Graph Data Management Systems in New Applications Domains

Mikko Halin
Introduction

● Presentation is based on two papers
  ○ Graph Data Management Systems for New Application Domains - Philippe Cudré-Mauroux, Sameh Elnikety
  ○ Graph data management and mining: A survey of algorithms and applications - Aggarwal, Charu C., and Haixun Wang

● What is graph data management and what kind of problems can be solved with graphs?
Introduction

- Graphs and early applications
- Graph data management
- Graphs in new application domains
- Conclusions
- Own thoughts
Graphs

- A graph $G$ is a non-empty finite set of vertices $V$ which can be represented as nodes or points and edges $E$ (arcs or lines) that connect the vertices together.
- Graphs can be used in various applications where information that has some kind of relationship is being analyzed, optimized or understood.
Early applications: Bridges in Königsberg

- Problem: Is it possible to walk across every bridge only once?
- In 1736 Leonhard Euler transformed the map to a graph
  - If a graph has an even number of vertices and only the start and endpoint of the path have an odd number of edges it is possible to draw a path where every edge was crossed only once
Other early applications

Travelling salesman problem

- The problem: Find the shortest route between a number of cities.
  - Trying every possible route combination is very time consuming if there are many nodes.
- Fastest route could look entirely different depending on the weights of the edges.
  - Weights could be determined based on distance, traffic and for example information about road construction sites which location could change in different points of time.

Four colour theorem

- Is it possible to colour a map of countries with different colours so that no two adjacent countries have the same colour?
- In 1852 Francis Guthrie and August De Morgan assumed that four colours would be enough for any map.
- In 1976 Kenneth Appel and Wolfgang Haken used graph theory and computers to verify that the four colour theorem was true.
Graph Data Management

- Graph data management has been studied a long time by database researchers.
- New application domains have caused renewed interest in graph data management. For example, Social Networks and Web of Data.
- Graphs in modern applications are too large and dynamic for legacy graph data management applications to handle.
Graph Data Management

● What makes Big Data problems Graph Data Management problems?
  ○ Big Data analysis problems have typically three characteristics: Volume, variety and velocity
  ○ Finding relationships between different data items might be needed to acquire information from Big Data

● New application domains in graph data management have Big Data related problems
  ○ Disk-resident data sets need a good algorithmic design where nodes and edges of the graph are not accessed randomly
  ○ Massive graphs require summarization of the underlying graphs in order to process the data
Graph Data Management

- **Graph Management Algorithms**
  - Algorithms for managing and indexing large volumes of the graph data.

- **Graph Mining Algorithms**
  - Algorithms used to extract patterns, trends, classes, and clusters from graphs.

- **Applications of Graph Data Management and Mining**
  - Various application domains in which graph data management and mining algorithms are required.
Graph Databases

- It is possible to represent and store a graph in relational table
  - Relational database systems and the SQL language are not designed to handle connection patterns between data
- Graph Database systems provide specific features to handle graphs
  - For example breadth-first search, shortest path algorithms and insert, delete and modify data in a graph
Graph Databases

- Storage mechanism
  - Some use native graph storage which is optimized for storing and managing graphs. Others use relational or object-oriented databases, or some other general purpose data store

- Processing engine
  - Index-free adjacency - Connected nodes physically “point” to each other in the database (Native graph processing)
  - Processing is done some other way where the database behaves like a graph database from the user's perspective

ASCII art representation in Cypher: (emil)<-[::KNOWS]-(jim)-[::KNOWS]-(ian)-[::KNOWS]-(emil)

MATCH (a:Person)-[::KNOWS]->(b)-[::KNOWS]->(c), (a)-[::KNOWS]->(c)
WHERE a.name = 'Jim'
RETURN b, c
Graph Databases

Overview of graph databases
Graph Databases

- Properties added to the nodes and edges make it possible to perform queries that would be more complex in traditional databases
- Typical queries in graph databases
  - Finding the shortest path in a network
  - Finding minimum spanning tree
  - Finding graph planarity
  - Finding adjacency matrices
  - Finding connectedness
  - Finding the cycles in a graph
  - Searching an element in a data structure
    - Depth first search
    - Breadth first search
Graphs in New Application Domains

- Graphs are an intuitive way to present connected data.
- Examples of different domains where graphs are applied:
  - Social networking
  - Computer networking
  - Biological data
  - Chemical data
  - Software bug localization
  - Fraud detection
  - Route planning and navigation (Airline, logistics etc..)
  - Disease spreading prediction
  - Cyber security
Graphs in New Application Domains

Social Networks

- A social graph is built based on the information users provide
- Information is linked to each other and it has multiple properties
  - Nodes can represent people, media content or some other data
  - Edges represent relationships between the nodes for example friendship and tagging information
- Typical social network queries
  - Updates on individual nodes (e.g., to change the status of a given user)
  - Edge inserts (e.g., to connect users to new photos or webpages)
  - Shortest-path queries (e.g., to find how two users are related)
  - Reachability queries (e.g., to find all employees reporting directly or indirectly to the CTO)
Graphs in New Application Domains

Web of data

- Various applications provide data online through API:s
- If the data was created, transmitted and archived in a structured format it would be possible to create links between the pieces of data
- Linked Data Movement has made three rules for publishing data online so that connections can be made between them
  - For each resource published a Unique Resource Identifier (URI) is created
  - Publication of structured data at the URI corresponding to the resource
  - Links are included which relate each resource to similar online resource
Graphs in New Application Domains

Web of data

- Governmental agencies, companies and scientists are publishing Linked Data and users are bulk-exporting, querying and integrating data online
- Typical Web of Data queries
  - Large bulk inserts
  - Triple pattern look-ups (e.g., to find a particular node based on constraints on its direct neighbors and on its edges).
  - Distributed entity retrieval queries (e.g., to retrieve information from interlinked entities stored in separate databases).
  - Molecule queries (e.g., to retrieve all the nodes and edges directly or indirectly related to a given node).
  - Path queries (e.g., to identify all pairs of nodes connected by a given series of edges).
  - Inference queries (e.g., to retrieve nodes based on the logical properties defined by their edges).
Graphs in New Application Domains

● Graphs in different domains have different operations
  ○ Some operations in social networks and web of data are similar (scope queries and molecule queries)
  ○ Most operations are different in the two domains

● Social networks have punctual updates, transitive closures and betweenness queries

● Web of Data have edge lookups, joins and logical inference based on edges
Future Research

● Large-scale graphs - Recently research focus has mainly been in small and memory-resident graphs. Large-scale graphs are still hard to handle.
● Massive graph streams - New applications like social networks require new techniques to handle the large data streams. Data summarization and other methods need to be developed to analyze the data
Conclusions

- Graphs are a good way to present connected data
- Graphs have better performance when handling connected data than traditional databases
- Big Data problems are present and a research topic in graph data management
Graph-based Approach for Network Identity Modeling

- Natural persons and inorganic objects can be identified with a set of attributes in the digital world.
- Depending on the context it may be that only a single attribute might build the identity.
- Before identification the actors are called entities.
Graph-based Approach for Network Identity Modeling

- Network Identity Modeling using Graphs is a technique where data which is defined as relevant is added to a graph data-structure and relationships between the pieces of data are created.
- The focus is to create identity graphs within the context of information security.
- The techniques and models could be applied also in various scenarios.