Big Data Analytics Building Blocks; Simple Data Storage (SQLite)

Duen Horng (Polo) Chau
Georgia Tech

CSE6242 / CX4242

Aug 21, 2014

Partly based on materials by
Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos

What is Data & Visual Analytics?

No formal definition!

Polo's definition:
the *interdisciplinary* science of combining
computation techniques and
interactive visualization
to transform and model data to aid
discovery, decision making, etc.

What are the “ingredients”?

Need to worry (a lot) about storage, complex system design, scalability of algorithms, visualization techniques, etc.

Used to be “simpler” before the big data era (why?).

What is **big data**? Why care?

- **Many companies**' businesses are based on big data (Google, Facebook, Amazon, Apple, Symantec, LinkedIn, and many more)
- **Web search**
  - Rank webpages (PageRank algorithm)
  - Predict what you're going to type
- **Advertisement** (e.g., on Facebook)
  - Infer users' interest; show relevant ads
  - Infer what you like, based on what your friends like
- **Recommendation systems** (e.g., Netflix, Pandora, Amazon)
- Online education
- Health IT: patient records (EMR)
- Bio and Chemical modeling:
- Finance
- Cybersecruity
- Internet of Things (IoT)

Good news! Many big data jobs

- What jobs are hot?
- **“Data scientist”**
- Emphasize breadth of knowledge
- This course helps you learn some important skills

Big data analytics process and building blocks

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

Building blocks, not “steps”

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

- **Can skip some**
- **Can go back (two-way street)**
- Examples
  - Data types inform visualization design
  - Data informs choice of algorithms
  - Visualization informs data cleaning (dirty data)
  - Visualization informs algorithm design (user finds that results don’t make sense)

How big data affects the process?

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

- The **4V** of big data (now **5V**: **Value**)
  - **Volume**: “billions”, “petabytes” are common
  - **Velocity**: think Twitter, fraud detection, etc.
  - **Variety**: text (webpages), video (e.g., youtube), etc.
  - **Veracity**: uncertainty of data

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Schedule

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

Two analytics examples

NetProbe:
Fraud Detection in Online Auction

WWW 2007

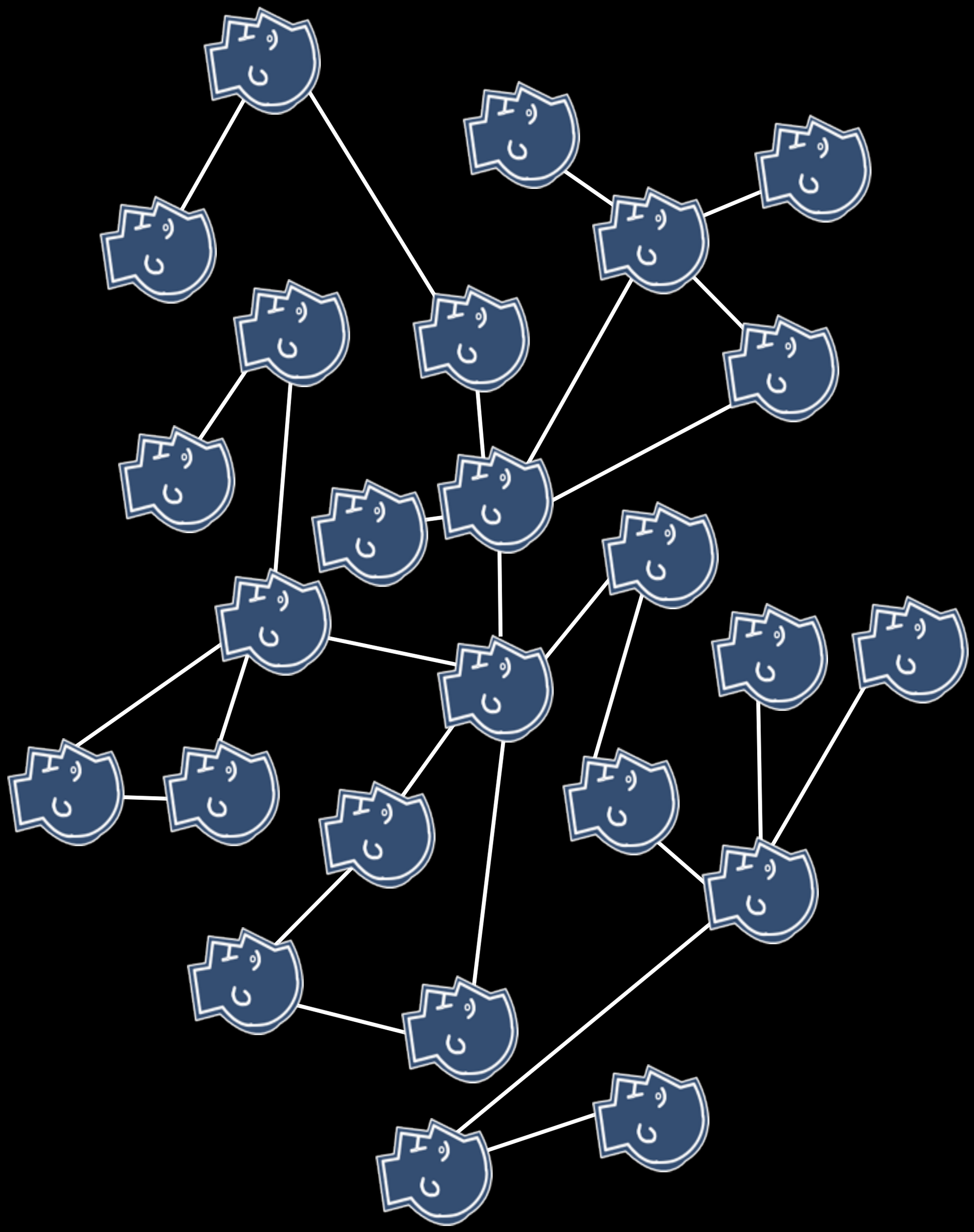
http://www.cs.cmu.edu/~dchau/papers/p201-pandit.pdf

