Advanced Processing and Analytics for Large-scale Graphs

Yasuhiro Iida, Yasunari Kishimoto, Yasuhiro Fujiwara, Hiroaki Shiokawa, Junya Arai, and Sotetsu Iwamura

Abstract

Expectations have been rising in recent years for capabilities to extract hidden knowledge in friendship relationships, product purchase relationships, and business transaction relationships for use in promoting sales, increasing work efficiency, and other such purposes. A graph is a data structure that represents data relationships. Data mining on graph data is called graph mining. Here, we describe advanced graph mining techniques that can be used to instantly discover hidden knowledge in large-scale graphs which lead to improved work efficiency through high-speed graph processing. We also present examples of graph mining applications.

Keywords: graph mining, clustering, label propagation

1. Introduction

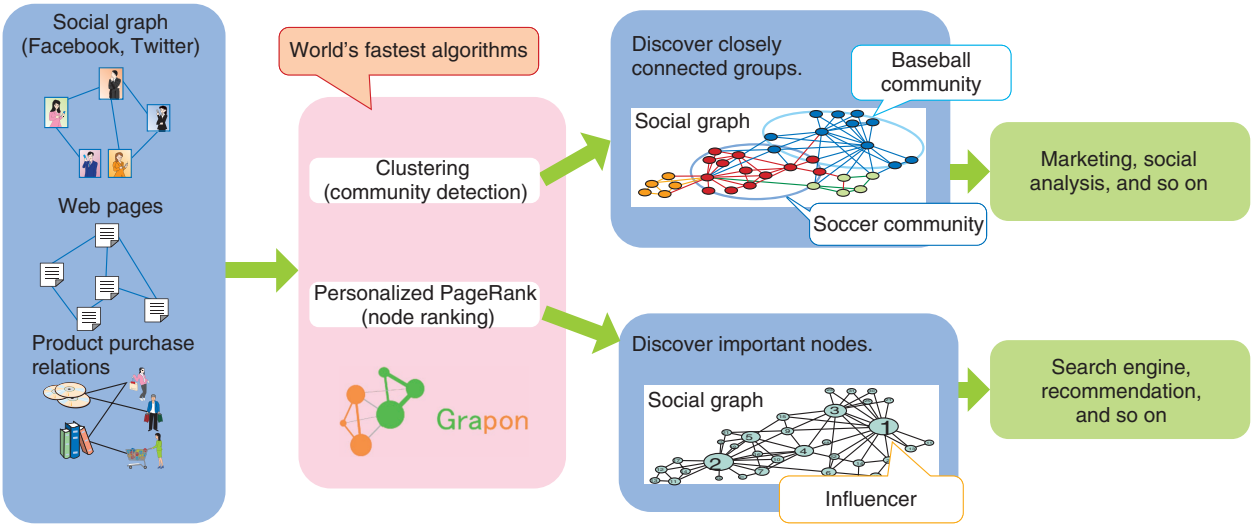
With the explosive growth of social media such as social networking services (SNSs) over the last few years, the data structures of big data have extended beyond the conventional simple table format. They now include graphs, which can represent the relationships among many kinds of information, including people, objects, and places. In a graph structure, data are represented as nodes, and the relationships between data are represented as edges. Graph structures can represent the relationships of many kinds of data that we see all around us, including the web pages and links on the Internet, the friendship relationships in SNSs, product and purchase relationships, and the relationships of roads and intersections.

The research and development (R&D) on graph technology being done at NTT laboratories includes the grouping of similar data (nodes) in large-scale graphs, the discovery of important nodes, and the suitable partitioning of large-scale graphs for handling on multiple machines. By utilizing these technologies in cooperation with other companies and academic society, we aim to produce new and more practical technology.

2. Graph mining technology for efficient analysis of large-scale graphs

Graph mining refers to the process of discovering hidden relationships in a graph such as a group of people who have strong friendship relationships or a group of people who have a high degree of influence in a certain area. Research on graph mining methods and applications has been increasing in recent years, and we are beginning to see uses in the business world as well. However, a huge amount of processing time is required when working with large graphs on the scale of the population of Japan. For such large-scale graphs, efficient graph mining that involves trial and error is not possible.

