ISSN 2186-7437

NII Shonan Meeting Report

No. 2018-10

Graph Database Systems: Bridging Theory, Practice, and Engineering

George Fletcher
Wook-Shin Han
Oskar van Rest

July 30 - August 2, 2018

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Graph Database Systems: Bridging Theory, Practice, and Engineering

Organizers:
George Fletcher (TU Eindhoven, the Netherlands)
Wook-Shin Han (Pohang University of Science and Technology, South Korea)
Oskar van Rest (Oracle, USA)

July 30 - August 2, 2018

1 Introduction

Recent years have seen great advances in the study of data management solutions for massive graph-structured data sets. This has been stimulated by the increasing availability of large graphs in a broad variety of application domains such as social networks, biological networks, linked open data, communications networks, and mobility networks. Consequently, there has been a marked rise in demand for scalable solutions for the principled management of graph data. Rapid progress has been made on our understanding of: the theoretical foundations of fundamental topics such as graph query languages, graph analytics, and graph modeling; the engineering foundations of efficient and scalable graph intensive systems; and, the practical application and engineering of graph data management solutions in industry.

Through these advancements the graph database research community has now reached a first stage of maturity. However, this understanding and acquired wisdom is distributed across various disparate subcommunities in the field. The time is right for a community "checkpoint", to share these experiences and insights across the rich and diverse areas of investigation in graph database systems. Indeed, a major outcome of this necessary checkpoint will be to consolidate our broad community understanding of the "first generation" of practical graph data management systems.

A second outcome of this taking stock and intense sharing of perspectives is to identify the major challenges and limiting factors in the realization of the next generation of graph database systems. Examples of such open challenges include: identifying appropriate graph schema formalisms and developing deeper our understanding of graph constraints (both in theory and practice); the efficient processing of recursive graph queries, such as the Regular Path Queries; developing practical syntaxes and engineering solutions for graph query languages supporting reasoning over data, e.g., in the property-graph model; practical human-in-the loop graph analytics and visual query methodologies; and, ensuring that we are asking the right questions driven by application domains and practical graph analysis.

1.1 Goals of the Meeting

The goal of this meeting was to take stock of the current state of the art in graph data management systems and to identify major open research challenges and directions, towards setting a community research agenda for the coming years. We placed particular focus on building bridges between advances in the theory, engineering, and practical deployment of graph databases. For this broad discussion, we brought together relevant leading researchers from both academia and industry, across these diverse subcommunities.

1.2 Workshop organization and outcomes

We aim to write a community white paper for peer-reviewed publication, indicating the most important open challenges in the field identified during the meeting. Having an open bottom-up shared vision for the field will stimulate research and industry progress, setting the research agenda for the graph data

management community for the coming years. The workshop also led to concrete action plans for international collaborations in research and longer-term international projects of broad ambition, including identifying concrete calls for proposals for research funding, to facilitate international collaboration.

2 List of Participants

- Marcelo Arenas, Pontifical Catholic University of Chile
- Peter Boncz, CWI
- Angela Bonifati, Lyon 1 University
- Hassan Chafi, Oracle Labs
- Hirokazu Chiba, Database Center for Life Sciences
- Laurent Daynes, Oracle
- George Fletcher, TU Eindhoven
- Claudio Gutierrez, University of Chile
- Wook-Shin Han, Pohang University of Science and Technology
- Olaf Hartig, Linköping University
- Jan Hidders, Vrije Universiteit Brussel
- Romans Kasperovics, SAP SE
- Arijit Khan, Nanyang Technological University
- Jaehun Lee, Samsung Electronics
- Young-Koo Lee, Kyung Hee University
- Tobias Lindaaker, Neo4j
- Jean-Pierre Lozi, Oracle Labs
- Wim Martens, University of Bayreuth
- Makoto Onizuka, Osaka University
- Stefan Plantikow, Neo4j
- Oskar van Rest, Oracle
- Semih Salihoglu, University of Waterloo
- Jiwon Seo, Hanyang University
- Juan Sequeda, Capsenta
- Hannes Voigt, Neo4j

- Peter Wood, Birkbeck, University of London
- Zhe Wu, Oracle
- Nikolay Yakovets, TU Eindhoven
- Ryota Yamanaka, Oracle
- Yuichiro Yasui, Nikkei Business Publications
- Yongluan Zhou, University of Copenhagen

3 Overview of Talks

3.1 Viewpoint Talks

Making Graph Database Systems Competent

Peter Boncz, CWI

Despite the fact that graph database functionality is ever better understood and eg powerful query languages have been defined, there is still a problem. Namely, current graph database systems are not competent. Where I define competent as reliable, performant and scalable. In this talk I describe my view on this sorry state of affairs and look for causes and solutions.

