CSE 6242 A / CS 4803 DVA

Feb 26, 2013

Graphs III

(Interactive) Applications

Duen Horng (Polo) Chau

Georgia Tech

Partly based on materials by

Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos, Le Song

Last Time

Node centrality

Algorithms

- Fast algorithm to count triangles
  - Per node, and for whole graph
  - Main idea: compute the **first few eigenvectors** of graph's adjacency matrix
  - Useful for computing clustering coefficients, etc.
- (Personalized) PageRank

How to build (interactive) applications?

Will show you example applications that uses **"diffusion-based" algorithms** (i.e., a node *influences* its neighbors)

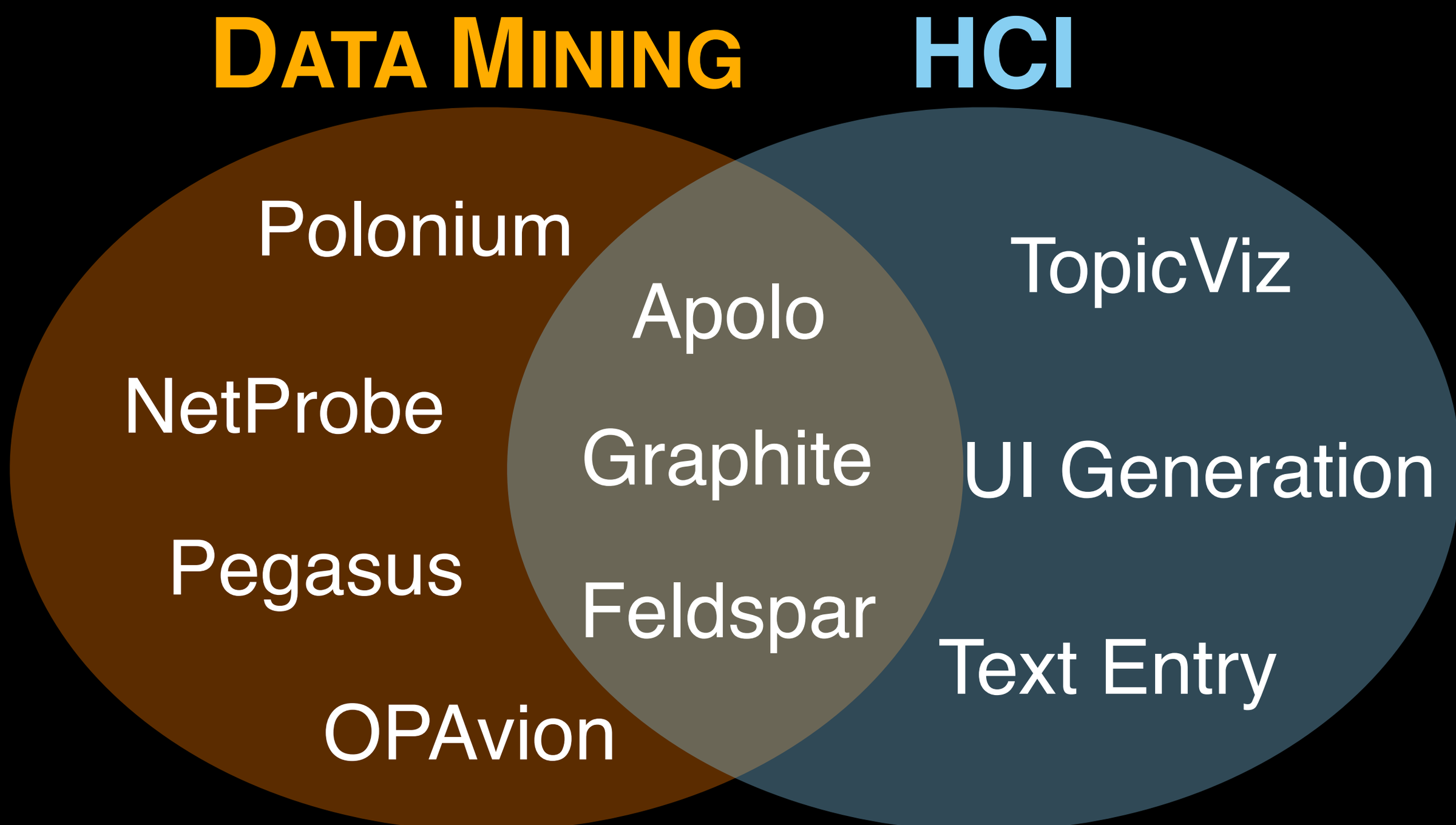
- You've already seen one: **Personalized PageRank** (= Random Walk with Restart)
- Will introduce **Belief Propagation** (powerful inference algorithm)
- "Spreading activation" or "degree of interest" in Human-Computer Interaction (HCI)

How to build (interactive) applications?

