Data & Visual Analytics

Duen Horng (Polo) Chau
Georgia Tech

CSE 6242 A / CS 4803 DVA
Jan 8, 2013
I Work with Large Graphs
I Work with **Large Graphs**

= Large Network Data
Internet
50 Billion Web Pages
Facebook
800 Million Users
Citation Network
250 Million Articles

Modified from well-formed.eigenfactor.org
Many More

**Twitter**
Who-follows-whom (500 million users)

**Amazon**
Who-buys-what (120 million users)

**AT&T Cellphone network**
Who-calls-whom (100 million users)

**Protein-protein interactions**
200 million possible interactions in human genome

# Large Graphs I Analyzed

<table>
<thead>
<tr>
<th>Graph</th>
<th>Nodes</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>YahooWeb</td>
<td>1.4 Billion</td>
<td>6 Billion</td>
</tr>
<tr>
<td>Symantec Machine-File Graph</td>
<td>1 Billion</td>
<td>37 Billion</td>
</tr>
<tr>
<td>Twitter</td>
<td>104 Million</td>
<td>3.7 Billion</td>
</tr>
<tr>
<td>Phone call network</td>
<td>30 Million</td>
<td>260 Million</td>
</tr>
</tbody>
</table>
7±2

Number of **items** an average human holds in **working memory**

*George Miller, 1956*
Data ➔ Insights
How to do that?

COMPUTATION + VISUALIZATION
## How to do that?

<table>
<thead>
<tr>
<th><strong>COMPUTATION</strong></th>
<th><strong>VISUALIZATION</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td>User-driven; iterative</td>
</tr>
<tr>
<td>Summarization, clustering, classification</td>
<td>Interaction, visualization</td>
</tr>
<tr>
<td>&gt;Millions of nodes</td>
<td>Thousands of nodes</td>
</tr>
</tbody>
</table>

Both develop methods for making sense of network data.
How to do that?

**COMPUTATION**  **VISUALIZATION**

- Automatic
- Summarization, clustering, classification
- >Millions of nodes
How to do that?

<table>
<thead>
<tr>
<th><strong>COMPUTATION</strong></th>
<th><strong>VISUALIZATION</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td></td>
</tr>
<tr>
<td>Summarization,</td>
<td></td>
</tr>
<tr>
<td>clustering,</td>
<td></td>
</tr>
<tr>
<td>classification</td>
<td></td>
</tr>
<tr>
<td>&gt;Millions of</td>
<td></td>
</tr>
<tr>
<td>nodes</td>
<td></td>
</tr>
</tbody>
</table>
How to do that?

<table>
<thead>
<tr>
<th>COMPUTATION</th>
<th>VISUALIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td>User-driven; iterative</td>
</tr>
<tr>
<td>Summarization, clustering, classification</td>
<td>Interaction, visualization</td>
</tr>
<tr>
<td>Interaction, visualization</td>
<td>Thousands of nodes</td>
</tr>
</tbody>
</table>
How to do that?

**COMPUTATION**

![Diagram of interconnected nodes]

**VISUALIZATION**

- User-driven; iterative
- Interaction, visualization
- Thousands of nodes
How to do that?

**COMPUTATION**

- Automatic
- Summarization, clustering, classification

**VISUALIZATION**

- User-driven; iterative
- Interaction, visualization
- Thousands of nodes
“Computers are incredibly fast, accurate, and stupid.

Human beings are incredibly slow, inaccurate, and brilliant.

Together they are powerful beyond imagination.”
Logistics

Course homepage  
[poloclub.gatech.edu/cse6242/](poloclub.gatech.edu/cse6242/)

Discussion, Q&A, find teammates  
Piazza (link on homepage)

Submission  
T-Square
Course Goals

• Learn a broad class of scalable visual and computation techniques and tools, for typical data types

• Learn how to combine both kinds of methods (how they complement each other)

• Gain practical know-how

• Gain breadth of knowledge
Course Expectation

• Overview of scalable visual and computation techniques and tools

• Gain knowledge & experience (useful for jobs, research)

• Experience with designing and developing an interactive analysis tool
Schedule

See course homepage
Course Staff

**Instructor**
Duen Horng (Polo) Chau
Assistant Professor, CSE
Thu 3-4pm, Klaus 1324

**TA**
Parikshit Ram
PhD student, CSE
Mon 4-5pm, Klaus 1315
Grading

- 3 homework assignments (40%)
- End-to-end analysis
- Techniques (computation and vis)
- Hadoop (+ other “big data” tools)
- Group project (50%) -- 2 to 3 people
- Participation (10%) -- in class, on Piazza