

Graph Cube: On Warehousing and OLAP Multidimensional Networks

Peixiang Zhao[†], Xiaolei Li[‡], Dong Xin[§], Jiawei Han[†]

[†]Department of Computer Science, UIUC

[‡]Groupon Inc.

[§]Google Cooperation

[†]pzhao4@illinois.edu, hanj@cs.illinois.edu

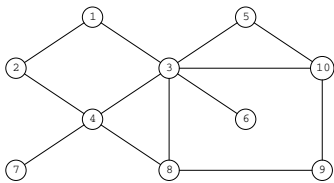
[‡]me@xiaolei.org, [§]dongxin@gmail.com

June 16th, 2011

- 1 Introduction
- 2 The Graph Cube Model
- 3 OLAP on Graph Cube
 - Cuboid Query
 - Crossboid Query
- 4 Implementing Graph Cube
- 5 Experiment
- 6 Conclusion

- Recent years have seen an astounding growth of networks in a wide spectrum of application domains
 - Communication networks
 - Social networks
 - Biological networks
 - The Web
- **Multidimensional networks**
 - ① An underlying **graph structure** comprising entities and relationships
 - ② **Multidimensional attributes** are specified and associated with entities of the network
- There exist considerable technology gaps in managing, querying and summarizing multidimensional networks effectively

A Sample Multidimensional Network



(a) Graph

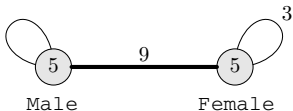
ID	Gender	Location	Profession	Income
1	Male	CA	Teacher	\$70,000
2	Female	WA	Teacher	\$65,000
3	Female	CA	Engineer	\$80,000
4	Female	NY	Teacher	\$90,000
5	Male	IL	Lawyer	\$80,000
6	Female	WA	Teacher	\$90,000
7	Male	NY	Lawyer	\$100,000
8	Male	IL	Engineer	\$75,000
9	Female	CA	Lawyer	\$120,000
10	Male	IL	Engineer	\$95,000

(b) Vertex Attribute Table

Figure: A Multidimensional Network Comprising a Graph Structure and a Multidimensional Vertex Attribute Table

- **Motivation:** *Can we extend decision support facilities on multidimensional networks?*
 - Data warehouses and OLAP are **advantageous** in the multidimensional network scenario
 - **Summarizing the massive networks** into different levels of granularity for more effective analysis and exploration
 - **Business Intelligence:** in Facebook and Twitter, advertisers and marketers take advantage of social networks within different multidimensional spaces to better promote their products via **social targeting** or **viral marketing**
- However, in multidimensional networks, much of the valuation and interest lies in the **network** itself!
 - Simple numeric value based group-by's in traditional data warehouses are no longer insightful and of limited usage, because the structural information of the networks is simply ignored

Network Aggregation v.s. Traditional Group-by

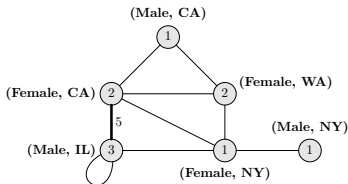


(a) Aggregate Network

Gender	COUNT(*)
Male	5
Female	5

(b) Aggregate Table

Figure: Multidimensional Network Aggregation v.s. Traditional RDB Aggregation (Group by Gender)



(a) Aggregate Network

Gender	Location	COUNT(*)
Male	CA	1
Female	CA	2
Female	WA	2
Male	IL	3
Male	NY	1
Female	NY	1

(b) Aggregate Table

Figure: Multidimensional Network Aggregation v.s. Traditional RDB Aggregation (Group by Gender and Location)

- **Graph Cube**

- A multidimensional network can be summarized to **aggregate networks** in coarser levels of granularity within different multidimensional spaces
 - Vertex coalescence
 - Structure summarization
- Different query models and OLAP solutions are proposed for multidimensional networks
 - Cuboid Queries
 - Crossboid Queries
- Efficient implementation is based on a combination of
 - Well-studied data cube implementation techniques
 - Special characteristics of multidimensional networks
- The **first** to systematically address warehousing and OLAP issues on large multidimensional networks

Multidimensional Network

A multidimensional network, \mathcal{N} , is a graph denoted as $\mathcal{N} = (V, E, A)$, where V is a set of vertices, $E \subseteq V \times V$ is a set of edges and $A = \{A_1, A_2, \dots, A_n\}$ is a set of n vertex-specific attributes, i.e., $\forall u \in V$, there is a tuple $A(u)$ of u , denoted as $A(u) = (A_1(u), A_2(u), \dots, A_n(u))$, where $A_i(u)$ is the value of u on i -th attribute, $1 \leq i \leq n$. A is called the **dimensions** of the network \mathcal{N} .

