

poloclub.github.io/#cse6242

CSE6242/CX4242: **Data** & **Visual** Analytics

Analytics Building Blocks

Duen Horng (Polo) Chau

Professor, College of Computing
Associate Director, MS Analytics
Georgia Tech

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

Building blocks. **Not Rigid “Steps”.**

Collection

Can skip some

Cleaning

Can go back (two-way street)

Integration

- **Data types** inform **visualization** design

Analysis

- **Data size** informs choice of **algorithms**

Visualization

- **Visualization** motivates more **data cleaning**

Presentation

- **Visualization** challenges algorithm assumptions

Dissemination

e.g., user finds that results don't make sense

How “big data” affects the process?

(Hint: almost **everything** is harder!)

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

The Vs of big data (3Vs originally, then 7, now 42)

Volume: “billions”, “petabytes” are common

Velocity: think X/Twitter, fraud detection, etc.

Variety: text (webpages), video (youtube)...

Veracity: uncertainty of data

Variability

Visualization

Value

<http://dataconomy.com/seven-vs-big-data/>

<https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html>

Two Example Projects

from Polo Club

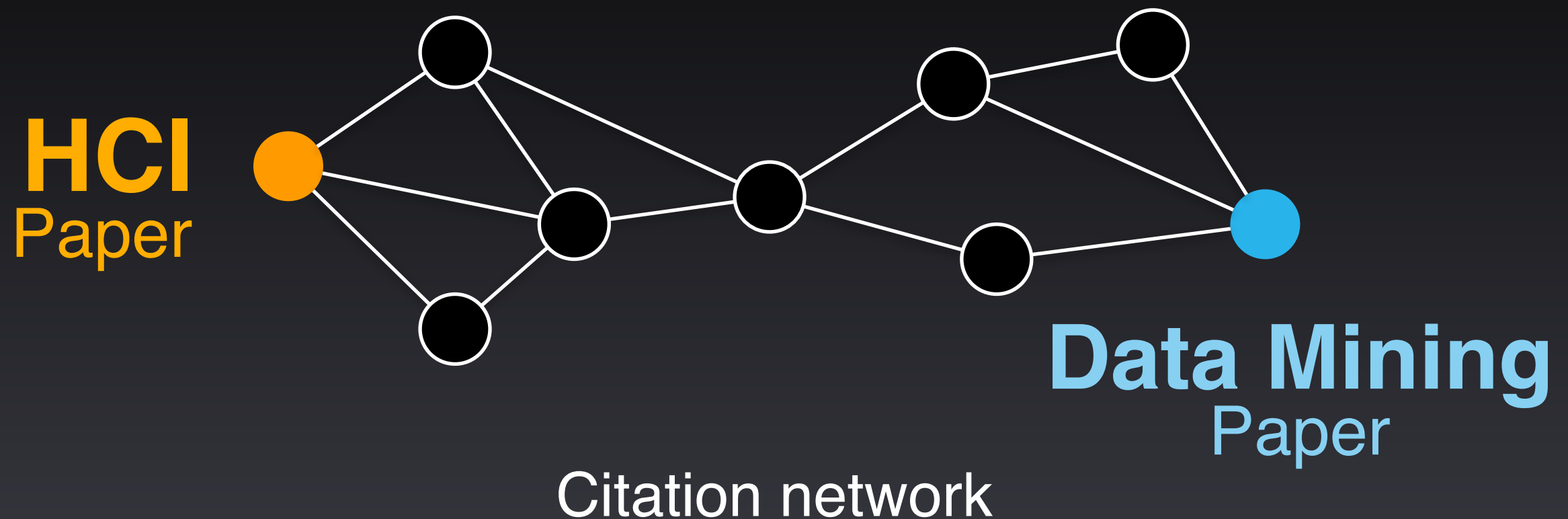
Apolo Graph Exploration: Machine Learning + Visualization

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. CHI 2011.