To address this problem, NTT laboratories have developed *Grapon*, a graph mining engine that is capable of performing clustering and ranking of large-scale graphs at high-speed (**Fig. 1**). Clustering (community detection) can be used to efficiently discover groups of people or things that have strong relationships in a graph. Such information can be used in various ways such as in developing marketing and product sales strategies. Personalized PageRank (node ranking) can be used to instantly rank people or things by importance according to how they are



*Facebook is a registered trademark of Facebook, Inc. in the United Sates and other countries.
Twitter is a registered trademark of Twitter, Inc. in the United Sates and other countries.

Fig. 1. Grapon, NTT's graph mining engine.

related. Such information is useful for making product recommendations.

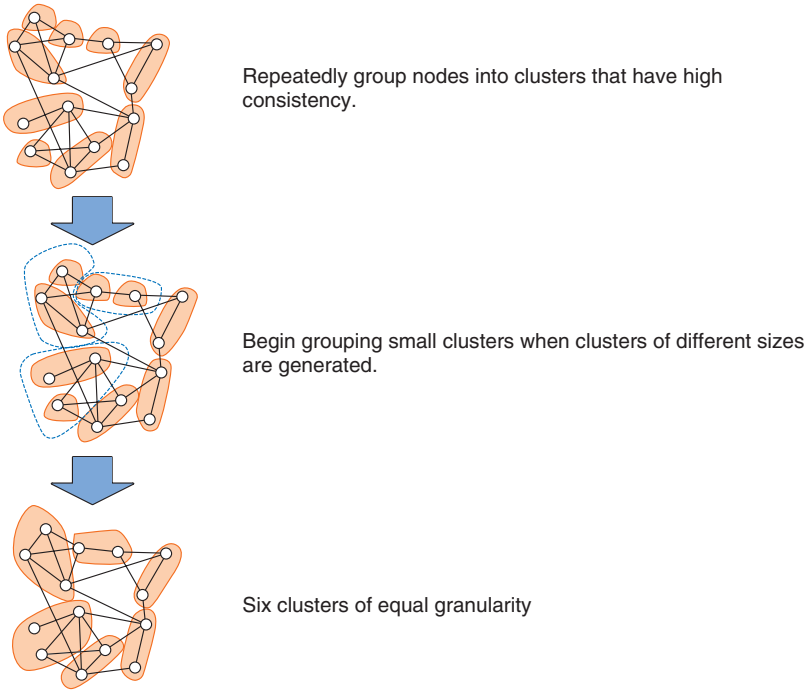
Grapon implements an algorithm that we devised and achieves the world's highest-speed graph processing. For clustering analysis in a graph on a scale of about 100 million nodes, for example, our algorithm reduces the processing time from several hours to about three minutes [1, 2]. We are currently moving forward with R&D to enable more kinds of data analysis and to develop even more advanced graph mining technology. In this article, we describe our most recent work, including techniques for partitioning graphs into clusters of equal granularity at high speed, techniques for efficiently discovering nodes that serve as mediators by connecting multiple groups in a graph, and techniques for instantly inferring node category membership.