NetProbe: The Problem

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

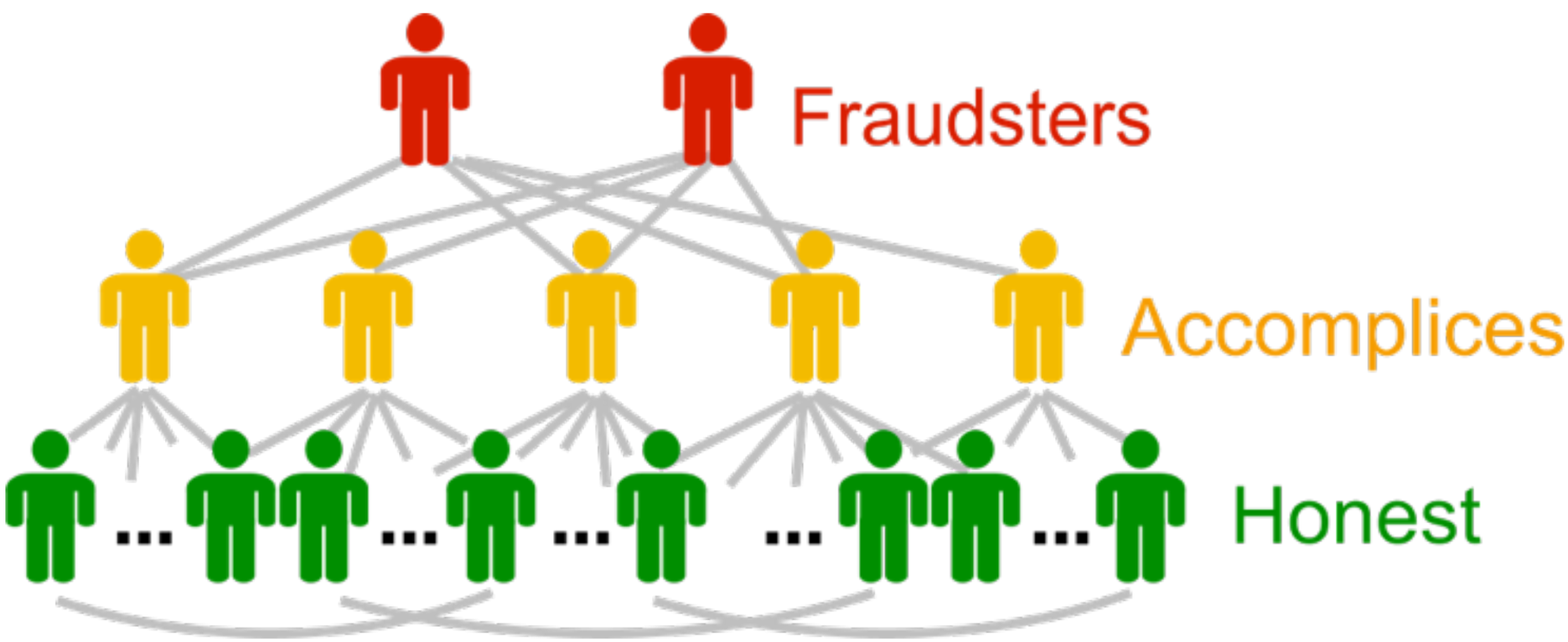
Auction fraud is **#3** online crime in 2010