BEAUTIFUL HAIRBALL
DEATH STAR
SPAGHETTI

Finding **More** Relevant Nodes

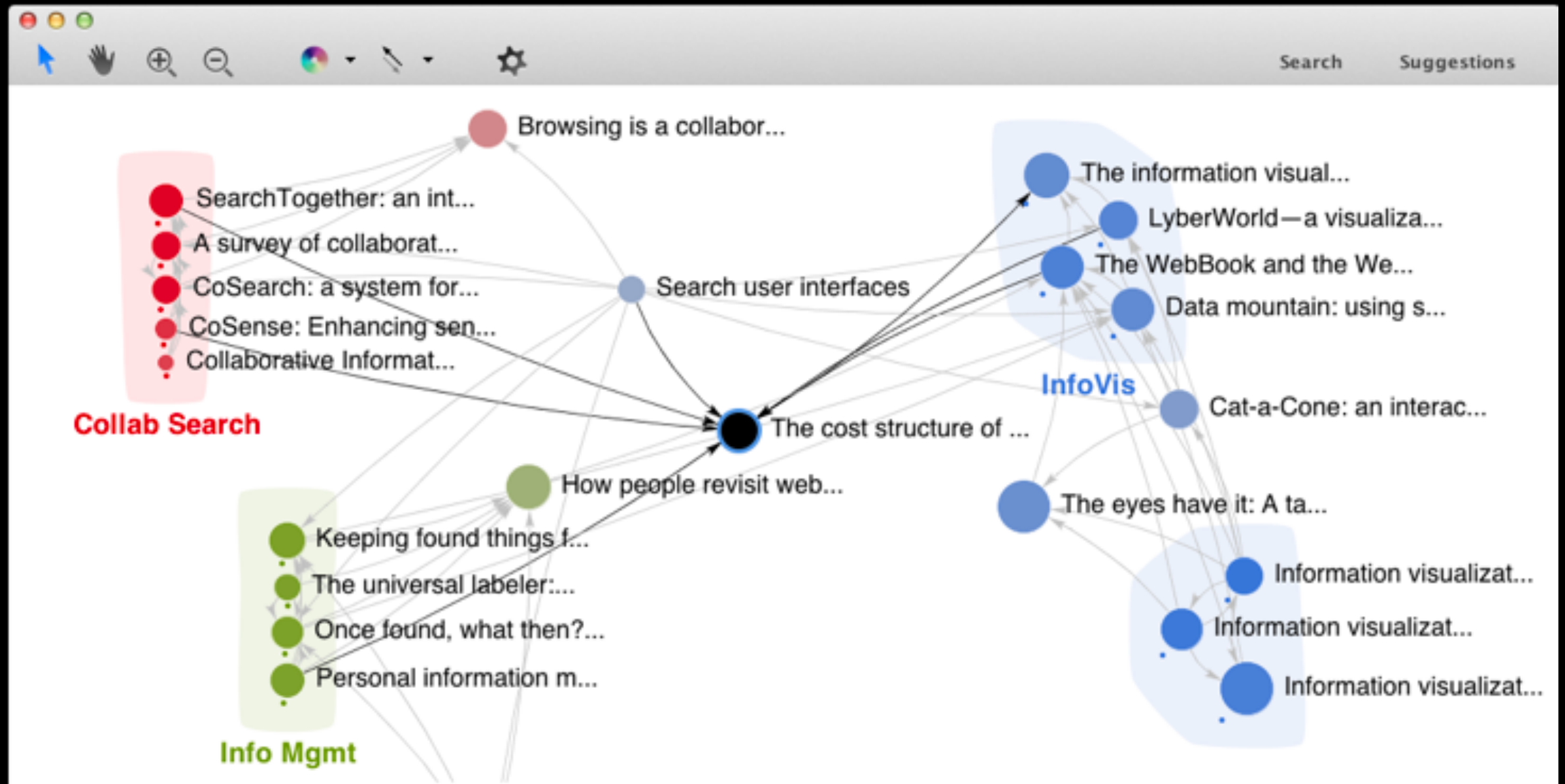


Apolo uses **guilt-by-association**
(Belief Propagation)

Demo: Mapping the Sensemaking Literature

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

Edges: 150k citations



The cost structure of sensemaking

Russell, D.M. and Stefik, M.J. and Pirolli, P. and Card, S.K.

245 citations 8 versions

PDF 1993



Suggestions

The information visualizer, an inf... 1991

The WebBook and the Web Forag... 1996

LyberWorld—a visualization user... 1994

The structure of the information... 1997

Information visualization 2009

"I'll get that off the audio": a cas... 1997

An organic user interface for sear... 1995

Using a landscape metaphor to re... 1993

Personal information management 2007

SearchTogether: an interface for c... 2007

Information foraging theory: Ada... 2007

Pirolli, P.	107
-------------	-----

Investigating behavioral variability... 2007

White, R.W. and Drucker, S.M. 79

Jigsaw: Supporting investigative... 2008

Stasko, J. and Görg, C. and Liu, Z. 71

The cost-of-knowledge character... 1994

Card, S.K. and Pirolli, P. and Mackinlay, J.D. 54

Collaborative conceptual design:... 1996

Potts, C. and Catledge, L.	45
----------------------------	----

googlescholar.db

- The cost structure of sen...

The cost structure of sensemaking

PDF 1993

Russell, D.M. and Stefik, M.J. and Pirolli, P. and Card, S.K.

