a conclusion of this survey, there is a need for more researches to be conducted to improve a comprehensive Graph Mining System, especially for very big Graph.

1. Introduction

Graph Mining is a research area that is developed as a response to the enhancement of Graph database from the researcher. Graph Mining research is estimated to start since the beginning of 1994. The development is in line with a need of using data that is saved in the form of graph. The data is being saved in the form of graph since graph can be used to save a complex data. Later, the evolution technology of data storage and data processing, resulted in the emergence of a phenomenon of Big Data. This phenomenon states that data that can be produced by any transaction occurred in this world has reached Peta byte every day. And along with that phenomenon, the term of "Big Graph Mining" occurs, it is because the data saved in the form of graph is getting bigger and bigger, it reaches Tera and Peta byte. [1], [2].

The publication of researches related to Big Graph Mining is started, estimated, since 2009 and it is pioneered by U Kang and friends [3]. After that, the development of researches in this field is improving fast. It is identified from the number of publications that related to Big Graph Mining. Of course, it is because there are many fields using the result of this enormous data mining. For example, Afra Abnar et. al, wrote SSRM: Social Structural Role Mining for Dynamic Social Network. This paper explains that Graph Mining can provide information relating to someone's participation rate among his social network and the influence to the related community dynamics [4]. Also S Kutty et.al, [5] implementing Graph Mining to decide a personal compatibility. So that the information can be used to match a leader and a deputy in an organization, or students with their advisor, and so on.

There are many other examples about how to use Graph Mining, for example: a research conducted by S Gao and Li to solve a problem of CRM [2], this research is conducted to analyze the kind of protein molecule, social network analysis, web link analysis, and also business process.

The objective of this paper is to conceive the development of Big Graph Mining Research in the last 7 years, and the issues related to Big Graph Mining, so it can summarize the problems occur and the proposed solution, to solve problems in this field. For that reason, it is hoped that there will be a conclusion for any opportunity and the field for the next research, as mentioned earlier.

2. Graph Mining and Big Graph Mining

2.1. Graph Mining

Graph is a group of nodes and edges, where every node and /or edge has a label. In a graph, edge is a connector between node [6], [7]. Graph, can be said as diagram that picture the relation between objects, and the relation is represented by edges and object is represented by node. Label in a node explain the name of object and label in an edge can inform the type of relation between objects. In math, the notation of graph is **G (V, E)** where V is a node and E is an edge. The condition that must be fulfilled for a graph is V cannot be an empty set. However, E can be an empty set. As a result, a graph can consist of only a group of node with no edge at all. Notation |V| states a number of nodes in a graph, and |E |states the number of edge.

Graph Mining is an improvement form of data mining. So, basically, both graph mining and data mining have the same purpose; creating information or knowledge from numerous data in the form of pattern that might hide in the data. The difference, graph mining can be done in a graph data. This is in line with the appearance of graph database technology that apparently becoming trend in this last decade [6], [7]. Specifically, graph mining is conducted to create information or knowledge in the form of subgraph patterns from its graph. The mined graph can be groups of small graphs, it usually called as a transaction graph, and it can be a big size single graph that consists of so many nodes and edges.

In general, the techniques of graph mining clustered into 3 (three) big techniques; frequent substructure discovery or it is usually named frequent subgraph discovery or frequent pattern discovery, classification and clustering. To explain the same technique, some researchers use the term of graph analysis to substitute Graph Mining. It is related to the point of view that to get pattern or sub graph that contains certain information, the technique used in the graph is the technique of graph structure analysis and / or spectral analysis. The structure analysis strongly related to graph diameter view and the space between nodes. These two parameters are commonly used to count Page Rank and finding the connected components. While spectral analysis is strongly related to counting eigen value and eigen vector calculation from the existed metrics as a representation of a graph. The calculation of eigen value and eigen vector calculation is usually applied to detect community, count triangle pattern in a graph, etc. [6].

2.2. Big Graph Mining

The term of Big Graph Mining, as mentioned at the beginning of the paper is first introduced by U Kang et.al. This term is used to show that the important point is not only in the graph mining, but also for a big sized graph, that can reach trillion nodes and edges. This term is also used as a differentiation of Big Data Mining that is used to mine very big data, but not for a graph data. [6]–[8].