Graph Databases: Is it a feature or an Index?

Hassan Chafi, Oracle Labs

During our talk, we will explore the evolution of Oracle Labs thinking on Graph Analytics and Databases. In particular, we question the future of Graph "Only" databases as traditional multi-model databases absorb the graph model from one side and vertical applications embed graph techniques from another side. Is there a large future for so called Graph Native (or Only) databases?

The R versus NR data problem

Claudio Gutierrez, University of Chile

Relational data is mathematically equivalent to graph data. The main challenge of graph data management is to show that it has an intrinsic niche different from the successful relational model. I will argue that this niche is not a better and simple representation of certain features of data (like paths, etc.), but two intrinsic characteristics of massive data of the era opened by the Web: openness and incompleteness.

Beware, Graph Swamp Ahead!

Juan Sequeda, Capsenta

One of the latest hypes around graphs is to create Enterprise Knowledge Graphs, a fancy way of saying data integration. Similar to evolution from "data lake" to "data swamp", if we are not careful, we are doomed to create Graph Swamps: naive transformation of source databases into graphs without an understanding of a schema. I will argue that we need to understand the socio-technical aspects of data integration with graphs with an important focus on: mappings from relational data to graph and graph schemas.

Are graph databases following the same path that object-oriented or xml databases have followed?

Wook-Shin Han, Pohang University of Science and Technology

In this talk, I review the claims of object-oriented and XML databases and analyze their problems. Their major claims have not been substantiated successfully, and some of the claims are now re-echoed by graph databases. Finally, I conclude my talk with very simple questions. 1) Can graph databases be used with minimal efforts for applications which currently use RDBs? 2) What are killer applications?; Are they ubiquitous?; and who needs fast graph traversal?

Worst Case Guidance – Role of Query Complexity Analysis in Query Language Design

Hannes Voigt, Neo4j

G-CORE and other QL Proposals made a strong point about guaranteeing tractable evaluation complexity for all its features. We should not forget that asymptotic complexity analysis is a deliberate generalization of the worst case. Should the abstract worst case be our main guidance for a practical language? Let's remind ourselves want it means to transfer theory knowledge back to practice and which questions shall arise in that process. This talk provides a primer.

3.2 Regular Talks

From Theory to Practice in Subgraph Query Evaluation

Semih Salihoglu, University of Waterloo

Recently, a new class of join algorithms that evaluate queries by one attribute at a time, instead of traditional table(s) at a time. Perhaps the most important application of these algorithms are in the context of complex, cyclic subgraph queries in graph processing, where they correspond to vertex at a time matching of queries. Theoretically, these algorithms have been shown to be worst-case optimal (WCO), yet the existing theory fails to give further advice on how to pick which vertices to pick to match next. We have been studying these algorithms in different contexts, serial and distributed, and for both one-time and continuous

subgraph queries. I will present both some theoretical and performance results we obtained about the behavior of these algorithms, focusing on the distributed setting.

Towards GQL and beyond - Consolidation, trends, and future opportunities for graph querying

Stefan Plantikow, Neo4j

This talk will report on recent developments around the standardization of a next generation property graph query language (GQL) and then present currently discussed and potential future features of such a language, as well as open research questions.

An Analytical Study of Large SPARQL Query Logs

Wim Martens, University of Bayreuth

I will present an overview of features of queries in SPARQL query logs. The logs are from 2009 - 2017 and come from a broad range of sources (biological, geographical, museum, semantic web). Our analysis is not limited to the use of keywords or the number of triples per query, but includes more investigations, like a classification of graph patterns according to their structure, and a structural analysis of property paths. If time permits, I can also talk about connections between theory and practice.