3. Partitioning graphs using equal granularity

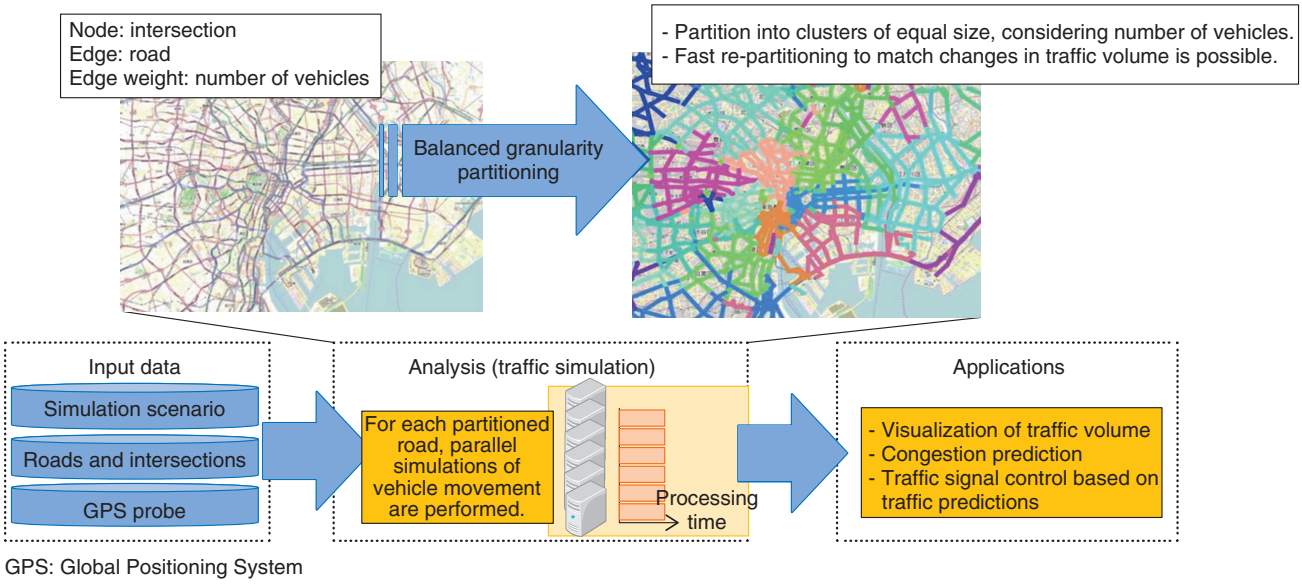
Graph partitioning is an important technique with large-scale graphs since such graphs should be stored in multiple storage systems, and they are computed in a parallel manner by using multiple machines. We have developed a novel graph partitioning algorithm called *balanced granularity partitioning* that divides graphs into same-sized partitions without sacrificing clustering quality (clustering property) [3]. Maintaining the clustering property means that strongly connected nodes are not split apart but are put into the same cluster. Also, if the clusters are formed so they have about the same size, weights can be assigned to each edge, and the sum of the edge weights can be used as a basis for partitioning.

The balanced granularity partitioning first generates a lot of small-sized clusters as an intermediate step. Those small clusters are then merged into the same partition in order to build final partitions that have the same cluster size (**Fig. 2(a)**). For example, we can apply our balanced granularity partitioning for traffic congestion prediction by using road networks. In this case, roads and their intersections are respectively represented as nodes and edges in a graph. In addition, we can also represent the amount of traffic on a road by using a weighted edge in the graphs. For example, if there is a road network in an urban area, we can expect its traffic conditions to change dramatically from time to time. Obviously, if the traffic congestion prediction and simulation system requires a large runtime, the system will not make sense to users. Hence, to predict traffic conditions in near real time, it is important to reduce the prediction and simulation runtime by using parallel computing on multiple machines and to use graph partitioning techniques that balance the workload across the machines.

The equal-granularity clustering technique enables sufficiently fast parallel processing of simulations by



(a) Example of processing to obtain six clusters of equal granularity



(b) Traffic congestion prediction using balanced granularity partitioning

Fig. 2. Partitioning graphs using equal granularity.

fast partitioning of the graph to match the number of parallel simulations, and it evens out the processing load by producing partitions of equal granularity (**Fig. 2(b)**).

We evaluated our equal-granularity clustering technique in a traffic simulator done in collaboration with