In general, a problem that must be faced with Big Data phenomenon are problems that are called Volume, Velocity and Variety. Volume is a problem related to growing capacity; which grows bigger and bigger as a result of the technology in data storage. Velocity, it is related with the speed of the high data changing. For example, there are millions of new websites in the internet, and trillions of online transactions, and statuses in the social media emerge every day. So, it can be imagined how fast the changing of the data is. [1], [9]. While variety is related with many kinds of data format that grows from time to time because of the informatics technology. People can make digital photo, sound, and

text easily. For that reason, Big Graph Mining facing 3V problem as well as Big Data Mining. Nevertheless, Big Graph Mining is more difficult because it has to mine graph structure contained complex relation between nodes.

3. The Development of Graph Mining Research

To solve problems in Big Graph Mining, there are many solutions proposed by researchers. In this paper survey, the solutions are grouped into three big parts, they are; solution through algorithm modification, solution through system building and solution related to preprocess. These three types of solution focused on the problem of volume of data or scalability. Not considering data variety and velocity. The table below shows the group of solution proposed by researchers.

3.1. Solution Using Algorithm Modification

In general, this solution is proposed because of the algorithm complexity in graph mining can be categorized as high complexity algorithm or NP-Hard Problem. Consequently, if this algorithm is used to manage a very big data, will needs big resources. The process will need long time and big memory. Based on the survey, the improvement of research that modify algorithm can be grouped into two groups, they are graph mining and graph partition. Graph Mining usually focused to search the pattern of frequent subgraph, and graph partition algorithm is aimed to make big graph become small subgraphs so it can be managed in parallel.

Table 1. Development of Algorithm Modification

No	Year	Author	Goal	Description
1	1994	DJ Cook & LB Holder (SUBDUE)[7]	Frequent Subgraph Discovery	Using Minimum Description Length (MDL)
2	2004	Kuramochi et al. (GREW) [10]	Making mining process faster.	Using Heuristic algorithm in canonical graph. Shaping bigger subgraph candidate based on smaller subgraphs and fulfill the criteria in frequent subgraph, to make smaller opportunity for infrequent subgraph to appear. Contraction mechanism; cutting infrequent subgraph.
3	2005	Kuramochi et al. (HSIGRAM & VSIGRAM)[11]	Frequent Subgraph Discovery	Implementing BFS and DFS techniques in the process of finding frequent subgraph candidate

Table 1. Cont.

4	2010	A Khan et al.[12]	Frequent Subgraph Discovery	Using proximity pattern concept. Proximity pattern is counted by probabilistic algorithm FP-Growth, where the complexity of algorithm is lower than isomorph graph technique.
5	2010	Kambatla et al.[13]	Making process to count page rank, shortest path and clustering using K-Means algorithm faster	Algorithm modification to partially synchronized. It is aimed to make it faster than the common synchronization.
6	2012	Sun et al.[14]	Making graph matching faster	Replacing the function of index to become exploration and join subgraph, so there is no need of memory to save and index management in data mining process.
7	2012	Zhao et al. (SAHAD)[15]	Frequent Subgraph Discovery	Algorithm modification to be implemented an in MapReduce framework.
8	2012	Z Zeng et al.[16]	Breaking big graph to minimize computer communication during the mining.	Using aggregation and Stepwise Minimizing Ratio Cut.
9	2013	Pattabiraman et.al [17]	Accelerating the way to maximum clique in big and rare graph. Where the complexity is categorized into NP Hard Problem.	Using exact algorithm to compare the biggest click in any and heuristic approach to start the search based on the highest node
10	2013	J Han et al.[18]	Efficiency in frequent subgraph discovery process.	Defining the pattern that is searched, in the neighborhood. Using VID (Vertex Identifier List) in enumeration phase of solution candidate

Table 1. Cont.