Parallel SPARQL Query Processing and Optimization

Yongluan Zhou, University of Copenhagen

Existing parallel SPARQL query optimizers assume hash-based data partitioning and adopt plan enumeration algorithms with unnecessarily high complexity. Therefore, they cannot easily accommodate other partitioning methods and only consider an unnecessarily limited plan space. To address these problems, we first define a generic RDF data partitioning model to capture the common structure of various state-of-the-art RDF data partitioning methods. Then we propose a query plan enumeration algorithm that not only has an optimal efficiency, but also accommodates different data partitioning methods. Furthermore, based on a solid analysis of the complexity of the plan enumeration algorithm, we propose two new heuristic methods that can consider a much larger plan space than the existing methods, and at the same time can still confine the search space of the algorithm. An autonomous approach is proposed to choose one of the two methods by considering the structure and the size of a complex SPARQL query. We conduct extensive experiments using synthetic and a real-world dataset, which show the superiority of our algorithms in comparing to existing ones.

Certified Graph View Maintenance with Regular Datalog

Angela Bonifati, Lyon 1 University

My talk revolves around the design and implementation of a regular graph query engine, which is correct-by-construction by relying on interactive theorem proving techniques in Coq. The designed query engine is robust under updates and implements incremental graph view maintenance algorithms seen in the database literature.

ORM2: the best graph data model you never heard about

Jan Hidders, Vrije Universiteit Brussel

ORM (aka NIAM) is a data model that was originally introduced as an improved ER dialect with a formal semantics based on first order logic and a philosophical justification based on business-oriented communication. It has a long history of successful application in practice, as well as a tradition of academic papers of several decades, which makes it somewhat unique. Although never described as such by the ORM community, the semantics of ORM make it effectively a graph-oriented data model, and as such it can be an interesting source of inspiration for those interested in defining a standard schema definition language for graph databases, especially for those who would like it to resonate with business-oriented users.

Latest Challenges in Property Graph Language Design

Oskar van Rest, Oracle

The talk is split into two parts: the first part is a short introduction to existing property graph technologies at Oracle, while the second part presents some of our latest challenges regarding graph language design, namely, (1) graph views, (2) integration of procedural language with query language, and (3) handling structured vs. unstructured data.

Flexible querying of graph data

Peter Wood, Birkbeck, University of London

When users receive fewer answers than expected to a query, a flexible querying system can automatically rewrite the query into a set of new queries, using query approximation and relaxation methods. Each such rewritten query has a cost associated with it, and answers are returned to users in increasing order of this cost. The rewritings offer many opportunities for query optimisation.

Data Management for Emerging Problems in Large Networks

Arijit Khan, Nanyang Technological University

Application of networks and data management techniques for user-friendly,

approximate, and scalable querying, mining, and analytics of knowledge graphs, social networks, road networks, biological networks, and program flow graphs.

Real World Experience of using Graphs and Semantics for Enterprise Data Integration

Juan Sequeda, Capsenta

An early vision in Computer Science has been to create intelligent systems capable of reasoning on large amounts of data. Today, this vision can be delivered by integrating Relational Databases with the Semantic Web using the W3C standards: a graph data model (RDF), ontology language (OWL), mapping language (R2RML) and query language (SPARQL). The research community has successfully been showing how intelligent systems can be created with Semantic Web technologies, dubbed now as Knowledge Graphs. However, where is the mainstream industry adoption? What are the barriers to adoption? Are these engineering and social barriers or are they open scientific problems that need to be addressed? This talk will chronicle our journey of deploying Semantic Web technologies with real world users to address Business Intelligence and Data Integration needs, describe technical and social obstacles that are present in large organizations, and scientific and engineering challenges that require attention. Additionally, the talk will introduce the methodology used to to design ontologies and mappings for real world large scale database schemas.

Experiences with implementing G-CORE

Peter Boncz, CWI

In this talk I will report on work at CWI on implementing the G-CORE compositional graph query language defined by LDBC on Spark.

Learning from history of Graph Models and Query Languages

Tobias Lindaaker, Neo4j

Through history there has been a few different models and languages implementing what we today talk about as graph data processing. What can we learn from their successes and failures? Looking forward to the needs of data processing, what things can we anticipate today that would benefit changing in the field of graph data processing in order to meet those needs? This talk is inspired by conversations with other researchers and material others have published on the topic but aren't able to present themselves.

Fundamental Properties of the GraphQL Language

Olaf Hartig, Linköping University

GraphQL is a recently proposed, and increasingly adopted, conceptual framework for providing a new type of data access interface on the Web. The framework includes a new graph query language whose semantics has been specified

informally only. We have embarked on the formalization and the study of this language. In the talk I will provide an overview of the results of our work.