source: www.ic3.gov



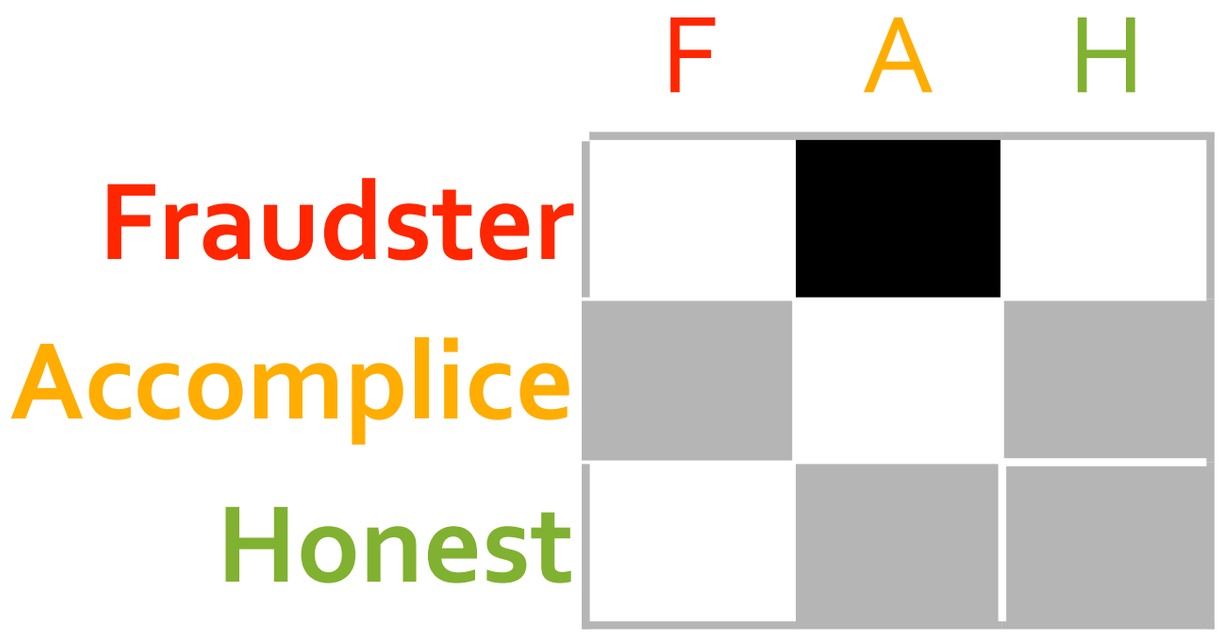
NetProbe: Key Ideas

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

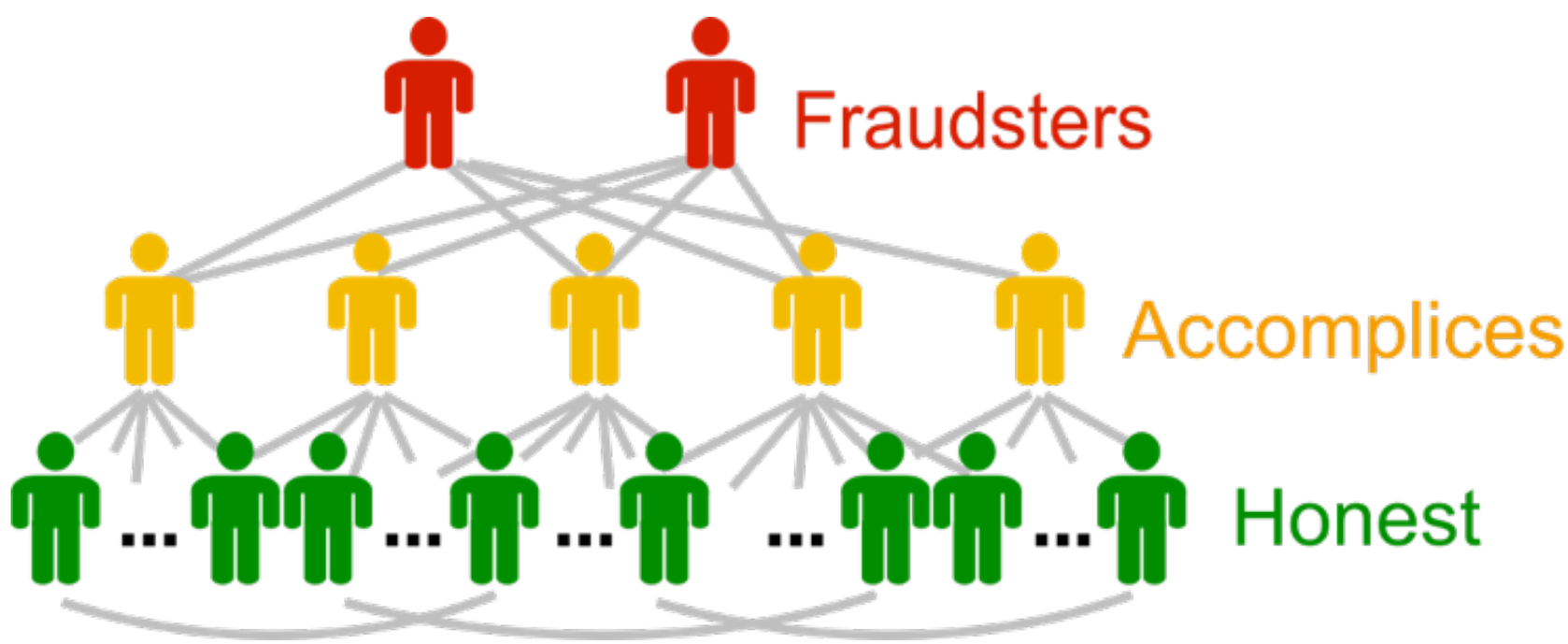


NetProbe: Key Ideas