245 citations 8 versions

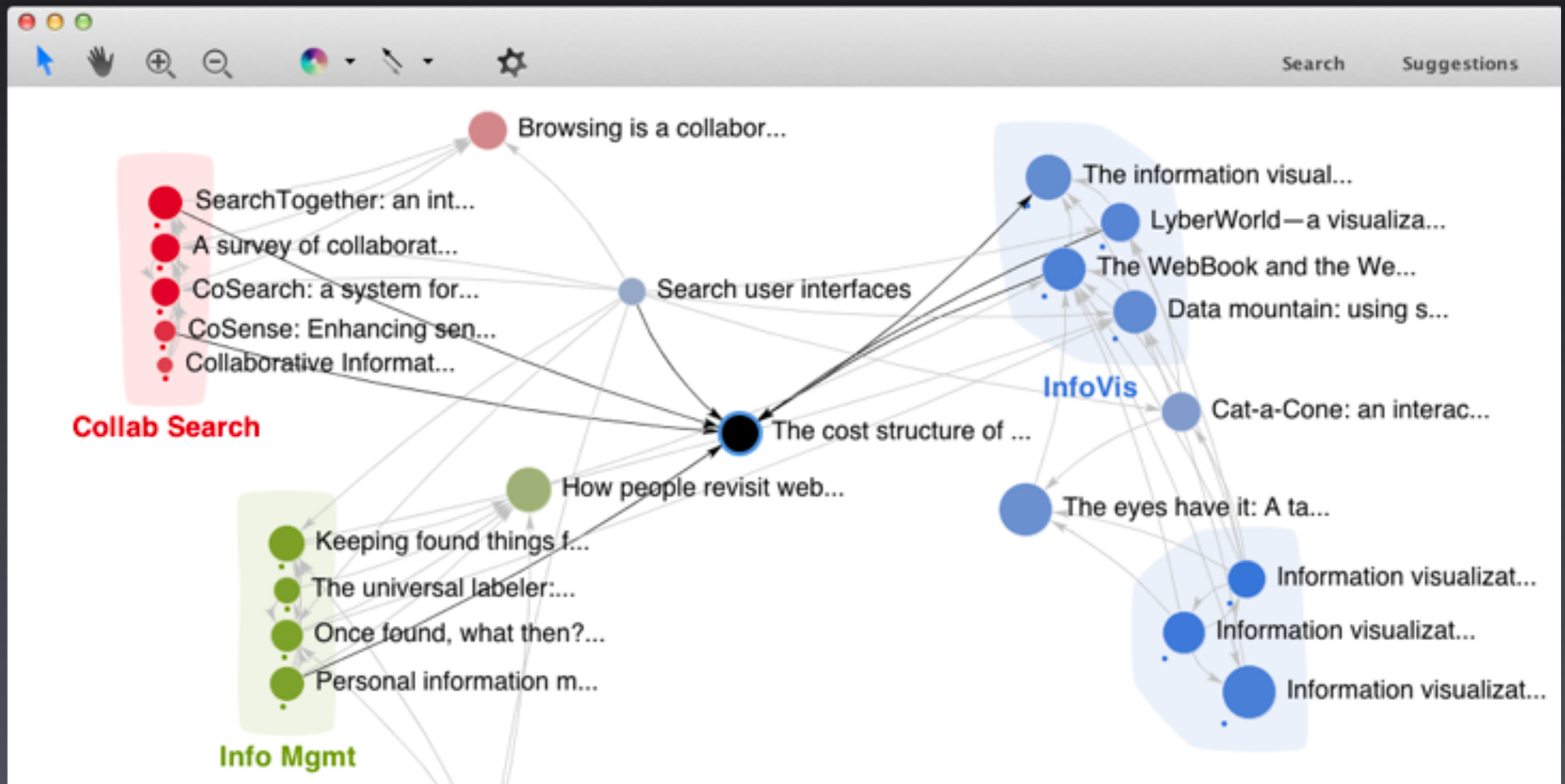


Key Ideas (Recap)



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

We developed **Argo Scholar** to replace Apolo

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, nkittur, 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.

Author Keywords

Sensemaking, large network, Belief Propagation

ACM Classification Keywords

H.3.3 Information Storage and Retrieval: Relevance feedback; H.5.2 Information Interfaces and Presentation: User