11	2014	U Kang et al (HEIGEN)[19]	Making spectral graph analysis	Adaptation to make algorithms to find k eigen value is able to be implemented in Hadoop environment

3.2. Big Graph Mining Development System

Since U Kang proposes PEGASUS as one of the system to analyze big graph in 2009, there are many researchers conducted related to the kind of system. Based on survey, there are 6 publications that mainly discuss distributed system to mine big graph in 2009-2014. Some of them proposes systems that can be implemented in a single machine, and the rest propose a system that adopt distributed programing. Table 2 is a summary of Big Graph Mining system that have been developed and the function of graph mining that can be applied in the system.

Table 2. The List of Big Graph Mining System

No	**Year**	**Author**	**Possible Graph Mining Operation.**	**Framework/Platform**
1	2009	U Kang et al. (PEGASUS)[3]	Counting Page Rank Counting the distance between nodes and the diameter of graph.	MapReduce
2	2010	Malewicz et al. (PREGEL) [20]	Counting Page Rank Finding the shortest path Verification of bipartite Semi Clustering graph	Bulk Synchronous Programming (BSP)
3	2011	U Kang et al. (HA-LFP)[21]	Frequent Pattern Discovery Anomaly Detection	MapReduce
4	2012	Yang et al. (SEDGE)[22]	Query in finding closest neighbor investigation of graph using random walk	-
5	2012	U Kang et al. (GBASE)[23]	Graph query	MapReduce
6	2012	Kalnis et al. (MIZAN)[24]	Counting Page Rank	BSP

3.3. Pre-processing Development

To solve problems related to big graph mining that can be categorized into preprocess group is effort to make the data smaller by compression or sampling. Based on the available publication, the researches related to this effort are only a few. Research related to sampling has ever been done by two groups of researchers, they are: Zou et.al, and Ahmed et.al. Zou and Holder research in 2010 proposed "random areas selection sampling" as sampling tool to reduce searching space in graph mining process. This technique is working by defining the value of **A** and **S**. **A** shows the number of sample areas and **S** define the number of nodes that will be taken as sample in graph. The process is started by taking **A** number of nodes as the first point for sample in each area. And then, the node that

close to each other will be added to the main node in every area, until the number of sample is **S** [25]. After 4 (four) years, there is an upcoming researcher who deal with sampling. In this research, Ahmed et.al proposed sampling model based on the edge sample using *gSH(p,q)* parameter, it stated that an edge that is not close to other edge is a sample with opportunity of *p*, or the opportunity of an edge to be a sample is *q* [26].

Next technique for preprocess is graph compression. The research related to this compression is done by making compression blocks based on the homogenous nodes or compact clusters based on specific regulation. The homogeneity can be in the form of compact cluster or non-compact cluster. This technique is equipped by metrics conversion of neighborhood in binary string, this binary string is the one that is made to compact [23]. The next is compression technique that has been done by Yongsub Lim with U Kang and Christous Faloutsos that proposed the improvement of compression method that has been done before and it is also adding some permutation to make the arrangement of compression blocks easier [27]. Both techniques can be implemented in parallel way.

4. Analysis

Based on exposed survey data, to answer big graph challenge, there are things that must be concern in the future. Many of the functions of big graph mining is related to frequent subgraph discovery and how to count Page Rank. While in there, is only few functions related to clustering and classification. Table 1 shows, from 11 developed algorithms, there are only 2 algorithms or functions that are aimed to make the clustering process more efficient. There are two things that boost this condition. First, because the big graph research is relatively new, so the researchers are still focused on building algorithm to find frequent subgraph. Second, the frequent subgraph discovery is a basis to do the classification and clustering [8], [28], [29], thus, if the algorithms to find these subgraphs is stable, then, it will be easier to elaborate it into classification or cluster algorithm for big graph. Therefore, there is a research opportunity to develop classification and clustering algorithm for big graph mining.