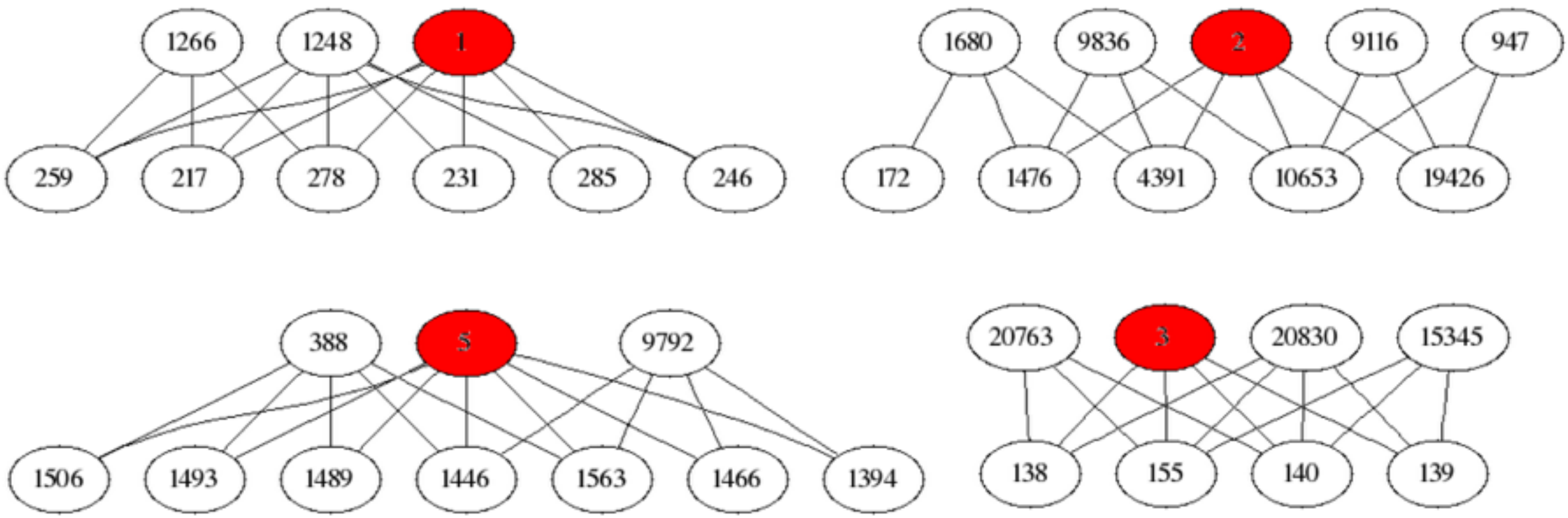
Use **Belief Propagation**



Darker means more likely



NetProbe: Main Results



ebay

USA TODAY

CNNMoney.com
A Service of CNN, Fortune & Money

THE WALL STREET JOURNAL.