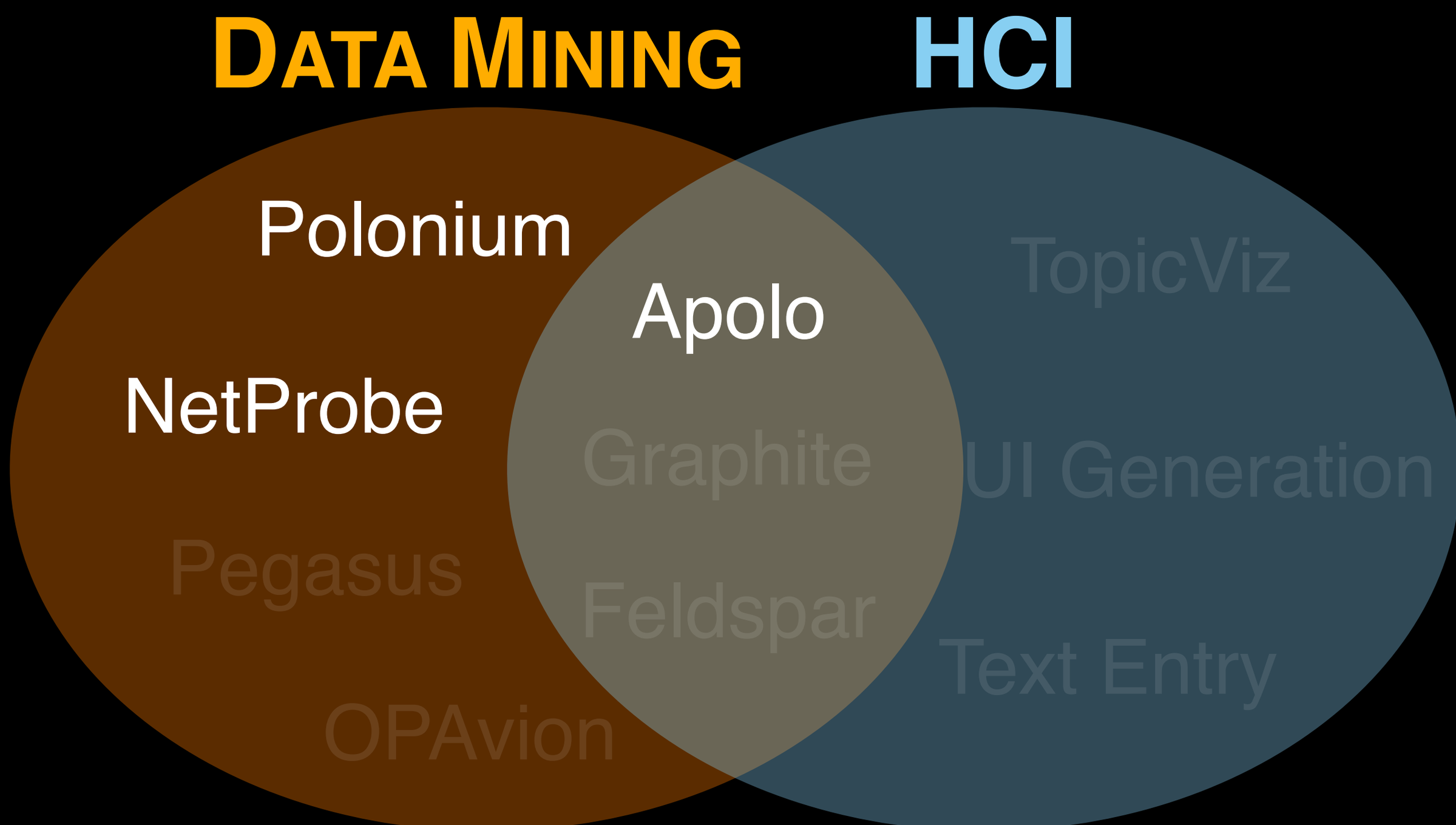
Why diffusion-based algorithms are widely used?