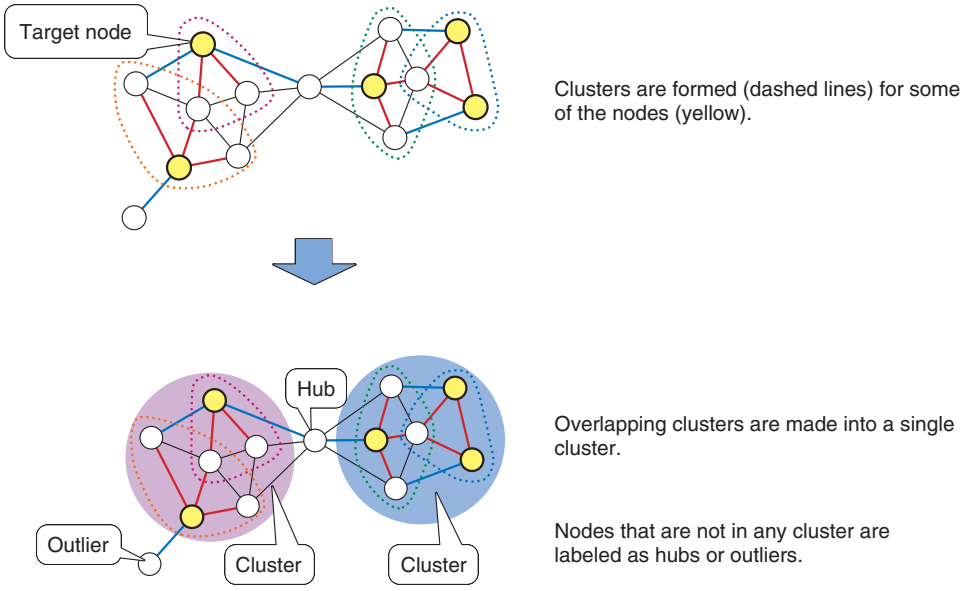


Fig. 3. Discovery of mediators.

NTT DATA Corporation, and we confirmed that our technique ran at least 10 times faster than METIS, the de-facto standard open-source graph partitioning software. Furthermore, we also experimentally verified that the equal-granularity clustering technique produced well-sized balanced partitions for large-scale graphs. NTT DATA Corporation is currently working on the application of this technique by conducting demonstration experiments for various use cases. Our technique is not limited to the above congestion prediction; we can also apply it to traffic volume visualization, traffic signal control, and other such purposes.

4. Efficient discovery of persons or things that serve as mediators

In recent years, techniques have been developed for finer graph analysis that goes beyond the discovery of groups of strongly connected nodes. One of the most important techniques is the Structural Clustering Algorithm for Networks (SCAN). SCAN is a graph clustering technique that tries to find clusters based on node similarities (i.e., structural similarities). Unlike traditional graph clustering techniques, SCAN finds not only clusters but also hubs, which are nodes that bridge multiple clusters, and outliers, which are noise components included in a graph. SCAN regards nodes that are not included in clusters as hubs or outliers. If a non-clustered node bridges multiple clusters, SCAN determines the node to be a hub; otherwise the node is an outlier. As mentioned previously, SCAN provides good clustering results; however, it requires substantial computation time to find all clusters, hubs, and outliers in large-scale graphs since SCAN has to iteratively compute all edges in the graph.

Hence, we developed a novel structural clustering algorithm that can analyze data at a practical high speed, even with large-scale graphs [4]. Specifically, our approach forms clusters of only some of the nodes rather than evaluating cluster membership for all nodes. If the clustering process produces clusters that intersect, the clusters are combined into a single cluster, and any nodes that do not belong to either cluster are regarded as either hubs or outliers (**Fig. 3**). By taking this approach, we can increase the processing speed by a factor of 20 compared to the conventional method without loss of accuracy. Discovery of clusters, hubs, and outliers at the same time takes more runtime than simple cluster analysis, but this large increase in speed makes it possible to perform the processing for a graph with hundreds of thousands of nodes in a matter of seconds.

If we apply this technique to friend relationships, for example, a person who does not belong to any group is classified as a hub. As a mediator, that person is considered to contribute to multiple groups. Hence,

