Data Engineering Research Group

4 faculty members
Reynold Cheng
David Cheung
Ben Kao
Nikos Mamoulis

20 research students (10 PhD, 10 MPhil)
Success Stories

Papers in the past 5 years

- 58 in top DB and DM conferences (9 SIGMOD, 15 VLDB, 17 ICDE, 3 EDBT, 4 CIKM, 3 SIGKDD, 6 ICDM)
- 28 in top DB and DM journals (5 TODS, 7 VLDBJ, 15 TKDE, 1 TKDD)

PhD alumni with faculty positions

(Rutgers, HKPolyU, Mexico State U, Aalborg U, Macau U, Renmin U)
Reynold Cheng

Background

HKU (BSc, MPhil 95-00), Purdue (PhD, 00-05), HKPolyU (Asst. Prof, 05-08)

Research

Database management, uncertainty management, data mining, spatial databases
Data Uncertainty

Data is often imprecise and erroneous. Handle data uncertainty, or service quality can be degraded!
The ORION Database

- Treat data uncertainty as a first-class citizen

- A probabilistic query provides answers with probabilities (e.g., Mary has a 80% chance to be in HKU)
<table>
<thead>
<tr>
<th>Queries in ORION</th>
<th>CREATE table T( k INTEGER primary key, a UNCERTAIN);</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create a table with <strong>UNCERTAIN</strong> type</td>
<td>Insert into T values (2, '(g, μ, σ)');</td>
</tr>
<tr>
<td>Insert Gaussian pdf (μ,σ)</td>
<td>Display uncertain info. of a if a &gt; 5</td>
</tr>
<tr>
<td>Display uncertain info. of a if a &gt; 5</td>
<td>SELECT a FROM T where a &gt; 5;</td>
</tr>
<tr>
<td>Equality join of uncertain attributes (=% returns probability of equality)</td>
<td>SELECT R.k, S.k, R.a =% S.a FROM R,S WHERE R.a = S.a;</td>
</tr>
<tr>
<td>Entities with prob. giving min value of a (e.g., {(3,0.5), (5,0.3), (11,0.2)})</td>
<td>SELECT Emin(T.a) from T;</td>
</tr>
<tr>
<td>Min value of a for table T (UNCERTAIN)</td>
<td>SELECT Vmin(T.a) from T;</td>
</tr>
</tbody>
</table>
David Cheung

Background

CUHK (BSc), Simon Fraser (MSc, PhD 83-88)

Research

Security and authentication in outsourced databases; data interoperability theory; queries on community networks
Outsourcing Data Mining Tasks

- Frequent itemset mining
Integrity concern

• Is the result correct?

• Scenario 1:
  • Honest but careless service provider
  • Example: incorrect implementation of mining algorithm, mistakes in settings

• Scenario 2:
  • Lazy service provider
  • Example: just execute on a sampled database to save cost

• Scenario 3:
  • Malicious service provider
  • Example: paid by a competitor of the data owner to return a wrong result; or provider/network falls victim of a malicious attack
Solution: artificial itemset planting

1. Generate an artificial database DB' so that the frequent itemsets L' in DB' are controlled and known to the data owner.

2. Service provider works on combined database.

3. Verify L'.
Ben Kao

Background

HKU (BSc 86-89), Princeton-Stanford (PhD 89-95)

Research

Database Systems, Information Retrieval, Data Mining
Finding Key Moments in Social Networks

“Distance” between two Facebook users over a 1-year period. (They are disconnected before Day 178 and finally became friends on Day 365.)
How did they \((u \text{ and } v)\) become friends?

• To understand how friendships are established, we need to study the events that happened at certain “key moments”. For example, what happened (Events (a), (b) or (c) above) that led to the shortening of two users’ distance from each other?

• But first, we need to discover those “key moments” so we know which “snapshots” of the Facebook graph we should look at.
Evolving Graph Sequence (EGS) Processing

• We model the dynamics of a social network as a (big) sequence of (big) evolving graph snapshots.

• We study efficient graph algorithms for identifying key moments (snapshots at which sharp changes in certain key measures are observed).

• Such key moments help social network analysts investigate the various properties of gigantic social networks.
Nikos Mamoulis

Background

UPatras (BSc-MSc 90-95), HKUST (PhD 97-00), CWI (00-01)

Research

Spatial Databases, Managing and Mining Complex Data Types, Privacy and Security, Information Retrieval.
Snippets of Data Subjects in Databases

Web

Christos Faloutsos
SCS CSD Professor's affiliations, research, projects, publications and teaching.
www.cs.cmu.edu/~christos/ - 9k

Michalis Faloutsos
The Homepage of Michalis Faloutsos ... Interesting and Miscellaneous Links · Fun pictures · Other Faloutsos on the web; The Teach-To-Learn Initiative:
www.cs.ucr.edu/~michalis/ - 5k

Petros Faloutsos
Courses · Press Coverage · Publications · Research Highlights · Awards · MAGIX Lab · Curriculum Vitae · Family · Other Faloutsos on Web.
www.cs.ucla.edu/~pfal/ - 4k

DBLP database

Author: Christos Faloutsos
Co-Author(s): Michalis Faloutsos, Petros Faloutsos.

Author: Michalis Faloutsos
Paper: The QBIC Project: Querying Images by Content, Using, Color, Texture and Shape.
Co-Author(s): Carlton W. Niblack, Dragutin Petkovic, Peter Yanker.

Author: Petros Faloutsos
Paper: Efficient and Effective Querying by Image Content.
Co-Author(s): N. Roussopoulos, T. Sellis.

Author: Christos Faloutsos
Co-Author(s): Anindo Banerjee, Rajesh Pankaj.

Co-Author(s): Aiguo Fei.

results of “Faloutsos”
Data Subject Schema Graph

...based on database schema
Object Summary of a Given Entity

...based on DS Schema Graph and actual data
Software Engineering Group

Prof. T.H. Tse (PhD LSE)

Research

Software Engineering: program testing, debugging, and analysis with application on object-oriented software, concurrent systems, pervasive computing, service-oriented applications, graphic applications, and numerical programs.