Analytics Building Blocks

Duen Horng (Polo) Chau
Associate Professor, College of Computing
Associate Director, MS Analytics
Machine Learning Area Leader, College of Computing
Georgia Tech

Partly based on materials by
Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos
Collection
Cleaning
Integration
Analysis
Visualization
Presentation
Dissemination
Building blocks. Not Rigid “Steps”.

Can skip some

Can go back (two-way street)

- **Data types** inform **visualization** design
- **Data size** informs choice of **algorithms**
- **Visualization** motivates more **data cleaning**
- **Visualization** challenges algorithm assumptions
e.g., user finds that results don’t make sense
How “big data” affects the process? (Hint: almost everything is harder!)

<table>
<thead>
<tr>
<th>Collection</th>
<th>The Vs of big data (3Vs originally, then 7, now 42)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cleaning</td>
<td><strong>Volume</strong>: “billions”, “petabytes” are common</td>
</tr>
<tr>
<td>Integration</td>
<td><strong>Velocity</strong>: think Twitter, fraud detection, etc.</td>
</tr>
<tr>
<td>Analysis</td>
<td><strong>Variety</strong>: text (webpages), video (youtube)…</td>
</tr>
<tr>
<td>Visualization</td>
<td><strong>Veracity</strong>: uncertainty of data</td>
</tr>
<tr>
<td>Presentation</td>
<td><strong>Variability</strong></td>
</tr>
<tr>
<td>Dissemination</td>
<td><strong>Visualization</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Value</strong></td>
</tr>
</tbody>
</table>

http://www.ibmbigdatahub.com/infographic/four-vs-big-data
http://dataconomy.com/seven-vs-big-data/
https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx
Two Example Projects
from Polo Club
Duen Horng (Polo) Chau, Aniket Kittur, Jason I. Hong, Christos Faloutsos. CHI 2011.
Finding More Relevant Nodes

Citation network

HCI Paper

Data Mining Paper
Finding More Relevant Nodes

HCI Paper

Data Mining Paper

Citation network
Finding More Relevant Nodes

Apolo uses **guilt-by-association** (Belief Propagation)

**HCI**
**Data Mining**

Citation network
**Demo:** Mapping the **Sensemaking Literature**

**Nodes:** 80k papers from Google Scholar (node size: #citation)

**Edges:** 150k citations

---

**The cost structure of sensemaking**


245 citations  8 versions
The cost structure of sensemaking


245 citations  8 versions
The cost structure of sensemaking
Key Ideas (Recap)

Specify **exemplars**

Find **other** relevant nodes (BP)
What did **Apolo** go through?

**Collection**  Scrape Google Scholar. No API. 😞

**Cleaning**

**Integration**

**Analysis**  Design inference algorithm  
(Which nodes to show next?)

**Visualization**  Interactive visualization you just saw

**Presentation**  Paper, talks, lectures

**Dissemination**  You will a new Apolo prototype  
(called Argo)
Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning

Duen Horng (Polo) Chau, Aniket Kittur, Jason I. Hong, Christos Faloutsos
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{dchau, n kittur, jasonh, christos}@cs.cmu.edu

ABSTRACT
Extracting useful knowledge from large network datasets has become a fundamental challenge in many domains, from scientific literature to social networks and the web. We introduce Apolo, a system that uses a mixed-initiative approach—combining visualization, rich user interaction and machine learning—to guide the user to incrementally and interactively explore large network data and make sense of it. Apolo engages the user in bottom-up sensemaking to gradually build up an understanding over time by starting small, rather than starting big and drilling down. Apolo also helps users find relevant information by specifying exemplars, and then using a machine learning method called Belief Propagation to infer which other nodes may be of interest. We evaluated Apolo with twelve participants in a between-subjects study, with the task being to find relevant new papers to update an existing survey paper. Using expert judges, participants using Apolo found significantly more relevant papers. Subjective feedback of Apolo was also very positive.

Figure 1. Apolo displaying citation network data around the article The Cost Structure of Sensemaking. The user gradually builds up a mental model of the research areas around the article by manually inspecting some neighboring articles in the visualization and specifying them as exemplar articles (with colored dots underneath) for some ad hoc groups, and instructs Apolo to find more articles relevant to them.
NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks. Shashank Pandit, Duen Horng (Polo) Chau, Samuel Wang, Christos Faloutsos. WWW 2007
Find **bad sellers** (fraudsters) on eBay who don’t deliver their items.

NetProbe: The Problem

Non-delivery fraud is a common auction fraud

NetProbe: Key Ideas

- Fraudsters *fabricate their reputation* by “trading” with their accomplices
- Fake transactions form *near bipartite cores*
- How to detect them?
NetProbe: Key Ideas

Use Belief Propagation

Fraudster
Accomplice
Honest

Darker means more likely
NetProbe: Main Results
alisher

Registration: Aug 13, 2006  
Location: United States

Fraudsters: 95%
Accomplice: 4%
Honest: 1%

Suspected fraudster - this user has been behaving much like the other suspects by trading with the similar sets of possible accomplices.
<table>
<thead>
<tr>
<th>Process</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection</td>
<td>Scraping (built a “scraper”/“crawler”)</td>
</tr>
<tr>
<td>Cleaning</td>
<td></td>
</tr>
<tr>
<td>Integration</td>
<td></td>
</tr>
<tr>
<td>Analysis</td>
<td>Design detection algorithm</td>
</tr>
<tr>
<td>Visualization</td>
<td>Paper, talks, lectures</td>
</tr>
<tr>
<td>Presentation</td>
<td>Not released</td>
</tr>
<tr>
<td>Dissemination</td>
<td></td>
</tr>
</tbody>
</table>
NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks

Shashank Pandit, Duen Horng Chau, Samuel Wang, Christos Faloutsos
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{shashank, dchau, samuelwang, christos}@cs.cmu.edu

ABSTRACT

Given a large online network of online auction users and their histories of transactions, how can we spot anomalies and auction fraud? This paper describes the design and implementation of NetProbe, a system that we propose for solving this problem. NetProbe models auction users and transactions as a Markov Random Field tuned to detect the suspicious patterns that fraudsters create, and employs a Belief Propagation mechanism to detect likely fraudsters. Our experiments show that NetProbe is both efficient and effective for fraud detection. We report experiments on synthetic graphs with as many as 7,000 nodes and 30,000 edges, where NetProbe was able to spot fraudulent nodes with over 90% precision and recall, within a matter of seconds. We also report experiments on a real dataset crawled from eBay, with nearly 700,000 transactions between more than 66,000 users, where NetProbe was highly effective at unearthing hidden networks of fraudsters, within a realistic response time of about 6 minutes. For scenarios where the underlying data is dynamic in nature, we propose Incremental NetProbe, which is an approximate, but fast, variant of NetProbe. Our experiments prove that Incremental NetProbe.
Homework 1 (out next week; tasks subject to change)

- Simple “End-to-end” analysis
- Collect data using API
- Store in SQLite database
- Create graph from data
- Analyze, using SQL queries (e.g., create graph’s degree distribution)
- Visualize graph using **Gephi**
- Describe your discoveries