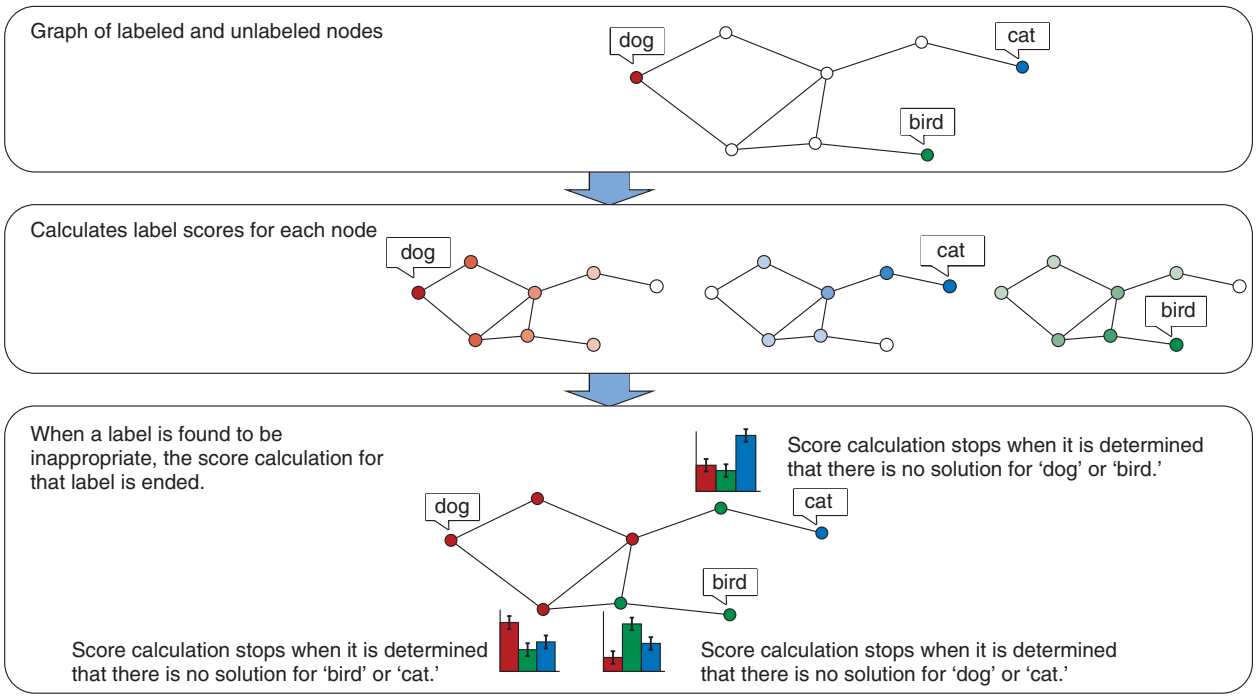


Fig. 4. Categorizing.

he/she might be expected to spread information about products to multiple groups and to serve as a route for conveying information from one group to another. In addition, since persons who are classified as hubs do not belong to any particular group, they can also be considered as candidates for work that requires a neutral position such as evaluation or investigation tasks. Another example is analyzing a graph of the transaction relationships among companies to identify companies that could serve as mediators. This would provide useful information that could be used in business decisions such as selecting companies that have influence in multiple industries for commercial cooperation. The existence of such possible applications has motivated us to continue evaluating the applicability of this technology.

5. Efficient categorizing of people and objects

One important task in data analysis is categorizing. Products can be classified according to customer interests and preferences such as health-consciousness or a preference for upscale products. News articles can be classified by subject such as politics or entertainment. One node classification technique for graph mining is label propagation. Each node in the graph is assigned label information, and nodes that have different labels are regarded as belonging to different categories.

Categorizing based on label propagation involves assigning labels to unlabeled nodes when given a small number of labeled nodes. Consider a graph that relates documents by similarity and has nodes that are labeled *dog*, *cat*, and *bird*, for example. The unlabeled nodes can be assigned labels according to their similarity to the labeled nodes. To do this, a score is calculated for each label, and the label with the highest score is applied to that node. A problem with this calculation is that the calculation may be recursive, depending on the structure of the graph, so the computational cost can be very high. We have developed a fast label propagation technique that can analyze large-scale graphs at a practical high speed [5].

Conventional label propagation involves evaluating each label to determine if it is the correct label, so the computational load is very high. Our approach is to halt the calculation of label scores when it is determined that a particular label is clearly not correct (**Fig. 4**). By doing so, we achieved an increase in speed by a factor of from 2 to 400 relative to the conventional technique. The processing for node categorization by label propagation usually takes longer

