Data & Visual Analytics

Duen Horng (Polo) Chau

Georgia Tech

CSE6242 / CX4242
Who is Polo?

Polo Chau
Associate Director, MS Analytics
Assistant Prof, CSE
Adjunct Assistant Prof, IC
POLO CHAU

Associate Director, MS in Analytics
Assistant Professor, School of Computational Science & Engineering
Adjunct Assistant Professor, School of Interactive Computing
College of Computing
Georgia Tech

Admin: Carolyn Young    Financial Manager: Arlene Washington
polo@gatech.edu    www.cc.gatech.edu/~dchau
Office: Klaus 1324    404-385-7682
Google Scholar (h-index: 18)    YouTube videos

<table>
<thead>
<tr>
<th>POSITIONS</th>
</tr>
</thead>
</table>
| May 2014 - Associate Director  
MS in Analytics, Georgia Tech |
| Aug 2012 - Assistant Professor  
School of Computational Science & Engineering, Georgia Tech |
| Dec 2012 - Adjunct Assistant Professor  
School of Interactive Computing, Georgia Tech |

NIH MD2K Center of Excellence, Co-PI
MD2K
Visualization at Georgia Tech

IDEA workshop at KDD 2013, 2014, 2015
TAs

Meera Manohar Kamath
Gopi Krishnan Nambiar
Siddharth Rajendra Raja
Ramakrishnan (Ramki) Kannan
Akanksha

Office hours listed on course homepage.
We work with (really) large data.
Internet
50 Billion Web Pages

www.worldwidewebsize.com  www.opte.org
Facebook
1.2 Billion Users

Modified from Marc_Smith, flickr
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 Networks We 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>
Number of items an average human holds in working memory

George Miller, 1956
Data

Insights
How to do that?

COMPUTATION + HUMAN INTUITION
How to do that?

<table>
<thead>
<tr>
<th>COMPUTATION</th>
<th>INTERACTIVE VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td>User-driven; iterative</td>
</tr>
<tr>
<td>Summarization,</td>
<td>Interaction, visualization</td>
</tr>
<tr>
<td>clustering,</td>
<td></td>
</tr>
<tr>
<td>classification</td>
<td></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** | **INTERACTIVE VIS**

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

**COMPUTATION**  **INTERACTIVE VIS**

Automatic

Summarization, clustering, classification

> Millions of nodes
How to do that?

<table>
<thead>
<tr>
<th>COMPUTATION</th>
<th>INTERACTIVE VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>User-driven; iterative</td>
</tr>
<tr>
<td></td>
<td>Interaction, visualization</td>
</tr>
<tr>
<td></td>
<td>Thousands of nodes</td>
</tr>
</tbody>
</table>
How to do that?

**COMPUTATION**

- Automatic
- Summarization, clustering, classification

**INTERACTIVE VIS**

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

**COMPUTATION**

- Automatic
- Summarization, clustering, classification

**INTERACTIVE VIS**

- 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.”

(Einstein might or might not have said this.)
Our Approach for Big Data Analytics

<table>
<thead>
<tr>
<th>DATA MINING</th>
<th>HCI</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 items</td>
<td>Thousands of items</td>
</tr>
</tbody>
</table>

Our research combines the **Best of Both Worlds**
Polonium

**Patented** with Symantec

Finds malware from **37 billion** file relationships

Serving **120 million** users worldwide

Published at SDM’11
MARCO
Detecting Fake Yelp Reviews

Best papers of SDM 2014
(top data mining conference)
Latent Gesture

CS undergraduate students: Prem Saravanan, Samuel Clarke

won prestigious Astronaut Scholarship
Insider Trading Detection with Securities and Exchange Commission (SEC)
NetProbe

Auction Fraud Detection on eBay
Apolo: Machine Learning + Visualization
Explore million-scale graphs in real time
Logistics

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

Discussion, Q&A, find teammates  
Piazza

Assignment Submission  
T-Square (for submissions only; use Piazza for discussion)
Course Goals & Expectation

• Learn scalable **visual** and **computation** techniques and tools, for typical data types

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

• Work on **real data & problem**

• Learn **practical** know-how

• Gain **breath** of knowledge
Course Goals

- Learn scalable **visual** and **computation** techniques and tools, for typical data types
- Learn how to **combine** both kinds of methods (how they complement each other)
- Work on **real data & problem**
- Learn **practical** know-how (useful for jobs, research)
- Gain **breath** of knowledge
Schedule

See course homepage
poloclub.gatech.edu/cse6242/
Grading

• 4 homework assignments (50%)
  • End-to-end analysis
  • Techniques (computation and vis)
  • “Big data” tools, e.g, Hadoop, Spark, etc.
• Group project (50%) -- 3 to 4 people
From Previous Classes…

- Class projects turned into papers at top conferences (KDD, IUI, etc.)
- Projects as portfolio pieces on CV
- Increased job and internship opportunities
  - Former students sent me “thank you” notes
Course Expectation

• Actively Participate in class! Ask questions during class, and on Piazza

• Polo will reserve last 5-10min of every lecture for Q&A