In addition, it is also important to note, that one of the main problems in big graph mining is the very big size of the data. A graph can consist of billion points and edges and the size of the data can reach Terabyte or Petabyte scale. For that reasons, the complexity of algorithm in frequent subgraph search has become an important parameter and it should be addressed carefully in designing algorithm. From some researches, it is found that to obtain frequent subgraphs, there are 2-steps-following data mining technique. First step is building the possibility of subgraphs forms that can be concluded as a candidate of frequently subgraphs and second step is counting subgraphs frequency and compare it to data that has been saved in database. Almost all algorithm that is developed is aimed to make the first step more efficient. There are only HSIGRAM and VSIGRAM are able to make subgraph search algorithm optimizes its searching technique, they are BFS and DFS. While other techniques modify algorithm to be more efficient, so it can be implemented and distributed. [12], [18], [19]. Seeing this condition, the improvement of algorithm is still needed in frequent subgraphs discovery. It is important to make the algorithm more effective and efficient, especially for the second step. As done by [12] who proposed the graph match by non-exact way, but by using proximity pattern technique. It is aimed to make mining process to be more precise because there is a threshold similarity between two subgraphs. This technique is better than isomorphism technique that match the graph in exact way. So there are many methods and technique that need to be investigated to reduce the complexity of the algorithm. And this is the research opportunities that are still open.

Research related to big graph mining using distributed system is also challenging. U Kang uses the development of distributed programming technology to answer the challenge big graph mining. [29]. This also the trigger for other researchers to build graph mining system so the process of graph mining can be done in distributed way and it is possible to use some machine. The strengths of distributed graph mining system is to make each computer's job a lot easier, so it is wished that the process of mining can be faster. However, there is also challenge in doing this, it is not easy to cut graph without sacrificing relational information between nodes. It is basically because the main strength of graph is

to save complex data relation. And if a graph is successfully cut and distributed to be mined, then, the next challenge would be how to minimize the communication between computers during the process. Since the communication level in mining process is high, basically, this mining process will be time consuming. So graph partitioning is also challenging problem in developing big graph mining using distributed system.

In developing Big Graph Mining system, U Kang et.al use MapReduce framework, so the problem of subgraphs distribution should be adapted with the current condition. In PEGASUS system, there is GIM-V (Generalized Matrix Vector Multiplication) that can be applied as iterative [3] . While in HA-LFP (Hadoop Line Graph Fixed Point), it is proposed to make a graph data mapping to be line graph so it can be proceed using Belief Propagation algorithm in iterative way [21]. U Kang adds graph compression feature in GBASE system in handling big graph. It is aimed to make the mining process faster, not only using distribution system, but also the appearance of graph compression [23]. Meanwhile, PREGEL system, uses vertex-centric model of computation and distributed programming uses BSP paradigm [20]. However, there is no explanation about how this system breaks the graph so that the distributed process can be executed. Aforementioned, the challenge in using distributed system in graph mining is a way or technique to break the graph. System that has been made to optimize the graph breaking are SEDGE [30] and MIZAN [24]. SEDGE optimizes graph partition by Dynamic Partitioning as it is needed by graph query, while MIZAN optimizes graph partition by considering the distribution of graph grade that is being analysed.

It can be said, to face the challenge in doing big graph mining, some researchers have proposed solutions by building the distributed system by using MapReduce and model BSP system. For that reason, it is still need more effort to adapt algorithms of graphs mining so it can be implemented based on framework and programming model used. However, as it is found in algorithm big graph mining research development, there are only few systems that are equipped by graph mining function that is aimed to classify graph and to mine graph. Nevertheless, it can be concluded that researches are still needed to solve those problems and to be implemented in distributed system.

Now, from pre-process point of view. So far, pre-process research for big graph mining can be grouped into two parts; sampling and compression technique. As mentioned earlier, sampling is aimed to make data processed small, but still represent the population that represented by the sample. It is important to consider the problems in graph data sampling. In single graph, the problem can be even bigger because each node is connected one to another, so the sampling process by taking random point or edge should keep the relation inside it.

Sampling process using random areas selection sampling technique proposed by Zao et.al, still contains some weaknesses; it does not guarantee that the data taken is random. User who use this technique should reconsider the starting point in taking the random data, whether it represents each area of the graph or not, and it should not gather in only small area. In algorithm proposed by Zao et.al, there is no checking mechanism or guaranty for this point. It seems that the researcher need knowledge about graph radius to choose the first point in every area that will be represented. However, for the second sampling approach that takes edges as a sample, it is still need to be considered how to guarantee that the graph structure would not be changed. Thus, the principle that sample is representing the population is valid. Based on questions, we can formulate the problems related to sampling technique in big graph, whether it is possible to take sample both from node and edges as combination.