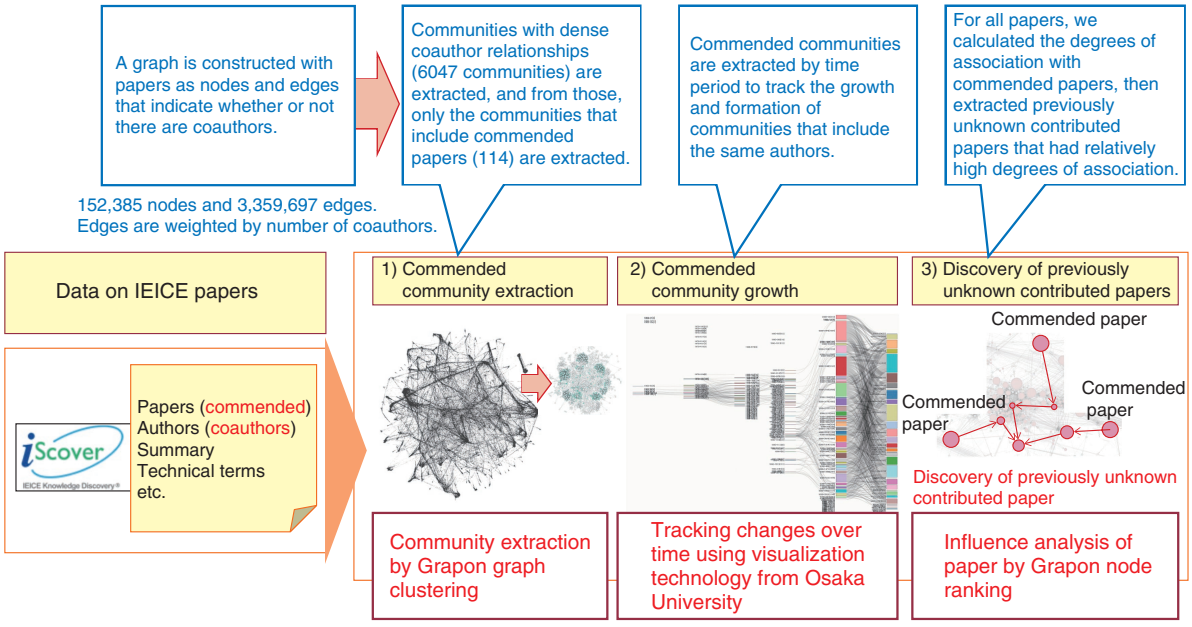


Fig. 5. Discovery of community trends and previously unknown contributed papers.

than cluster processing, but the increase in speed that we have achieved makes it possible to perform the processing for a graph with tens of thousands of nodes in just a few seconds.

Potential applications of this technique include classification of customers in a graph of purchase relationships and co-marketing of products that belong to the same category. Other possibilities include categorizing companies that are in the supply chain of a corporate group by capital grouping in order to make business decisions such as choosing companies for collaboration or accepting companies as clients or customers. We are currently evaluating the applicability of this technique.

6. Verification of the utility of the technology

We are moving forward with R&D to verify and refine the technology and to derive new research topics by applying Grapon in various fields. An important example is our participation in the data analysis competition sponsored by the Joint Association Study Group of Management Science (JASMAC), a gathering of over 600 data analysis experts from academic and corporate organizations, in which we conducted trials that involved applying Grapon in the field of marketing. Those trials confirmed the utility of Grapon for high-speed extraction of products for recommendation and discovery of product combinations that are suitable for cross-selling from large-scale POS (point of sales) data.

Another example is our participation together with Osaka University in the I-Scover Challenge 2014, which involved deriving useful knowledge from the textual data of over 160,000 research papers of the Institute of Electronics, Information and Communication Engineers (IEICE). Our results in that event confirmed that Grapon was effective for recognizing trends in communities that produce commended papers from the co-authorship relationships among several tens of thousands of authors and for discovering previously unknown contributed papers, and that it analyzed such large-scale data at high speed (**Fig. 5**).

7. Toward analysis of data on larger scale

We are witnessing the ever-growing scale of graph-structured data; for example, the total number of web pages has reached a scale of billions of nodes. To deal with this situation, we have begun studies on parallel processing techniques for efficient graph mining on multiple CPUs (central processing units) rather than on a single powerful CPU. For example, data layout optimization by reassigning node identification numbers promotes the use of data within the CPU cache memory. This optimization allows us to obtain

maximum benefit from parallel processing by reducing the amount of communication between CPUs as well as achieving faster operation of single CPUs. Although there are open source tools for parallel graph mining, for example, GraphLab, Grapon achieves 60 times better performance than GraphLab by using the data layout optimization [6, 7]. We will continue our R&D on more scalable parallel processing on multiple CPUs to cope with the increasing graph scale.

8. Future developments

We have briefly described our R&D on graph mining technology to achieve more efficient and more complex analysis of data that continues to increase in scale. In the future, we will continue to refine the graph mining techniques and increase their ease-of-use by applying our technology in various fields with the objectives of creating new value and contributing to a smart society.