- Some (or all) dimension A_i could be * (ALL), representing a super-aggregation along A_i
- Given a set of n dimensions of a network, there exist 2^n multidimensional spaces (aggregations)
- The **measure** within each possible space is no longer a simple numeric value, but an **aggregate network**

The Graph Cube Model

Graph Cube

Given a multidimensional network $\mathcal{N} = (V, E, A)$, the graph cube is obtained by restructuring \mathcal{N} in all possible aggregations of A . For each possible aggregation A' of A , the grouping measure is an **aggregate network** G' w.r.t. A' .

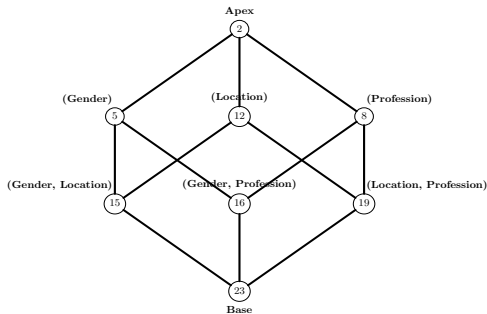
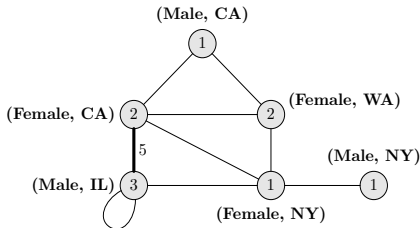
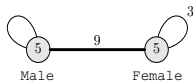


Figure: The Graph Cube Lattice

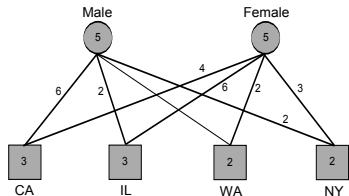
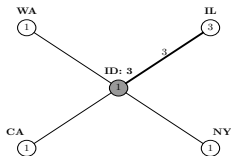
OLAP on Graph Cubes

- **Cuboid Query:** return as output the aggregate network corresponding to a specific aggregation of the dimensions of the multidimensional network
 - *What is the network structure between various **genders**?*
 - *What is the network structure between the various **gender and location** combinations?*

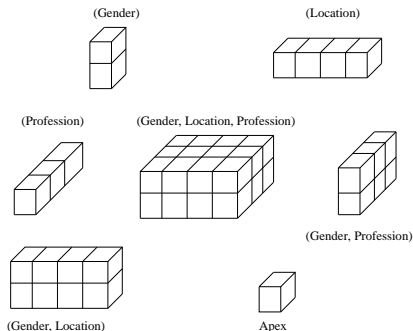


OLAP on Graph Cubes

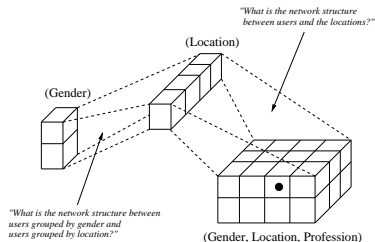
- A cuboid query is within a single multidimensional space, which follows the traditional OLAP model
- A **crossboid query** crosses multiple multidimensional spaces of the network, i.e., more than one cuboid is involved in a query
 - *What is the network structure between the user with ID = 3 and various locations?*
 - *What is the network structure between users grouped by **gender** v.s. users grouped by **location**?*



Cuboid Queries v.s. Crossboid Queries



(a) Traditional Cuboid Queries



(b) Crossboid Queries Straddling Multiple Cuboids

- **Objective:** compute the aggregate networks of different cuboids grouping on all possible dimension combinations of a multidimensional network
 - ① **Full materialization:** Best query response time, worst space cost
 - ② **No materialization:** Best space cost, worst query response time
 - ③ **Partial materialization:** A small portion of cuboids is materialized in order to balance the tradeoff between query response time and cube resource requirement