For preprocess research that is related with compression technique in algorithm, Slahsburn still need human intervention to decide homogeneity and compression rules. So, there is a questions, is it possible to do compression technique automatically by seeing the characterization of each graph that is inputted

The available preprocess technique only covers sampling technique and compression. Both are aimed to reduce the size of processed graph. However, if we take the analogy from data mining techniques, it is need to think about preprocess technique to guarantee the quality of graph data, so with a good graph data input, it is wished to have a good quality of pattern.

Above all, remembering that a system is built by considering input, process and output, so, in building big graph mining, we need to do the same thing. Consequently, it is quite understandable that there are researchers who only focused on data input. It is aimed to answer the challenge of Velocity and Variety. It can be grouped as a part of pre-process part. While the part that focused on the adjustment of algorithm is actually focused on the main process in analysing of big graph mining. From the available researches, there is no research that focuses on the output. Output means how to present the information based on analysis in an understandable way. For systems like PEGASUS, HA-LFP GBASE and many other systems that are implemented in Hadoop using Map Reduce concept, the output should follow the concept of information collaboration as defined in reduce step. The weakness from the available system is the output from this system is still displayed in the form of data without user-friendly visualization. Also for other systems beside PEGASUS, HA-LFP and GBASE, there is no effort to optimize the collaboration of information as output in every computer. It means, further researches are still needed to build big graph mining system to handle comprehensive big graph mining process, start from pre-process, main process and final process, even in visualization process.

5. Research Opportunity

Based on the result and discussion in part 4, it can be concluded that at least there are 3 (three) main topic research opportunity that can be managed. Firstly, the research that related to algorithm improvement. It is needed to elaborate existing algorithm to generate classification and clustering algorithms that can handle big graph, and can be implemented in distributed algorithm. It is in line with the tendency of graph size that grows exponentially from time to time and days to days. Secondly, the improvement of pre-process, peculiarly method or technique to reduce the size of graph using sampling or compression. Thirdly, the research related to developing distributed big graph mining system especially the system that can handle preprocess, main process and final process comprehensively. Each of the third topic is substantial topic and can be broken down into smaller studies. For example, research that is focused on how to guaranteed nature of randomness in graph sampling or studies that focus on developing an automated way in determining the compression threshold. Because sampling and compression are very important to reduce the graph size.

References

[1] Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” Commun. ACM, vol. 51, no. 1, pp. 107–113, 2008.
[2] S. Gao and M. M. Li, “Research of Data Graph Mining Based on Telecommunication Customers,” Appl. Mech. Mater., vol. 443, pp. 402–406, 2014.
[3] U. Kang, C. E. Tsourakakis, and C. Faloutsos, “Pegasus: A peta-scale graph mining system implementation and observations,” in ICDM’09. Ninth IEEE International Conference on Data Mining, 2009, 2009, pp. 229–238.
[4] A. Abnar, M. Takaffoli, R. Rabbany, and O. R. Zaiane, “SSRM: Structural social role mining for dynamic social networks,” in Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014, pp. 289–296.
[5] S. Kutty, R. Nayak, and L. Chen, “A people-to-people matching system using graph mining techniques,” World Wide Web, vol. 17, no. 3, pp. 311–349, 2014.
[6] C. C. Aggarwal and H. Wang, Managing and mining graph data, vol. 40. Springer, 2010.
[7] D. J. Cook and L. B. Holder, Mining graph data. John Wiley & Sons, 2006.
[8] Y. Liu, B. Wu, H. Wang, and P. Ma, “BPGM: A Big Graph Mining Tool,” Tsinghua Sci. Technol., vol. 1, p. 004, 2014.
[9] E. Ferrara, “A large-scale community structure analysis in Facebook,” EPJ Data Sci., vol. 1, no. 1, pp. 1–30, 2012.
[10] M. Kuramochi and G. Karypis, “Grew-a scalable frequent subgraph discovery algorithm,” in Data Mining, 2004. ICDM’04. Fourth IEEE International Conference on, 2004, pp. 439–442.