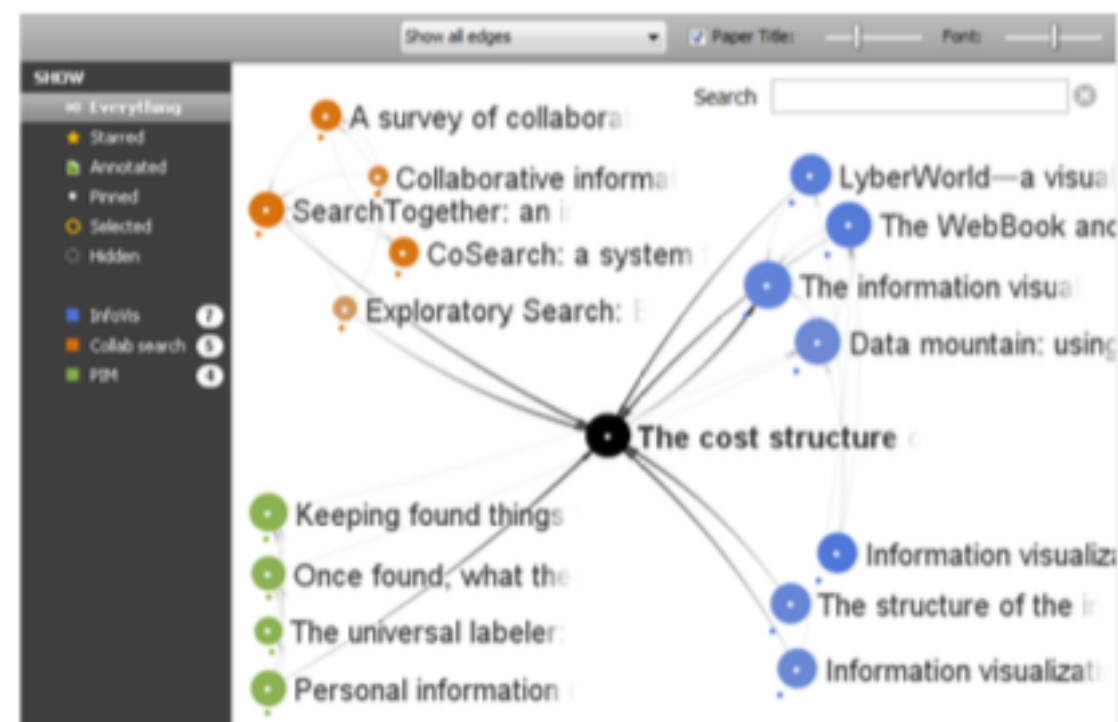


Figure 1. Apolo displaying citation network data around the article *The Cost Structure of Sensemaking*. The user gradually builds up a mental model of the research areas around the article by manually inspecting some neighboring articles in the visualization and specifying them as exemplar articles (with colored dots underneath) for some ad hoc groups, and instructs Apolo to find more articles relevant to them.

representation or schema of an information space that is useful for achieving the user's goal [31]. For example, a scientist interested in connecting her work to a new domain must build up a mental representation of the existing literature in the new domain to understand and contribute to it

NetProbe:

Fraud Detection in Online Auction

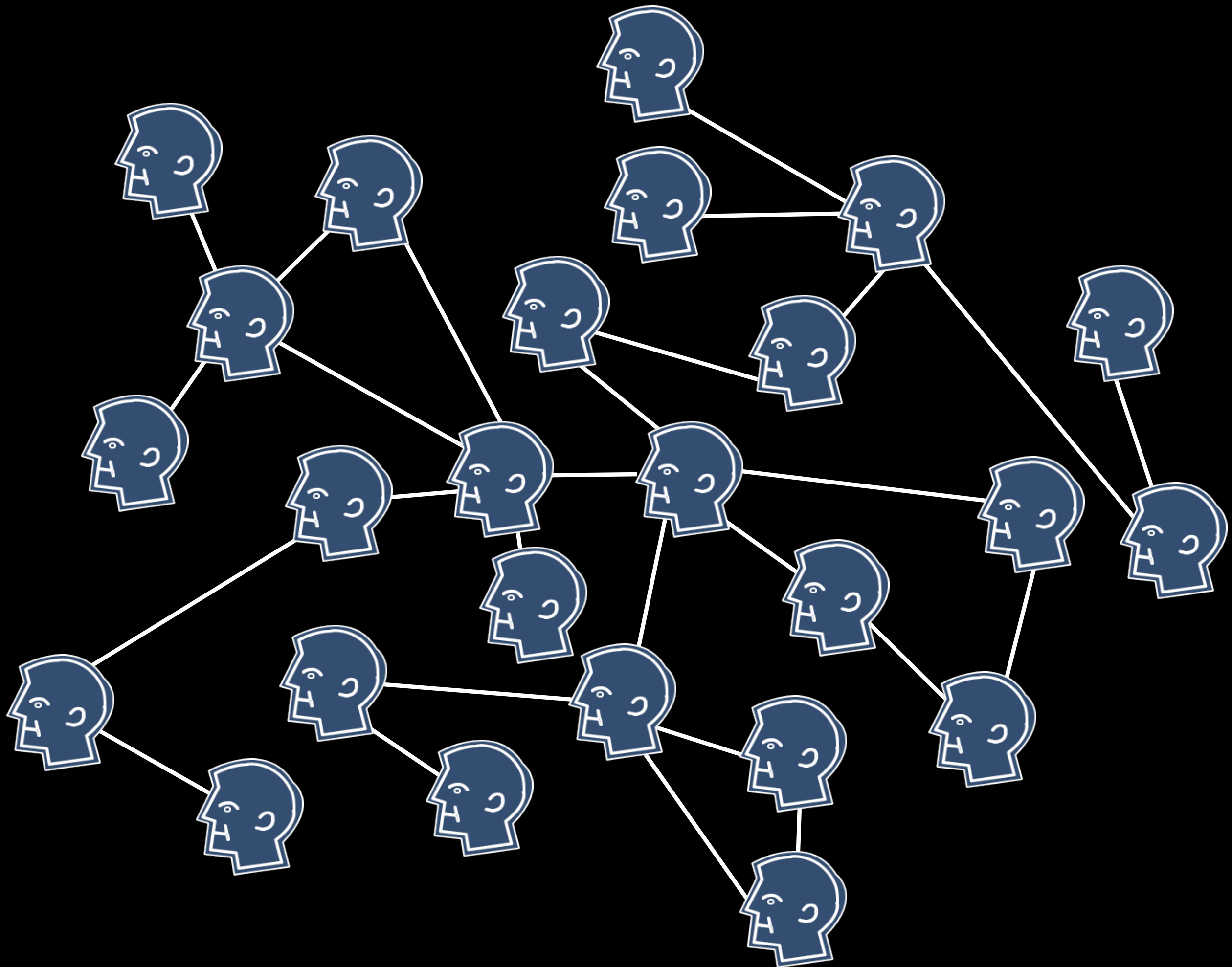


NetProbe: The Problem

Find **bad sellers** (fraudsters) on eBay who don't deliver their items