Graph Cube Implementation: Partial Materialization

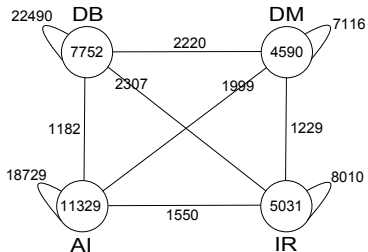
- **Problem:** To select a set S of k cuboids in the graph cube for materialization, such that the average time taken to evaluate the queries can be minimized
 - The partial materialization problem is NP-complete, reduced from *set-cover*
- **Greedy Algorithm:** Selecting k cuboids with the highest size-reduction benefit

Theorem

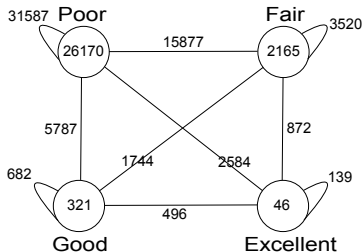
Let B_{greedy} be the benefit of k cuboids chosen by the greedy algorithm and let B_{opt} be the benefit of any optimal set of k cuboids. Then $B_{greedy} \leq (1 - 1/e) \times B_{opt}$ and this bound is tight

- **MinLevel Algorithm:** Materializing cuboids c , where $dim(c) = l_0$ indicating the level in the cube lattice at which we start materializing cuboids

- DBLP data set
 - A co-authorship graph with 28,702 authors as vertices and 66,832 coauthor relationships as edges
 - **Three dimensions:** name, area, productivity
 - area: DB, DM, AI, IR
 - productivity: Excellent, Good, Fair, Poor
- IMDB data set
 - A movie rating network with 116,164 vertices and 5,452,350 edges
 - **Seven dimensions:** Title, Year, Length, Budget, Rating, MPAA and Type
 - MPAA: G, PG, PG-13, R, NC-17, NR
 - Type: action, animation, comedy, drama, documentary, romance, short



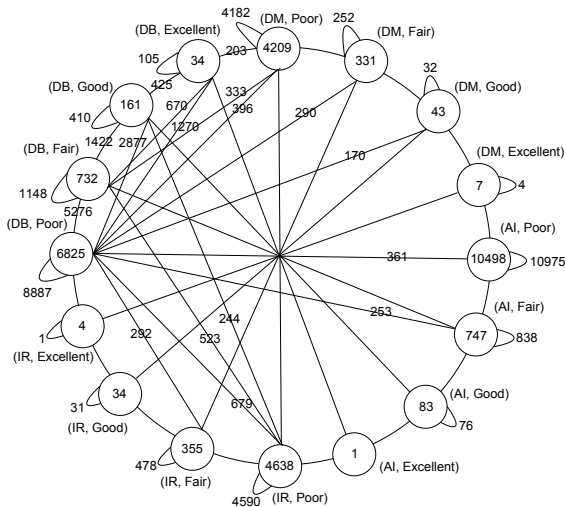
(c) (Area)



(d) (Productivity)

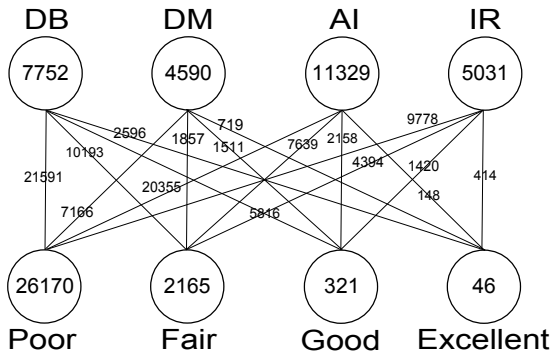
Figure: Cuboid Queries of the Graph Cube on DBLP Data Set

Effectiveness Evaluation



(a) (Area, Productivity)