References

[1] NTT press release issued on February 13, 2013 (in Japanese). http://www.ntt.co.jp/news2013/1302/130213b.html
[2] Y. Iida, Y. Kishimoto, Y. Fujiwara, H. Shiokawa, and M. Onizuka, "Finding Communities and Ranking in Large-scale Graphs: Fast Algorithms and Applications," Journal of JSAI, Vol. 29, No. 5, pp. 472–479, 2014.
[3] T. Fujimori, H. Shiokawa, and M. Onizuka, "Optimization of Graph Partitioning for Distributed Graph Processing," Proc. of DEIM Forum 2015 (the 7th Forum on Data Engineering and Information Management), E5-2, Fukushima, Japan, 2015.
[4] H. Shiokawa, Y. Fujiwara, and M. Onizuka, "SCAN++: Efficient Algorithm for Finding Clusters, Hubs and Outliers on Large-scale Graphs," The Proceedings of the VLDB Endowment (PVLDB), Vol. 8, No. 11, pp. 1178–1189, 2015.
[5] Y. Fujiwara and G. Irie, "Efficient Label Propagation," Proc. of ICML 2014 (the 31st International Conference on Machine Learning), pp. 784–792, Beijing, China, 2014.
[6] J. Arai, H. Shiokawa, T. Yamamuro, and M. Onizuki, "Scalable Parallel Graph Processing by Optimized Vertex Order," DEIM Forum 2015, E5-3, Fukushima, Japan, 2015.
[7] J. Arai, H. Shiokawa, T. Yamamuro, M. Onizuki, and S. Iwamura, "Rabbit Order: Just-in-time Parallel Reordering for Fast Graph Analysis," Proc. of IPDPS 2016 (the 30th IEEE International Parallel and Distributed Processing Symposium), Chicago, USA, to appear.

Yasuhiro Iida

Senior Research Engineer, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.S. and M.S. in applied physics engineering from the University of Tokyo in 1998 and 2000. In 2000, he joined NTT Information Sharing Platform Laboratories. His recent research area is data mining and data management. He is a member of the Association for Computing Machinery (ACM).

Yasunari Kishimoto

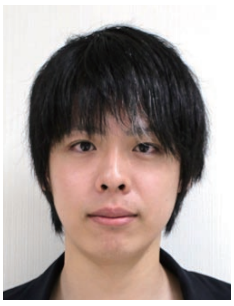
Research Engineer, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.E. and M.E. from Kyushu University, Fukuoka, in 1989 and 1991. He joined NTT in 1991 and studied directory systems, billing systems, and data mining. He is a member of the Information Processing Society of Japan (IPSJ).

Yasuhiro Fujiwara

Distinguished Technical Member, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.E. and M.E. from Waseda University, Tokyo, in 2001 and 2003, and a Ph.D. from the University of Tokyo in 2012. He joined NTT in 2003. His research interests include data mining, databases, natural language processing, and artificial intelligence. He is a member of IPSJ, IEICE, and the Database Society of Japan (DBSJ).



Hiroaki Shiokawa

Researcher, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.S. in information science and an M.E. and Ph.D. in engineering from University of Tsukuba in 2009, 2011, and 2015. He joined NTT in 2011 and has been studying graph data management, graph mining algorithms, distributed computing, and databases. He is a professional member of ACM.

Junya Arai

Researcher, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.S. and a Master of Information Science and Technology degree from the University of Tokyo in 2011 and 2013. He joined NTT in 2013 and has been studying parallel distributed processing and graph mining algorithms. He is a member of ACM and DBSJ.

Sotetsu Iwamura

Senior Research Engineer, Supervisor, Distributed Computing Technology Project, NTT Software Innovation Center.

He received a B.S., M.S., and Ph.D. in electronic engineering from the University of Tokyo in 1989, 1991, and 1994. He joined NTT Telecommunication Networks Laboratories in 1994. Since then, he has been researching mobile computing and high-speed computer networks based on asynchronous transfer mode (ATM), mobile computing, as well as an in-house cloud computing platform for NTT R&D. He is currently managing a research group on big data processing, such as fast data-mining algorithm, parallel-distributed processing architecture. He is a member of IPSJ.