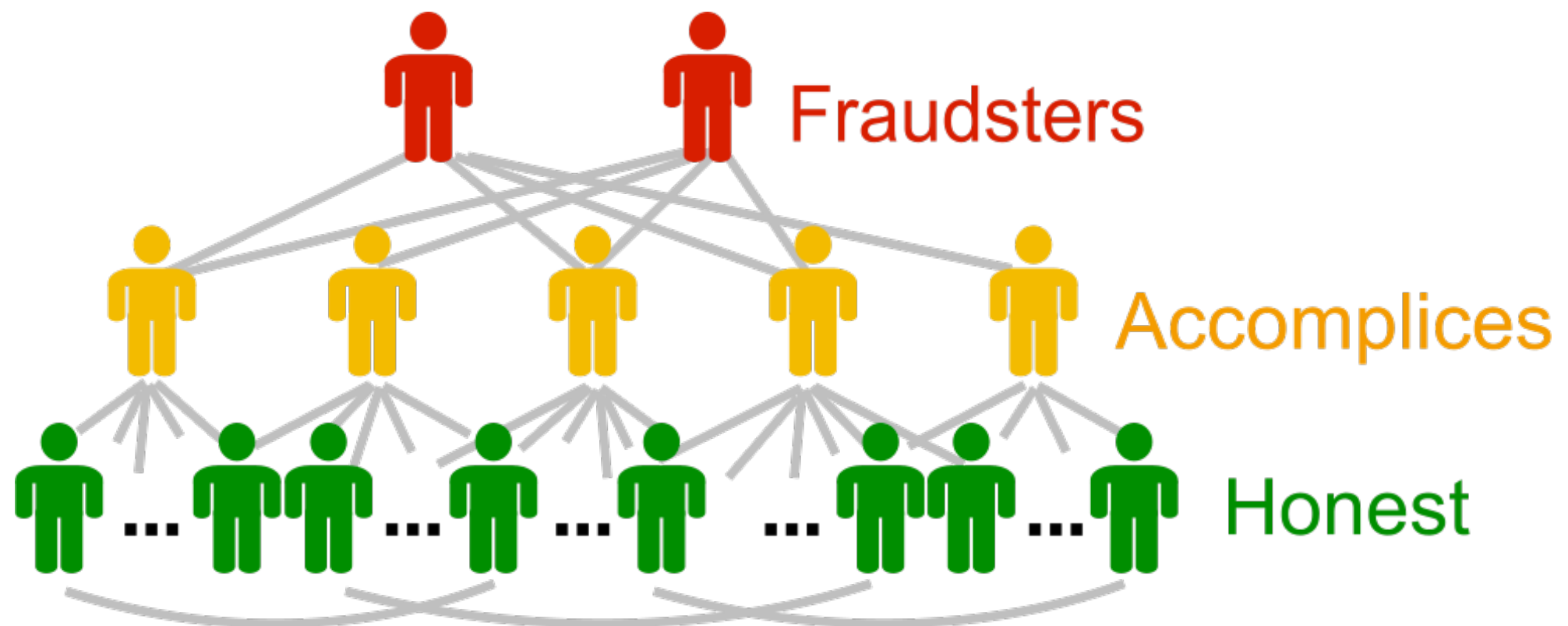


Non-delivery fraud is a common auction fraud



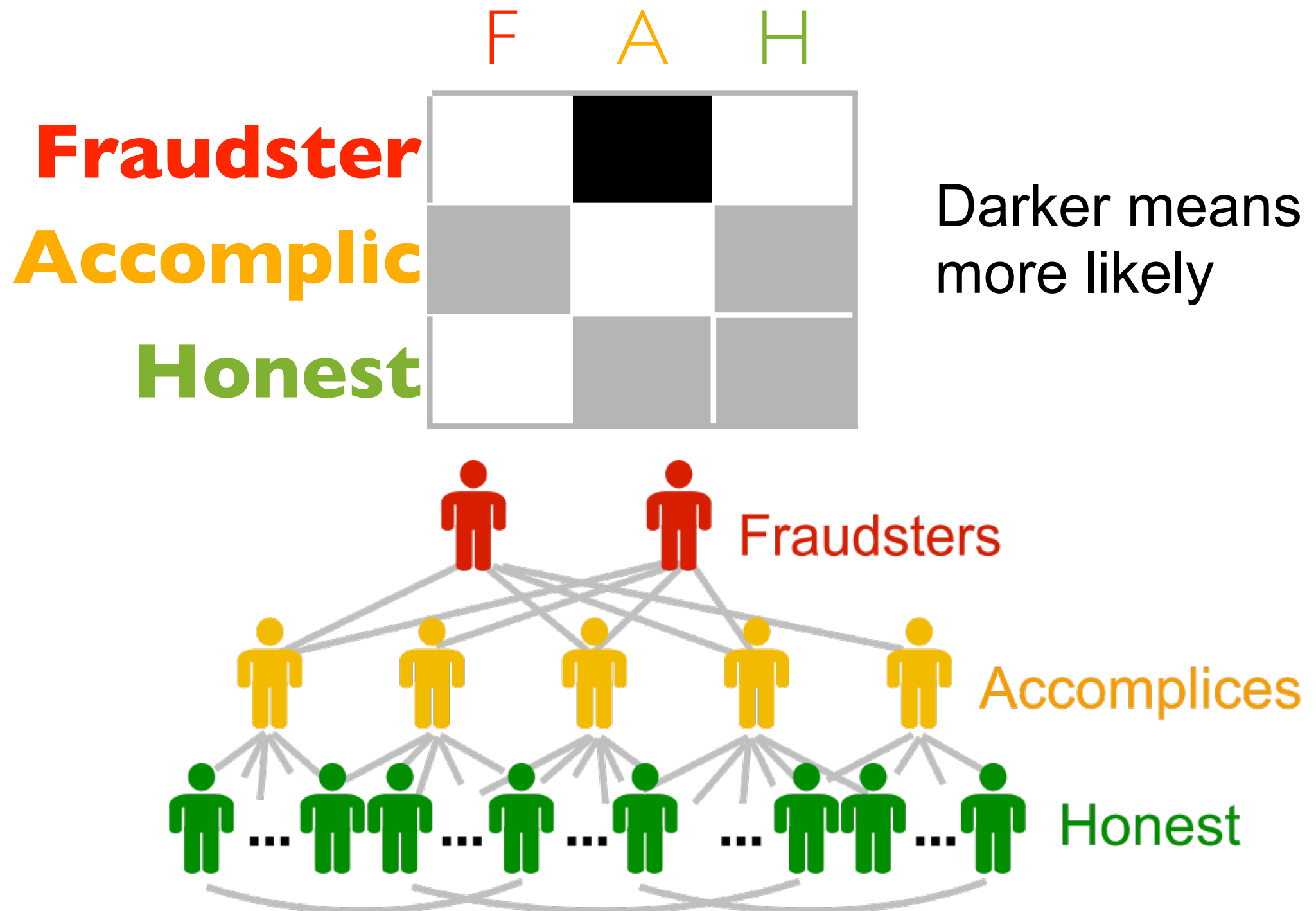
NetProbe: Key Ideas

