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

polo@gatech.edu, www.cc.gatech.edu/~dchau

Office: Klaus 1324, 404-385-7682

Google Scholar (h-index: 17)  
YouTube videos

## Positions

<table>
<thead>
<tr>
<th>Date</th>
<th>Position</th>
<th>Department</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 2014</td>
<td>Associate Director</td>
<td>MS in Analytics, Georgia Tech</td>
</tr>
<tr>
<td>Aug 2012</td>
<td>Assistant Professor</td>
<td>School of Computational Science &amp; Engineering, Georgia Tech</td>
</tr>
<tr>
<td>Dec 2012</td>
<td>Adjunct Assistant Professor</td>
<td>School of Interactive Computing, Georgia Tech</td>
</tr>
</tbody>
</table>

## Education

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

Polo Chau
Amir Afsharinejad
Yichen Wang
Chris Berlind
Meera Manohar Kamath

Office hours listed on course homepage.
We work with (really) large data.
Internet
50 Billion Web Pages
Facebook
800 Million 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><strong>COMPUTATION</strong></th>
<th><strong>INTERACTIVE VIS</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** | **INTERACTIVE VIS**
---|---
Automatic
Summarization, clustering, classification
>Millions of nodes
How to do that?

**COMPUTATION**

Automatic

Summarization, clustering, classification

>Millions of nodes

**INTERACTIVE VIS**
## How to do that?

<table>
<thead>
<tr>
<th><strong>COMPUTATION</strong></th>
<th><strong>INTERACTIVE VIS</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>User-driven; iterative</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Interaction, visualization</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Thousands of nodes</strong></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**

<table>
<thead>
<tr>
<th>Interaction, visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thousands of nodes</td>
</tr>
</tbody>
</table>

**INTERACTIVE VIS**

<table>
<thead>
<tr>
<th>User-driven; iterative</th>
</tr>
</thead>
</table>

> Millions of nodes

Automatic

Summarization, clustering, classification
“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

Our research combines the Best of Both Worlds

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

Human-Computer Interaction
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

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

Course homepage: polooclub.gatech.edu/cse6242/

Discussion, Q&A, find teammates: Piazza

Assignment Submission: T-Square (for submissions only; use Piazza for discussion)
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

• Projects from previous class turned into papers (KDD, IUI, etc.)
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
Schedule

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

• 4-5 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