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: 20)  
Follow @PoloChau  

POLO CHAU

Legal name:  
Duen Horng Chau

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

EDUCATION

Aug 2012  
Ph.D. Machine Learning  
Carnegie Mellon University  
Thesis: Data Mining Meets HCI: Making Sense of Large Graphs
TAs

Gopi Krishnan Nambiar
Nilaksh Das
Pradeep Rajendra
Ajitesh Jain
Vishakha Singh

Office hours to be listed on course homepage poloclub.gatech.edu/cse6242
We work with (really) large data.
Facebook
1.2 Billion 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
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>
7±2

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?

<table>
<thead>
<tr>
<th><strong>COMPUTATION</strong></th>
<th><strong>INTERACTIVE VIS</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 nodes</td>
<td></td>
</tr>
</tbody>
</table>
How to do that?

<table>
<thead>
<tr>
<th>COMPUTATION</th>
<th>INTERACTIVE VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td></td>
</tr>
<tr>
<td>Summarization, clustering, classification</td>
<td></td>
</tr>
<tr>
<td>&gt;Millions of nodes</td>
<td></td>
</tr>
<tr>
<td>COMPUTATION</td>
<td>INTERACTIVE VIS</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------</td>
</tr>
<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?

<table>
<thead>
<tr>
<th>COMPUTATION</th>
<th>INTERACTIVE VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Diagram of connected nodes]</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
“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

**DATA MINING** | **HCI**
---|---
Automatic | User-driven; iterative
Summarization, clustering, classification | Interaction, visualization
>Millions of items | Thousands of items

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

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

WIRED
engadget

YAHOO!

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

Auction Fraud Detection on eBay

THE WALL STREET JOURNAL.
Apolo: Machine Learning + Visualization

Explore million-scale graphs in real time
Logistics

Course homepage  
poloclub.gatech.edu/cse6242/

Discussion, Q&A,  
find teammates  
Piazza: goo.gl/9RYh5r  
or piazza.com/class/ijbuzug4gs43de

Assignment Submission  
T-Square  
(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)
- 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%) -- 4 to 5 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
What Polo expects from you

• Actively participate throughout the course!
• Ask questions **during class** and on **Piazza**
• Help out whenever you can, e.g., help answer questions on Piazza
• Polo will reserve last 5-10min of every lecture for Q&A
PASSAGE: A Travel Safety Assistant With Safe Path Recommendations For Pedestrians

Matthew Garvey
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
mgarvey6@gatech.edu

Meghna Natraj
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
mnatraj@gatech.edu

Nilaksh Das
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
nilakshdas@gatech.edu

Bhanu Verma
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
bhanuverma@gatech.edu

Jiaxing Su
College of Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
jiaxingsu@gatech.edu

Abstract
Atlanta has consistently ranked as one of the most dangerous cities in America with over 2.5 million crime events recorded within the past six years. People who commute by walking are highly susceptible to crime here. To address this problem, our group has developed a mobile application, PASSAGE, which uses crime data to find "safe paths" for pedestrians in Atlanta.

Figure 1: Paths recommended by PASSAGE

ACM Reference:
Jiaxing Su, Matthew Garvey, Meghna Natraj, Nilaksh Das, Bhanu Verma.
PASSAGE: A Travel Safety Assistant With Safe Path Recommendations For Pedestrians.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Aurigo: An Interactive Tour Planner for Personalized Itineraries

Alexandre Yahi; Antoine Chassang; Louis Raynaud; Hugo Duthil; Duen Horng (Polo) Chau
Georgia Institute of Technology
{alexandre.yahi, antoine.chassang, l raynaud, hduthil, polo}@gatech.edu

ABSTRACT
Planning personalized tour itineraries is a complex and challenging task for both humans and computers. Doing it manually is time-consuming; approaching it as an optimization problem is computationally NP hard. We present Aurigo, a tour planning system combining a recommendation algorithm with interactive visualization to create personalized itineraries. This hybrid approach enables Aurigo to take into account both quantitative and qualitative preferences of the user. We conducted a within-subject study with 10 participants, which demonstrated that Aurigo helped them find points of interest quickly. Most participants chose Aurigo over Google Maps as their preferred tools to create personalized itineraries. Aurigo may be integrated into review websites or social networks, to leverage their databases of reviews and ratings and provide better itinerary recommendations.

Author Keywords
User Interfaces; Visualization; Recommendation; Tour itinerary planning

ACM Classification Keywords

ISPARK: Interactive Visual Analytics for Fire Incidents and Station Placement

Subhajit Das, Andrea McCarter, Joe Minieri, Nandita Damaraju, Sriram Padmanabhan, Duen Horng (Polo) Chau
Georgia Tech
Atlanta, GA, USA
{das, andream, jminieri, nandita, sriramp, polo}@gatech.edu

ABSTRACT

In support of helping to reduce the response time of firefighters, and thus deaths, injuries, and property loss due to fires, we introduce ISPARK. The ISPARK system determines where fire stations should be located, analyzes the primary causes of fires, the existing infrastructure, and response times, by using visualizations which show the GIS mapping of fire stations on a dashboard. Incidents and response times are shown as additional layers, with clustering of fire incidents to determine predicted fire station locations, forecasting of fire incidents using regression, causal, infrastructure, and personnel analysis, creating an interactive, multi-faceted method for locating fire stations. A comparison of urban and rural fire incident response times is another dimension of this study. We demonstrate ISPARK's usage and benefits using a publicly available dataset describing 300,000 fire incidents in the states of Massachusetts and Maine. ISPARK is generalizable to other geographic areas.