- **Intuitive to interpret**
  uses “network effect”, homophily, etc.
- **Easy to implement**
  Math is relatively simple
- **Fast**
  run time linear to #edges, or better
- **Probabilistic** meaning

Applications to show you



Applications to show you

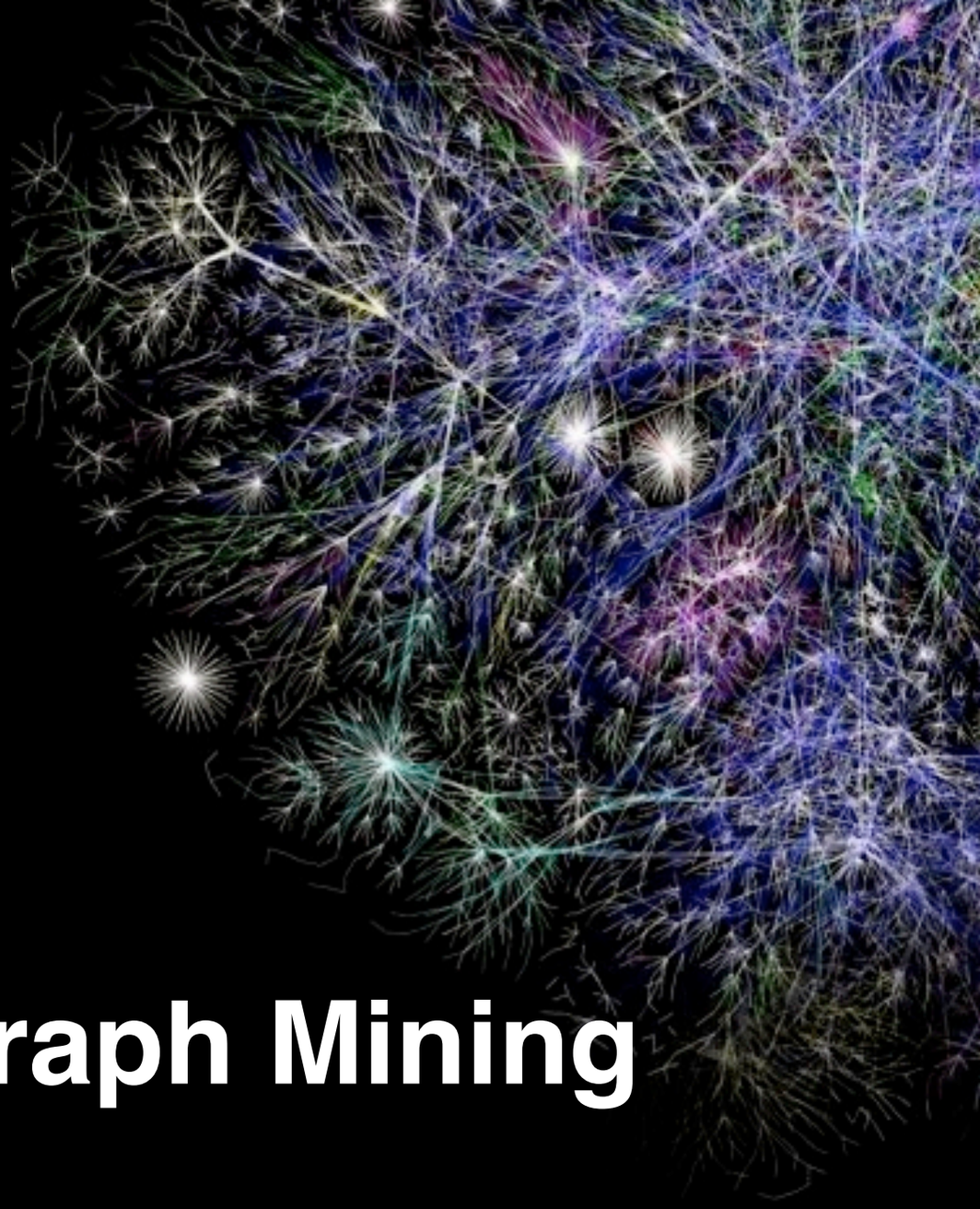


Attention Routing

Attention Routing

Polonium – Find Malware

NetProbe – Find Fraudsters

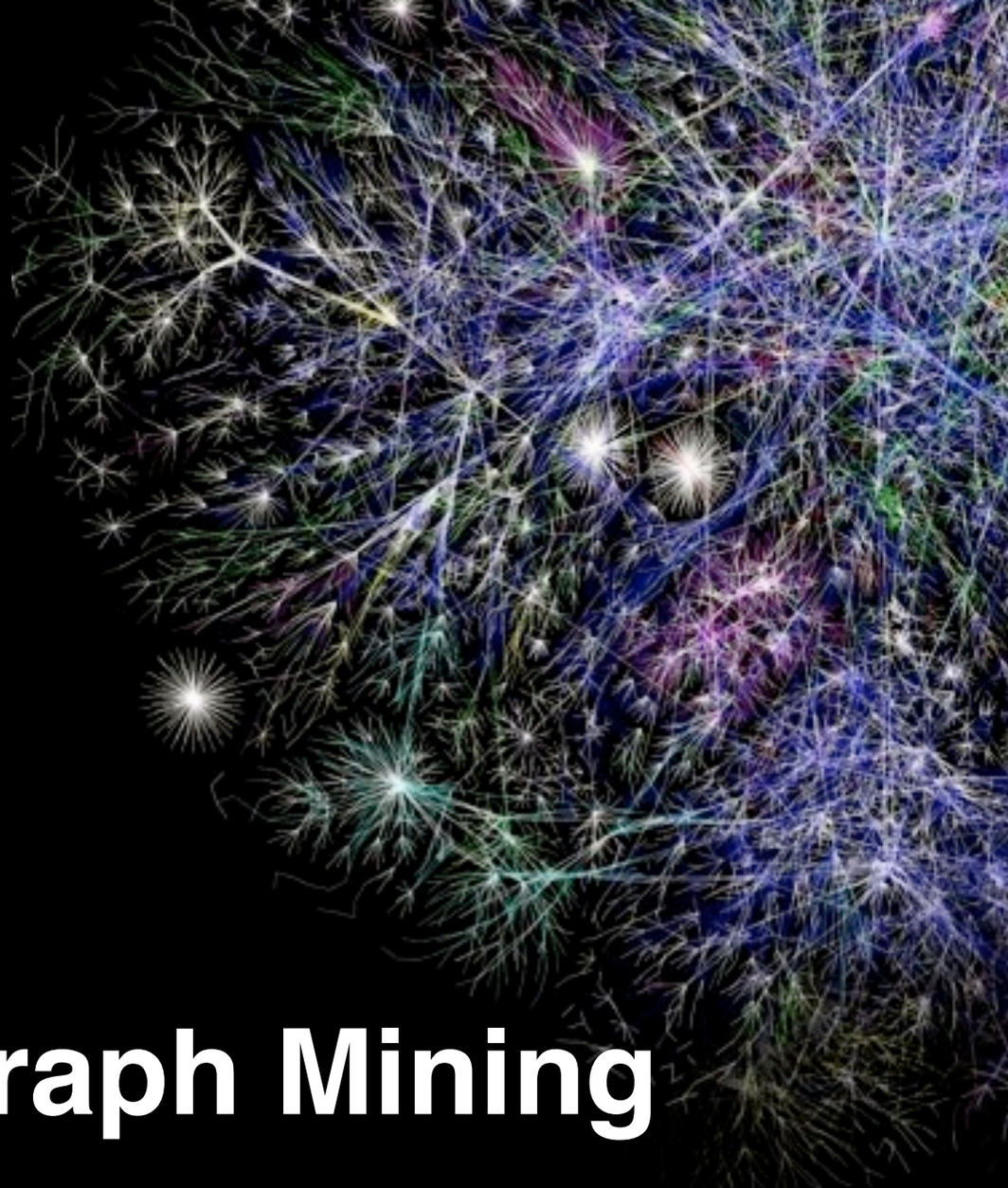


Attention Routing

Polonium – Find Malware

NetProbe – Find Fraudsters

Human-In-The-Loop Graph Mining



Attention Routing