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

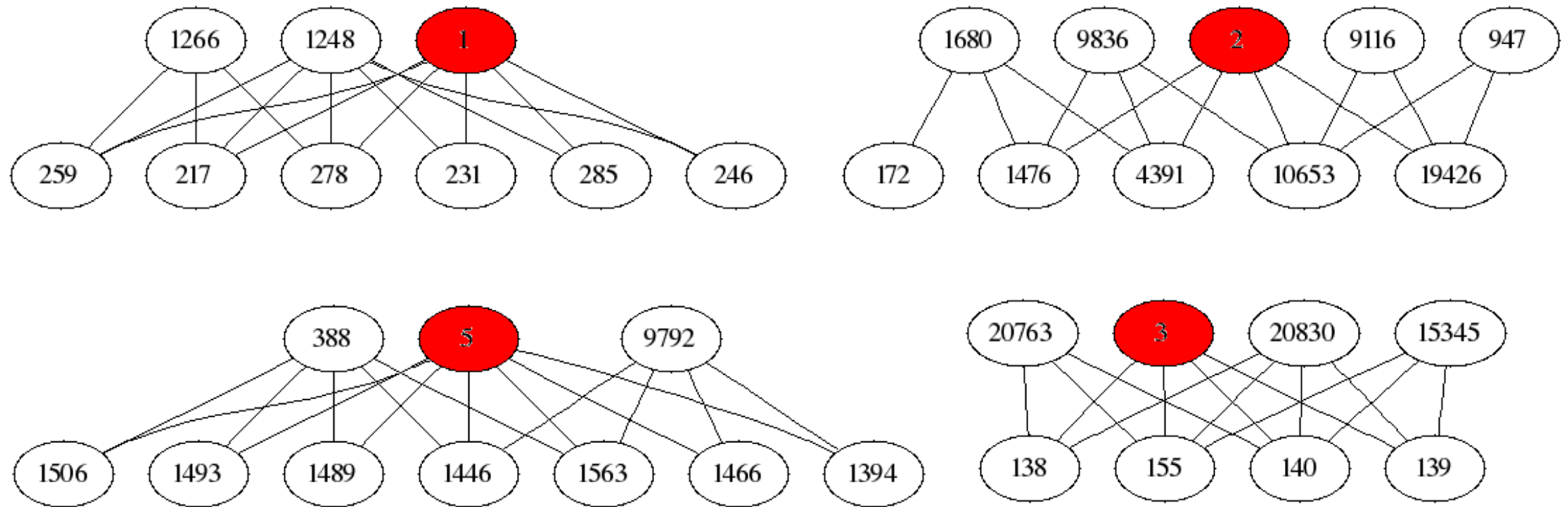


NetProbe: Key Ideas

Use Belief Propagation



NetProbe: Main Results





THE WALL STREET JOURNAL.