[11] M. Kuramochi and G. Karypis, "Finding frequent patterns in a large sparse graph*," Data Min. Knowl. Discov., vol. 11, no. 3, pp. 243–271, 2005.
[12] A. Khan, X. Yan, and K.-L. Wu, "Towards Proximity Pattern Mining in Large Graphs," in Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, Indianapolis, Indiana, USA, 2010, pp. 867–878.
[13] K. Kambatla, N. Rapolu, S. Jagannathan, and A. Grama, "Asynchronous algorithms in mapreduce," in Cluster Computing (CLUSTER), 2010 IEEE International Conference on, 2010, pp. 245–254.
[14] Z. Sun, H. Wang, H. Wang, B. Shao, and J. Li, "Efficient subgraph matching on billion node graphs," Proc. VLDB Endow., vol. 5, no. 9, pp. 788–799, 2012.
[15] Z. Zhao, G. Wang, A. R. Butt, M. Khan, V. A. Kumar, and M. V. Marathe, "Sahad: Subgraph analysis in massive networks using hadoop," in Parallel & Distributed Processing Symposium (IPDPS), 2012 IEEE 26th International, 2012, pp. 390–401.
[16] Z. Zeng, B. Wu, and H. Wang, "A parallel graph partitioning algorithm to speed up the large-scale distributed graph mining," in Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, 2012, pp. 61–68.
[17] B. Pattabiraman, M. M. A. Patwary, A. H. Gebremedhin, W. Liao, and A. Choudhary, "Fast algorithms for the maximum clique problem on massive sparse graphs," in Algorithms and Models for the Web Graph, Springer, 2013, pp. 156–169.
[18] J. Han and J.-R. Wen, "Mining Frequent Neighborhood Patterns in a Large Labeled Graph," in Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management, San Francisco, California, USA, 2013, pp. 259–268.
[19] U. Kang, B. Meeder, E. E. Papalexakis, and C. Faloutsos, "Heigen: Spectral analysis for billion-scale graphs," Knowl. Data Eng. IEEE Trans. On, vol. 26, no. 2, pp. 350–362, 2014.
[20] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski, "Pregel: a system for large-scale graph processing," in Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 2010, pp. 135–146.
[21] U. Kang, D. H. Chau, and C. Faloutsos, "Mining large graphs: Algorithms, inference, and discoveries," in Data Engineering (ICDE), 2011 IEEE 27th International Conference on, 2011, pp. 243–254.
[22] S. Yang, X. Yan, B. Zong, and A. Khan, "Towards effective partition management for large graphs," in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 2012, pp. 517–528.
[23] U. Kang, H. Tong, J. Sun, C.-Y. Lin, and C. Faloutsos, "Gbase: an efficient analysis platform for large graphs," VLDB J., vol. 21, no. 5, pp. 637–650, 2012.
[24] P. Kalnis, K. Awara, H. Jamjoom, and Z. Khayyat, "Mizan: Optimizing graph mining in large parallel systems," King Abdullah University of Science and Technology, 2012.
[25] R. Zou and L. B. Holder, "Frequent Subgraph Mining on a Single Large Graph Using Sampling Techniques," in Proceedings of the Eighth Workshop on Mining and Learning with Graphs, Washington, D.C., 2010, pp. 171–178.
[26] N. K. Ahmed, N. Duffield, J. Neville, and R. Kompella, "Graph Sample and Hold: A Framework for Big-Graph Analytics," ArXiv Prepr. ArXiv14033909, 2014.
[27] Y. Lim, U. Kang, and C. Faloutsos, "SlashBurn: Graph Compression and Mining beyond Caveman Communities," 2014.
[28] V. Krishna, N. Suri, and G. Athithan, "A comparative survey of algorithms for frequent subgraph discovery," Curr. Sci., vol. 100, no. 2, pp. 190–198, 2011.
[29] U. Kang, L. Akoglu, and D. H. Chau, "Big graph mining for the web and social media: algorithms, anomaly detection, and applications.," in WSDM, 2014, pp. 677–678.
[30] S. Yang, X. Yan, B. Zong, and A. Khan, "Towards effective partition management for large graphs," in Proceedings of the 2012 ACM SIGMOD International Conference on

Management of Data, 2012, pp. 517–528.