Polonium – Find Malware

NetProbe – Find Fraudsters

Human-In-The-Loop Graph Mining

Apolo – Mixed-Initiative Graph Sensemaking (ML + Vis)

Polonium: Web-Scale Malware Detection

SDM 2011

Polonium: Tera-Scale Graph Mining and Inference for Malware Detection

Typical Malware Detection Method

Signature-based detection

1. Collect malware
2. Generate **signatures**
3. Distribute to users
4. Scan computers for matches

What about "zero-day" malware?

No samples → No signatures → No detection

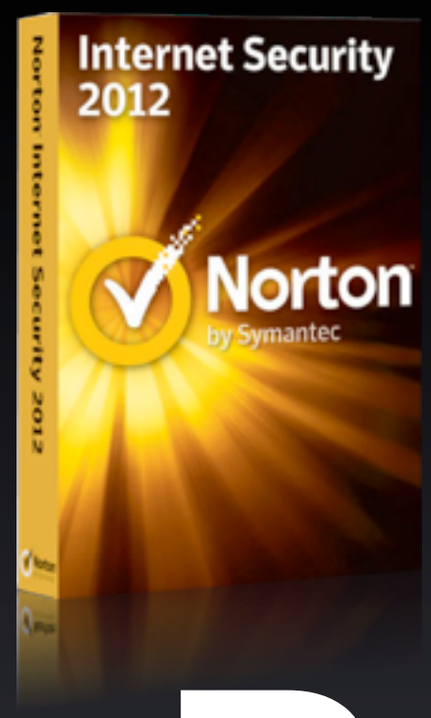
How to detect them early?