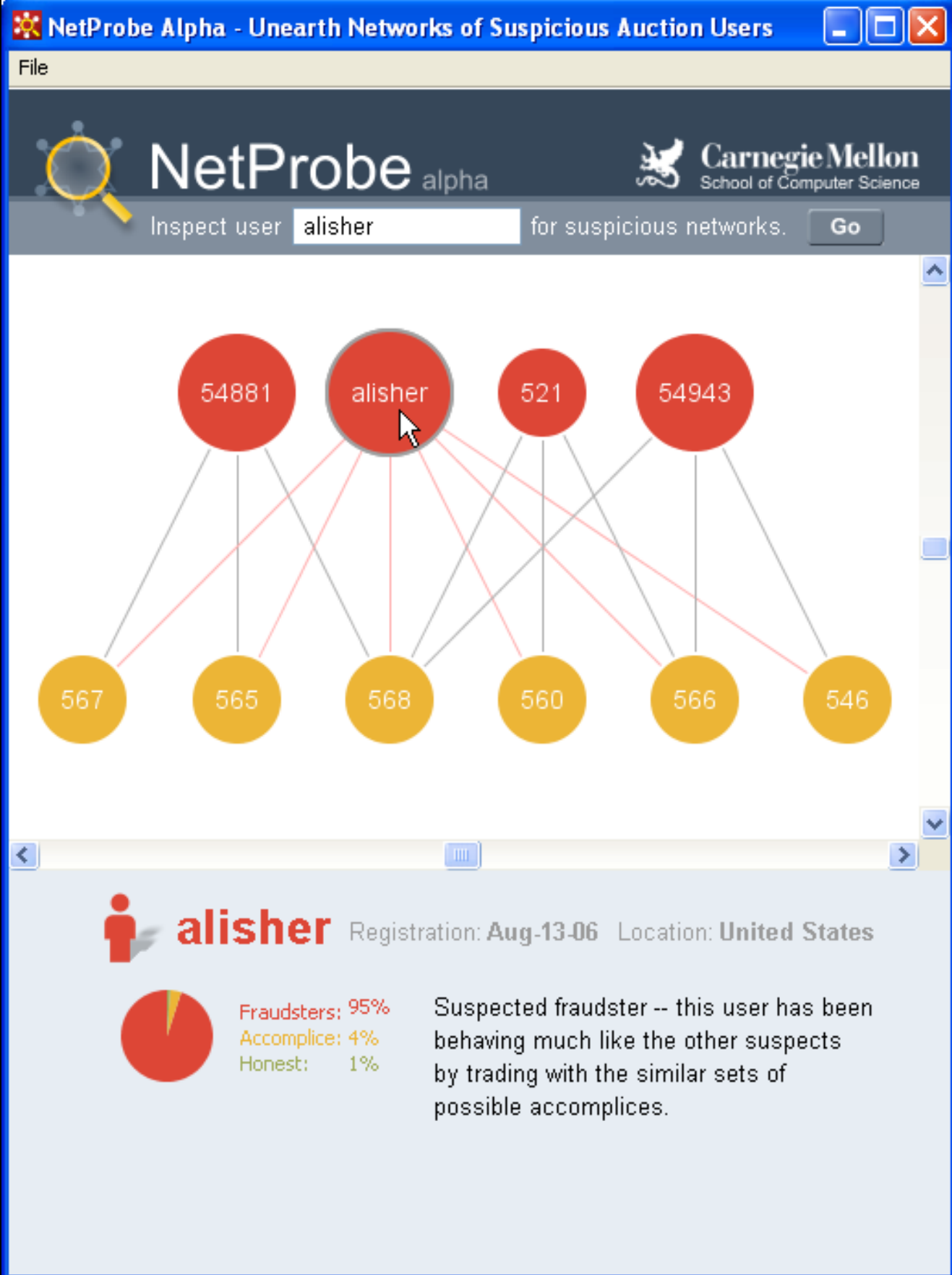
PITTSBURGH
TRIBUNE-REVIEW



Symantec™

“Belgian Police”





What did **NetProbe** go through?

Collection

Scraping (built a “scraper”/“crawler”)

Cleaning

Integration

Analysis

Design detection algorithm

Visualization

Presentation

Paper, talks, lectures

Dissemination

Not released

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 executes nearly doubly fast as compared to NetProbe, while retaining over 99% of its accuracy.

Categories and Subject Descriptors

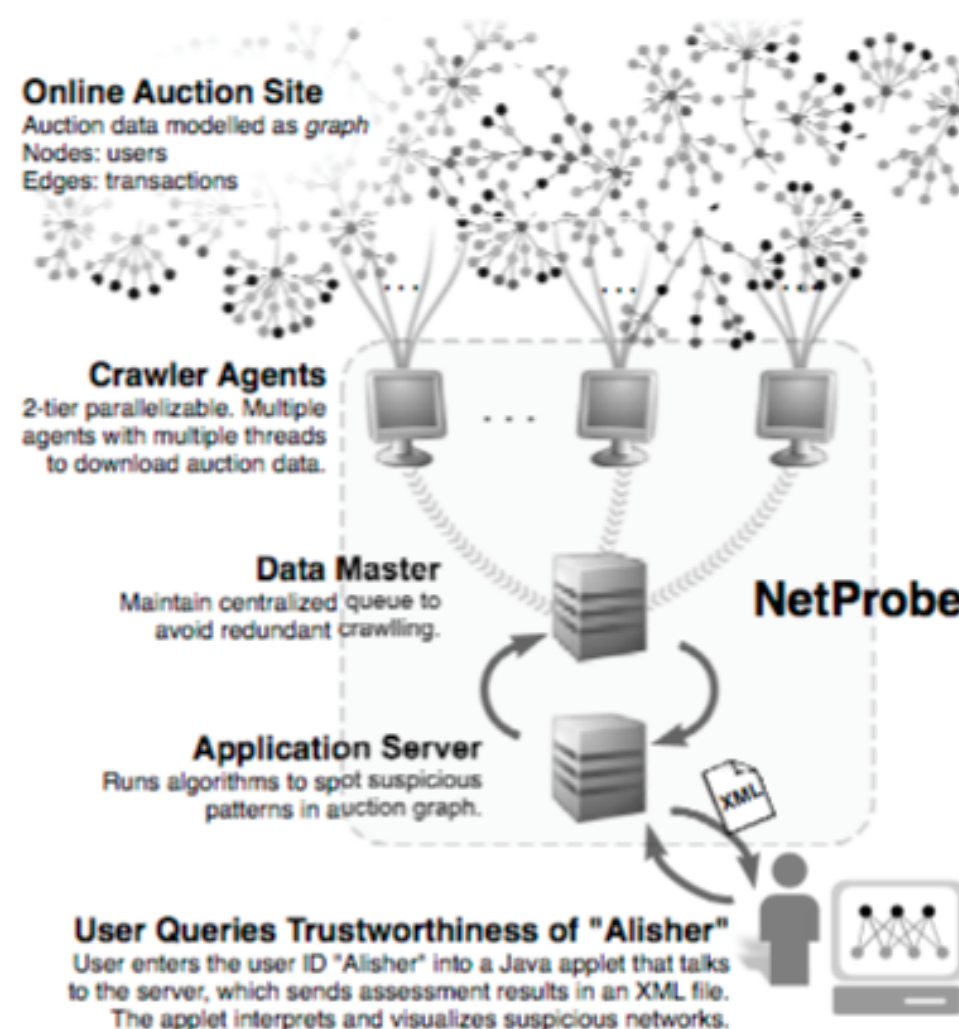


Figure 1: Overview of the NetProbe system

1. INTRODUCTION

Homework 1

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

- Simple “End-to-end” analysis
- Collect movie data via API
 - Store in SQLite database
- Create co-actor network from data
- Analyze, using SQL queries (e.g., create graph’s degree distribution)