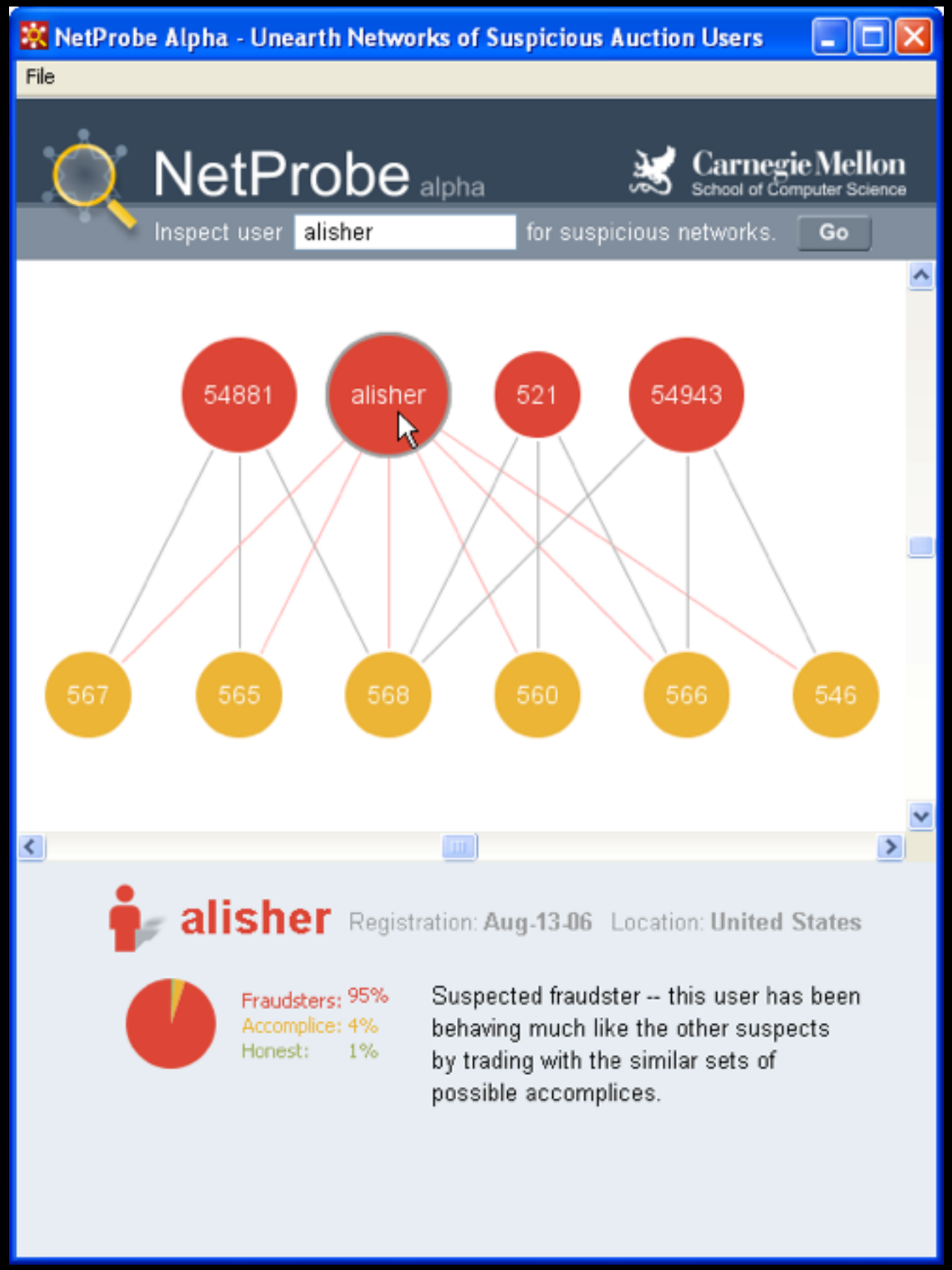
MSNBC

KDKA 2

PITTSBURGH TRIBUNE-REVIEW

Symantec

“Belgian Police”


NetProbe Alpha - Unearth Networks of Suspicious Auction Users
File
NetProbe alpha
Carnegie Mellon
School of Computer Science
Inspect user
alisher
for suspicious networks.
Go
54881
alisher
521
54943
567
565
568
560
566
546
alisher
Registration: Aug-13-06
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.