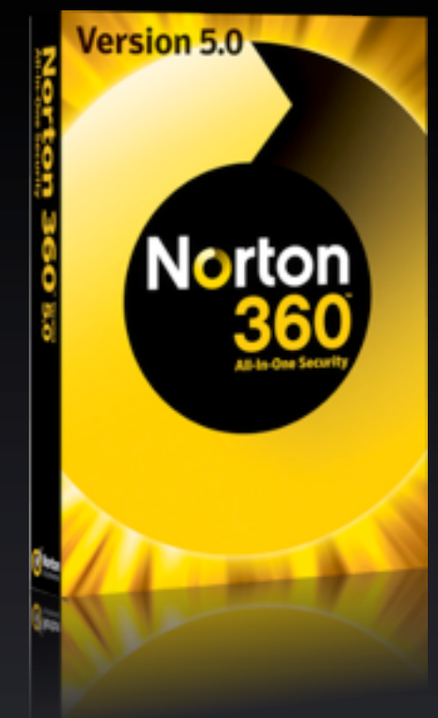
Reputation-Based Detection

Computes **reputation score** for each application

e.g., MSWord.exe

Poor reputation = Malware





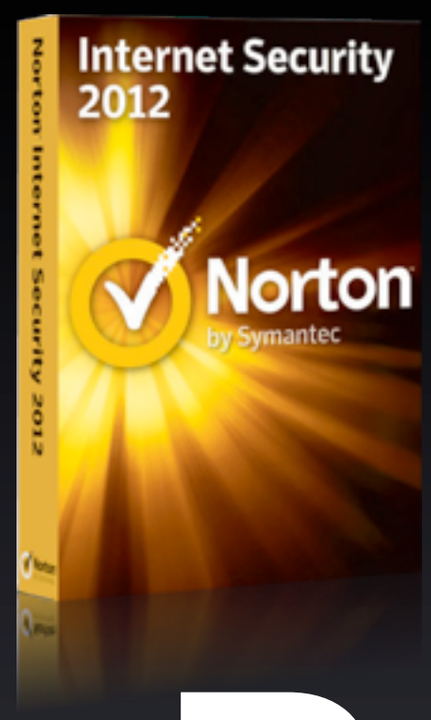
Polonium

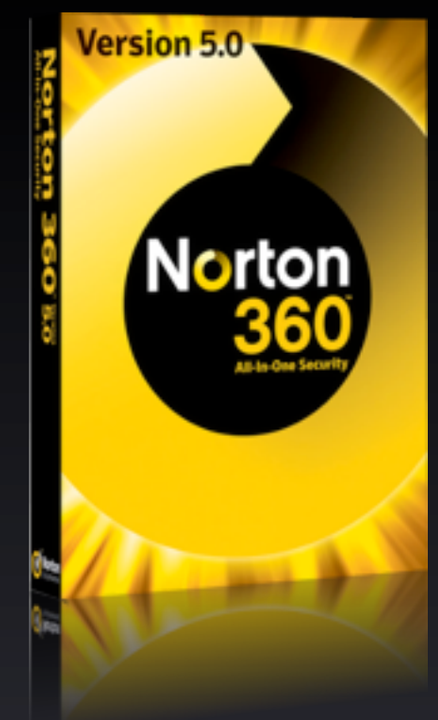
Patent-pending

I led initial **design** and **development**

Serving **120 million** users

Answered **trillions** of queries





Polonium

Propagation of leverage of network influence unearths malware

Patent-pending

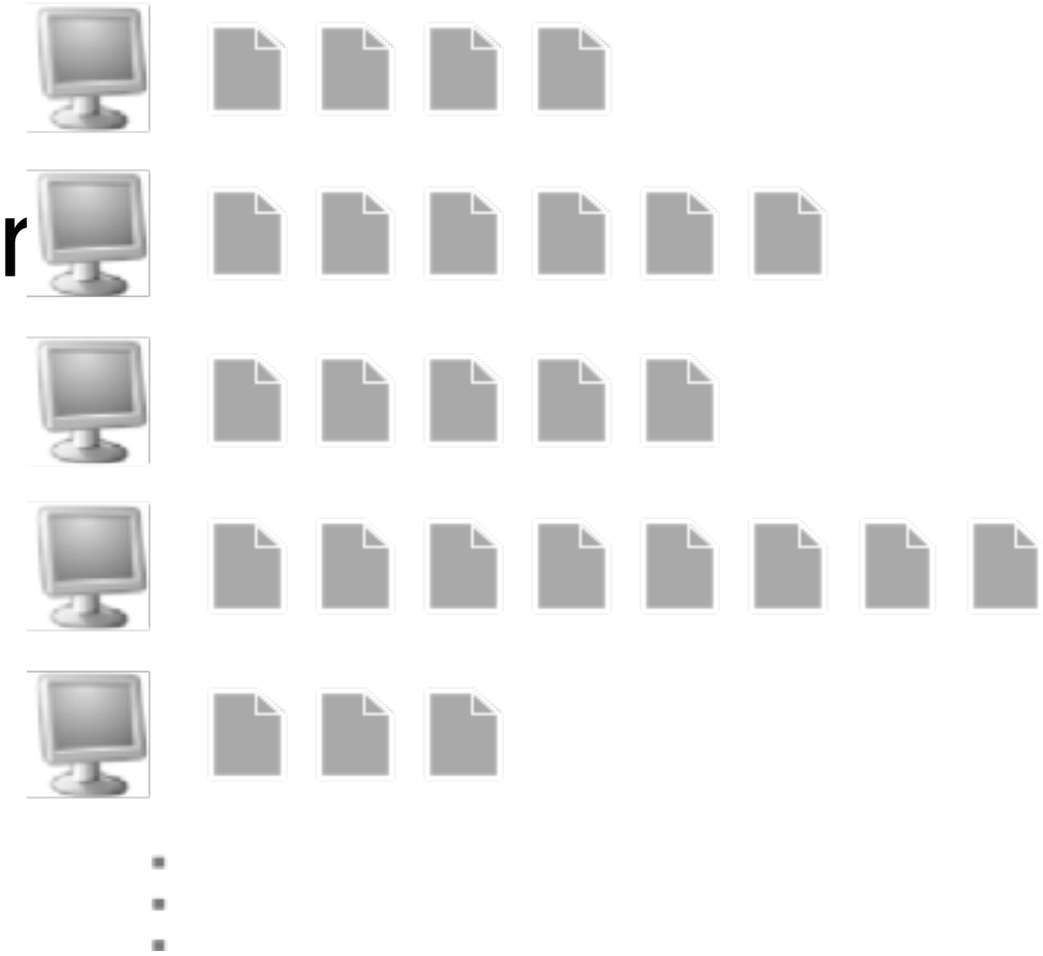
I led initial **design** and **development**

Serving **120 million** users

Answered **trillions** of queries

Polonium works with 60 Terabyte Data

