Big Data
Graph Data
Incomplete Information

Leonid Libkin
A bit about the group

• Ranges from 6 (never again) to one
• Right now, 4 postdocs
• Looking for new student (one, at most two)
• Key themes: data management (3 Vs of big data - volume, variety, veracity: scalability; relational, XML, graph data; incompleteness and inconsistency), foundations (as they are needed to handle those questions)
Past students/postdocs

• Mainly academic jobs (12 out of 14 have academic positions in places such as Paris, Singapore, Santiago, Warsaw, Bordeaux; one at IBM, one at Oracle)

• Several notable awards by students:
  • BCS Distinguished Dissertation Award
  • Cor Bayeen Award
  • EPSRC postdoctoral fellowship
  • ACM SIGMOD Honorable mention (2nd prize)
  • 8 (or more) best paper awards
• Main demands to students:
  • very good background
  • interest in what they are doing
• Flexibility with projects: there is always a choice, nothing is ever imposed
Big data and data management

- Data analytics only account for a small fraction of time invested in big data processing!

- **Data wrangling** (handling data before analyses can begin) can take up to 80% of the effort.
The 4 Vs

- **Volume, Velocity, Variety, Veracity**
- **Volume** - scalability (Wenfei will talk about it)
- **Variety** - graph data (XML is done and gone)
- **Veracity** - handling uncertainty
Graph Databases

Old techniques do not work.
New issues: combining data and topology
Essentially property graphs of products such as Neo4j
Incomplete Information

- **Practice**: incorrect answers (your laptop thinks that $|X|>|Y|$ and $X-Y=\emptyset$ are consistent!)

- **Theory**: computationally expensive notions of correctness

- It has been like that for 30+ years, until very recently

- Trying to break the *curse of incomplete information*
Why $N-1=0$ for all $N$

built into SQL standards - the tool most commonly used by data analysts
Why $N-1=0$ for all $N$

built into SQL standards - the tool most commonly used by data analysts

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>Difference R-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Why N-1=0 for all N

built into **SQL** standards - the tool most commonly used by data analysts

<table>
<thead>
<tr>
<th>R</th>
<th>S</th>
<th>Difference R-S</th>
</tr>
</thead>
</table>
| 1 | - | **SELECT R.A FROM R**  
WHERE R.A NOT IN  
(SELECT S.A FROM S)  
... |
| 2 |
| 3 |
| ... |
| N |
Why $N-1=0$ for all $N$

built into SQL standards - the tool most commonly used by data analysts

<table>
<thead>
<tr>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>

Difference $R-S$

$\text{SELECT R.A FROM R WHERE R.A NOT IN (SELECT S.A FROM S)}$

Answer: EMPTY for all $N$
Why $N-1=0$ for all $N$

built into SQL standards - the tool most commonly used by data analysts

<table>
<thead>
<tr>
<th>R</th>
<th>S</th>
<th>Difference R-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>SELECT R.A FROM R WHERE R.A NOT IN (SELECT S.A FROM S)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Answer: EMPTY for all $N$</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>So $N-1=0$ after all!</td>
</tr>
</tbody>
</table>
If interested...

• please come and talk to me
• libkin@inf.ed.ac.uk
• https://www.google.co.uk/#q=libkin