What analytics process does **NetProbe** go through?

Collection	Scraping (built a “scraper”/“crawler”)
Cleaning	
Integration	
Analysis	Design detection algorithm
Visualization	
Presentation	Paper, talks, lectures
Dissemination	Not released

Discovr movie app


search
search
about
help
email
share
tweet
Fantastic Four
Forever Young
The Incredible Hulk
Skyfall
Captain America
Clash of the Titans
Thor
Captain America: The First Avenger
Marvel's The Avengers
Iron Man
Man of Steel
Star Trek Into Darkness
Favorites
In Cinemas
Our Picks
rovi

What analytics process would you go through to build the app?

Collection	IMDB, Rotten tomatoes, youtube
Cleaning	May have duplicate trailers
Integration	
Analysis	Determine which movies are related
Visualization	
Presentation	
Dissemination	Mac app, iOS app

Homework 1 (out next week)

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

- Simple "End-to-end" analysis
- Collect data from Rotten Tomatoes (using API)
  - Movies (Actors, directors, **related** movies, etc.)
  - Store in SQLite database
- Transform data to movie-movie network
- Analyze, using SQL queries (e.g., create graph's degree distribution)
- Visualize, using **Gephi**
- Describe your discoveries

Data Collection, Simple Storage (SQLite) & Cleaning

Today:

Data Collection, Simple Storage (SQLite) & Cleaning

How to get data?

Download (where?)

API

Scrape/Crawl, or from equipment (e.g., sensors)

Low effort

High effort

Data you can just download

Yahoo Finance (csv)

StackOverflow (xml)

Yahoo Music (KDD cup)

Atlanta crime data (csv)

Soccer statistics

Data via API

CrunchBase (database about companies) - JSON

Twitter

Last.fm (Pandora has API?)

Flickr

Facebook

Rotten Tomatoes

iTunes

Data that needs scraping

Amazon (reviews, product info)

ESPN

Google Scholar

(eBay?)

Most popular embedded database in the world

iPhone (iOS), Android, Chrome (browsers), Mac, etc.

Self-contained: one file contains data + schema

Serverless: database right on your computer

Zero-configuration: no need to set up!

http://www.sqlite.org
http://www.sqlite.org/different.html

How does it work?

```
>sqlite3 database.db


sqlite> create table student(ssn integer, name text);
sqlite> .schema
CREATE TABLE student(ssn integer, name text);
```

ssn	name

How does it work?

```
insert into student values(111, "Smith");
insert into student values(222, "Johnson");
insert into student values(333, "Obama");
select * from student;
```

ssn	name
111	Smith
222	Johnson
333	Obama

How does it work?

```
create table takes
(ssn integer, course_id integer, grade integer);
```

ssn	course_id	grade

How does it work?

More than one tables - **joins**

E.g., create roster for this course

ssn	name
111	Smith
222	Johnson
333	Obama

ssn	course_id	grade
111	6242	100
222	6242	90
222	4000	80

How does it work?

```
select name from student, takes
where student.ssn = takes.ssn and
takes.course_id = 6242;
```

ssn	name
111	**Smith**
222	**Johnson**
333	Obama

ssn	course_id	grade
111	6242	100
222	6242	90
222	4000	80

SQL General Form

```
select a1, a2, ... an
from t1, t2, ... tm
where predicate
[order by ....]
[group by ...]
[having ...]
```

Find ssn and GPA for each student

```
select ssn, avg(grade)
from takes
group by ssn;
```

ssn	course_id	grade
111	6242	100
222	6242	90
222	4000	80

ssn	avg(grade)
111	**100**
222	**85**

What if slow?

Build an **index** to speed things up.
SQLite's indices use **B-tree** data structure.
O(logN) speed for adding/finding/deleting an item

```
create index student_ssn_index on
student(ssn);
```

Homework 1

Write simple scripts to import Rotten Tomatoes data into SQLite, and do some simple queries.

http://developer.rottentomatoes.com/docs/read/json/v10/Movie_Info