50 million machines anonymously reported their executable **files**



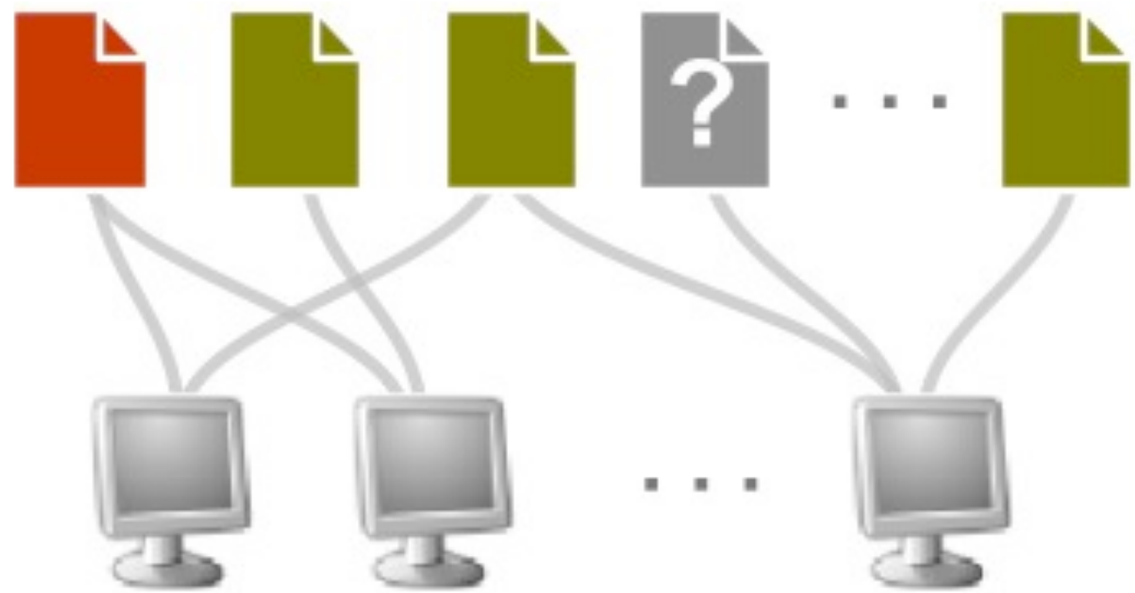
900 million unique files
(Identified by their cryptographic hash values)

Goal: label **malware** and **good** files

Why A Hard Problem?

Existing Research	Polonium
Small dataset	**Huge** dataset (60 terabytes)
Detects **specific** malware (e.g., worm, trojans)	Detects **all** types (needs a general method)
Many false alarms (>10%)	**Strict** (<1%)

Polonium: Problem Definition



Given

Undirected machine-file bipartite graph

37 billion edges , **1 billion** nodes (machines, files)

Some file labels from Symantec (**good** or **bad**)

Find

Labels for all **unknown** files

Where to Get Good and Bad Labels?