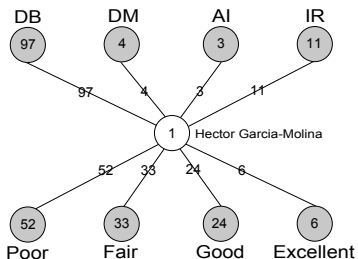
Figure: Cuboid Queries of the Graph Cube on DBLP Data Set



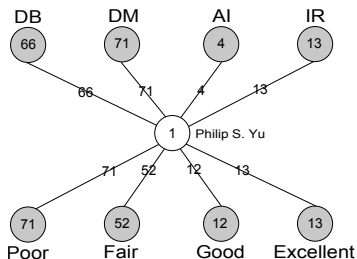
(a) Area \propto Productivity

Figure: Crossboid Queries of the Graph Cube on DBLP Data Set

Effectiveness Evaluation

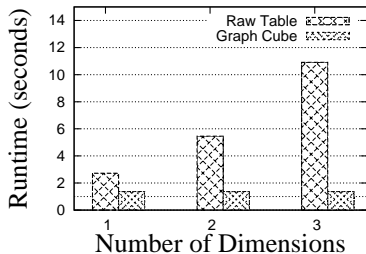


(a) Area \times Base \times Productivity for "Hector Garcia-Molina"

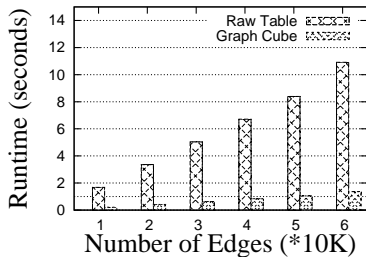


(b) Area \times Base \times Productivity for "Philip S. Yu"

Figure: Crossboid Queries of the Graph Cube on DBLP Data Set



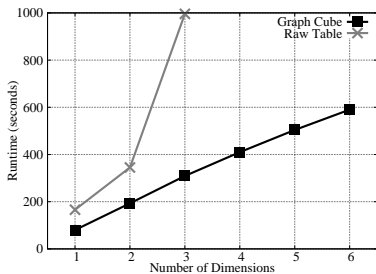
(a) Time v.s. # Dimensions



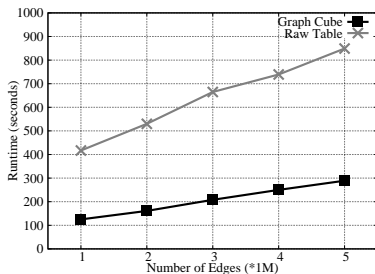
(b) Time v.s. # Edges

Figure: Full Materialization of Graph Cube for DBLP Data Set

Efficiency Evaluation

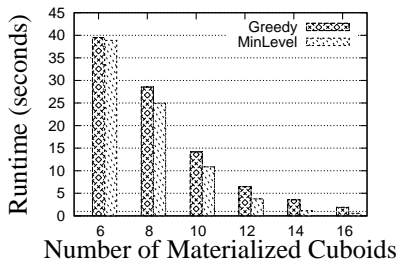


(a) Time v.s. # Dimensions

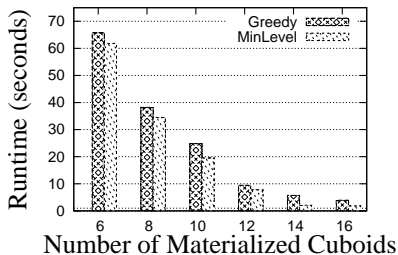


(b) Time v.s. # Edges

Figure: Full Materialization of Graph Cube for IMDB Data Set



(a) Cuboid Queries



(b) Crossboid Queries

Figure: Average Query Respond Time w.r.t. Different Partial Materialization Algorithms

- 1 This work seeks to enhance decision-support functionality on large multidimensional networks
- 2 **Graph cube:** A new data warehousing model is designed specifically for efficient aggregation on multidimensional networks
- 3 Different query models and OLAP solutions for Graph Cube are proposed and studied
 - **Crossboid queries** break the boundary of the traditional OLAP model by straddling multiple cuboids of the Graph Cube
- 4 The implementation of Graph Cube is discussed and the experimental results have demonstrated the power and efficacy of Graph Cube as **the first**, to the best of our knowledge, tool for warehousing and OLAP large multidimensional networks

Thank you