G2GML: Mapping from Semantic Graph to Property Graph

Ryota Yamanaka, Oracle& Hirokazu Chiba, Database Center for Life Sciences

In the life science domain, open data about genes, proteins, and pathways are increasingly available in RDF graphs. RDF structure, however, is not necessarily appropriate for graph representation/analysis using graph algorithms in an intuitive and effective way. Here we discuss a framework for mapping RDF to property graph, which will contribute to enhance usability of linked open data in graph databases.

Graph Abstraction

Hannes Voigt, Neo4j

The talk (1) points out the importance of composable query language capabilities for abstracting base data (e.g. twitter messages) into high-order concepts (e.g. conversations). Focusing on graph query languages, the talk (2) briefly reiterates graph construction and (3) aggregative graph construction as a foundation for graph abstraction and (4) extends this to graph summarization.

3.3 Poster and demo presentations

Participants gave poster and demo presentations during a plenary interactive session on the afternoon of Day 2.

4 Meeting Schedule

Day 0 (Sunday 29 July)
Check-in from 15:00. Welcome reception at 19:00.

Day 1 (full day)
The first day was dedicated to exchanges between participants in the form of presentations and free-form discussion with the goal of sharing results and establishing key objectives for the meeting.

Our preliminary proposal for the seminar goal was to write a community vision paper indicating the main open challenges in graph database systems (e.g., for submission to SIGMOD Record as a "workshop report" or "vision" paper).

Morning sessions of Day 1 were dedicated to kick-off talks representing a mix of academic and industrial viewpoints. The afternoon sessions were dedicated to research talks and discussion. We closed before dinner with a short session for discussion on planning and goals for the workshop.

- 09:00 – Introduction movie of NII Shonan meeting
- 09:10 – Workshop introduction by organizers
- 09:20 – Session 1: Viewpoint presentations
  - Peter Boncz, Hassan Chafi, Claudio Gutierrez, Juan Sequeda, Wook-Shin Han, Hannes Voigt
- 10:30 – Break
- 11:00 – Session 2: Individual talks (20 minutes each, including Q&A)
  - Semih Salihoglu, Stefan Plantikow, Wim Martens
- 12:00 – Lunch
- 13:30 – Group Photo shoot
- 14:00 – Session 3: Individual talks (20 minutes each, including Q&A)
  - Yongluan Zhou, Angela Bonifati, Jan Hidders
- 15:30 – Break
- 16:00 – Session 4: Individual talks (20 minutes each, including Q&A)
  - Oskar van Rest, Peter Wood, Arijit Khan, Juan Sequeda
- 17:30 – Planning discussion
- 18:00 – Break
- 18:30 – Dinner

Day 2 (full day)
The morning of the second day was dedicated to presentations by participants. The second half of the day we kicked off discussions and writing of the community challenges paper.

- 09:00 – Session 1: Individual talks (20 minutes each, including Q&A)
  - Peter Boncz, Tobias Lindaaker, Olaf Hartig, Ryota Yamanaka+Hirokazu Chiba, Hannes Voigt
- 10:40 – Session 2: Discussion of community paper
- 12:10 – Lunch
- 13:30 – Session 3: Poster and Demo presentations
- 15:00 – Break
- 16:00 – Session 4: Discussion of community paper, form writing groups
- 17:30 – Break
- 18:00 – Dinner

Day 3 (half day – group excursion in the afternoon)
The first morning session of the third day was dedicated to writing sessions. Then before lunch, we had a plenary discussion of progress on the paper. In the afternoon and evening we had the group excursion.

- 09:00 – Session 1: Writing session, in working groups
- 10:30 – Session 2: Plenary discussion of progress on community paper, regroup as necessary
- 12:30 – Lunch
- 13:25 – Meet in Lobby, for excursion (to Engakuji and Kenchoji Temples, in Kamakura)
- 18:15 – Dinner (off-site)

Day 4 (half day – seminar ends with lunch)
The last day we finalized the writing and the roadmap for concrete post-workshop action items. The goal was to achieve a more formal output, e.g., consolidate and extend the writing of the previous days to bring it to a form close to a final publication report. We also held a wrap-up discussion to identify post-workshop actionable items between participants (e.g., research directions, collaboration, project proposals).

- 09:00 – Session 1: Planning discussion, plenary
- 10:00 – Break
- 10:30 – Session 2: Plenary wrap-up discussions
- 12:00 – Lunch