Symantec has a **ground truth database** of **known-good** and **known-bad** files

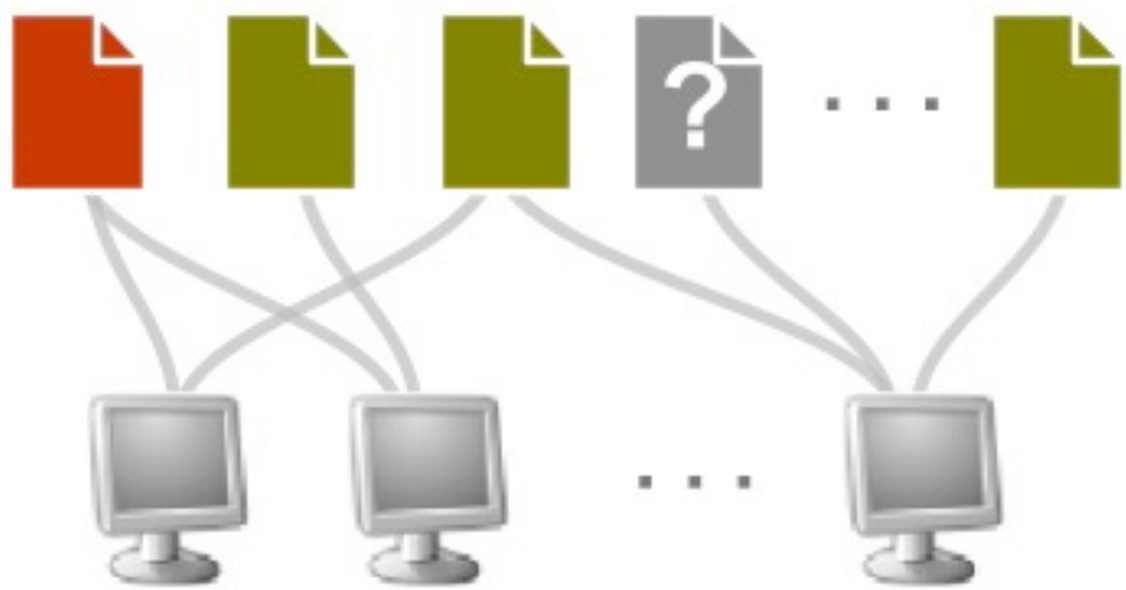
e.g., set **known-good** file's prior to **0.9**

How to Gauge Machine Reputation?

Computed using Symantec's proprietary formula;
a value between 0 and 1

Derived from anonymous aspects of machine's usage and behavior

How to propagate known information to the unknown?



