Data & Visual Analytics

Duen Horng (Polo) Chau
Georgia Tech

CSE6242 / CX4242
Aug 19, 2014
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

polo@gatech.edu         www.cc.gatech.edu/~dchau
Office: Klaus 1324   404-385-7682
Google Scholar (h-index: 17)     YouTube videos

Linkedin profile     Follow @PoloChau

POSITIONS

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

Excited to co-direct Georgia Tech’s MS Analytics, a unique 1-year program from GT’s world-class colleges of business, computing, and engineering.
Course Staff

Instructor: Polo Chau
Office hour: Tue, 11AM - 12PM, Klaus 1324

TAs:  Brian Kahng    Seungyeon Kim    Alan Zhang

Office hours: TBD
We work with (really) large data.
Internet
50 Billion Web Pages
Facebook
800 Million Users

Modified from Marc_Smith, flickr
Citation Network
250 Million Articles
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?

**COMPUTATION** | **INTERACTIVE VIS**

Both develop making sense of network data.

Automatic
User-driven; iterative
Summarization, clustering, classification
Interaction, visualization

> Millions of nodes

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.)
“Essentially, all models are wrong, but some are useful”

George Box
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
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

GIZMODO

Your Touchscreen Usage Is So Unique It Can Be Used as a Password

Andrew Liszewski
Filed to: SECURITY  4/17/14 1:20pm

engadget

YAHOO!
Insider Trading Detection with Securities and Exchange Commission (SEC)
NetProbe
Auction Fraud Detection on eBay
Apolo: Machine Learning + Visualization
Find relevant nodes in real time  (CHI’11)
CareFlow: Healthcare Visual & Data Analytics
## Logistics

<table>
<thead>
<tr>
<th>Course homepage</th>
<th><a href="poloclub.gatech.edu/cse6242/">poloclub.gatech.edu/cse6242/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussion, Q&amp;A, find teammates</td>
<td>Piazza</td>
</tr>
<tr>
<td>Submission</td>
<td>T-Square <em>(for submissions only; use Piazza for discussion)</em></td>
</tr>
</tbody>
</table>
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)

• Gain **practical** know-how

• Gain **breath** 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
Grading

• 3-4 homework assignments (40%)
• End-to-end analysis
• Techniques (computation and vis)
• Hadoop (+ other “big data” tools)
• Group project (50%) -- 3 to 4 people
• Participation (10%) -- in class, and on Piazza