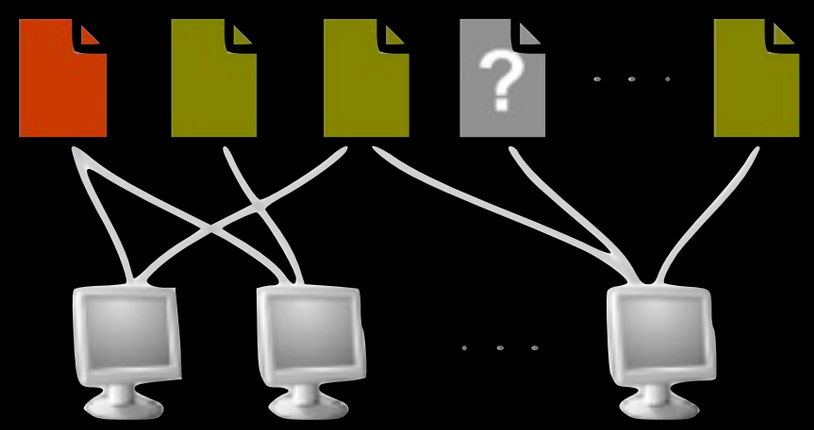
Key Idea: Guilt-by-Association

GOOD files likely appear on **GOOD machines**

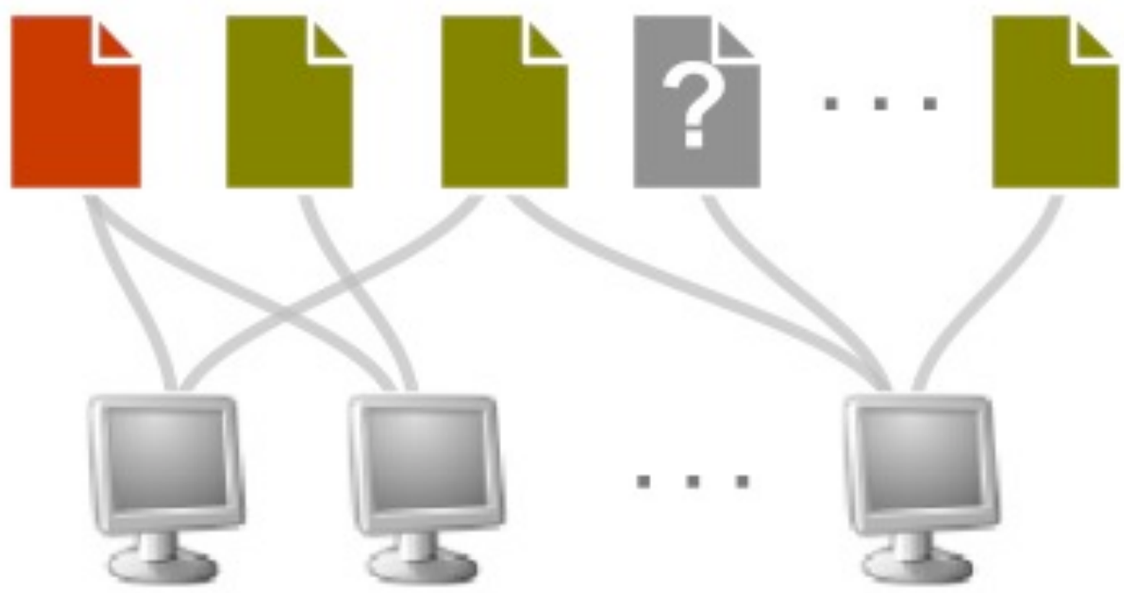
BAD files likely appear on **BAD machines**

Also known as **Homophily**

File \ Machine	Good	Bad
Good	0.9	0.1
Bad	0.1	0.9



How to propagate known information to the unknown?



Adapts **Belief Propagation (BP)**

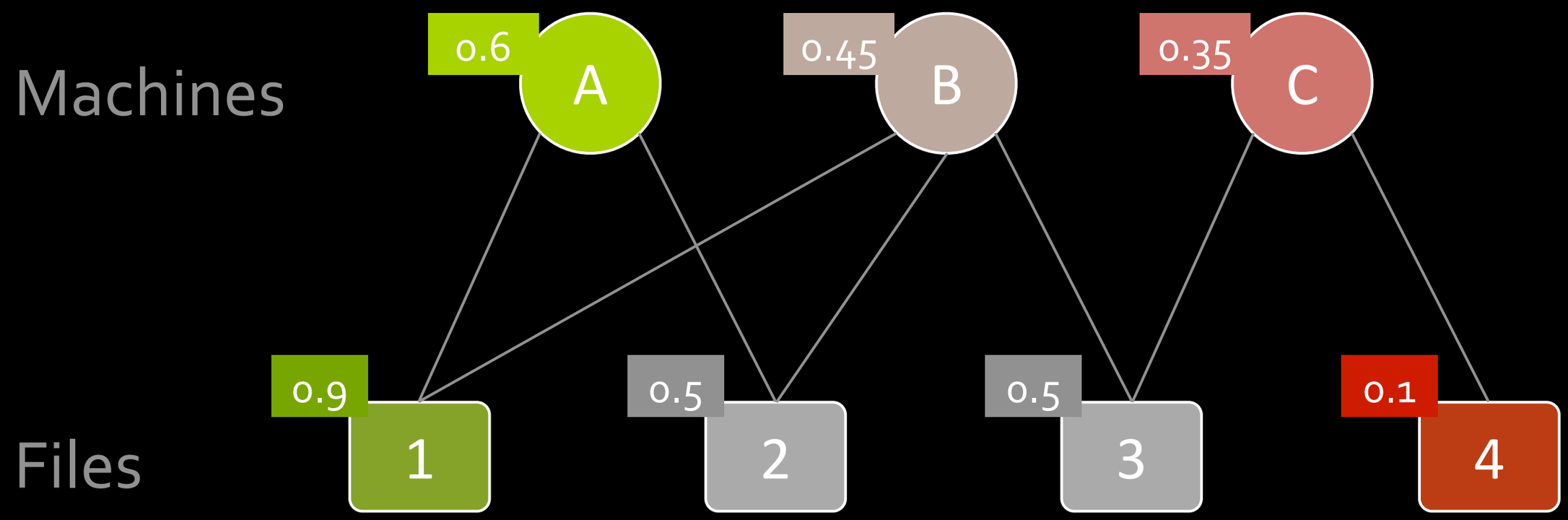
A powerful **inference** algorithm

Used in image processing, computer vision, error-correcting codes, etc.

Example

Propagating Reputation

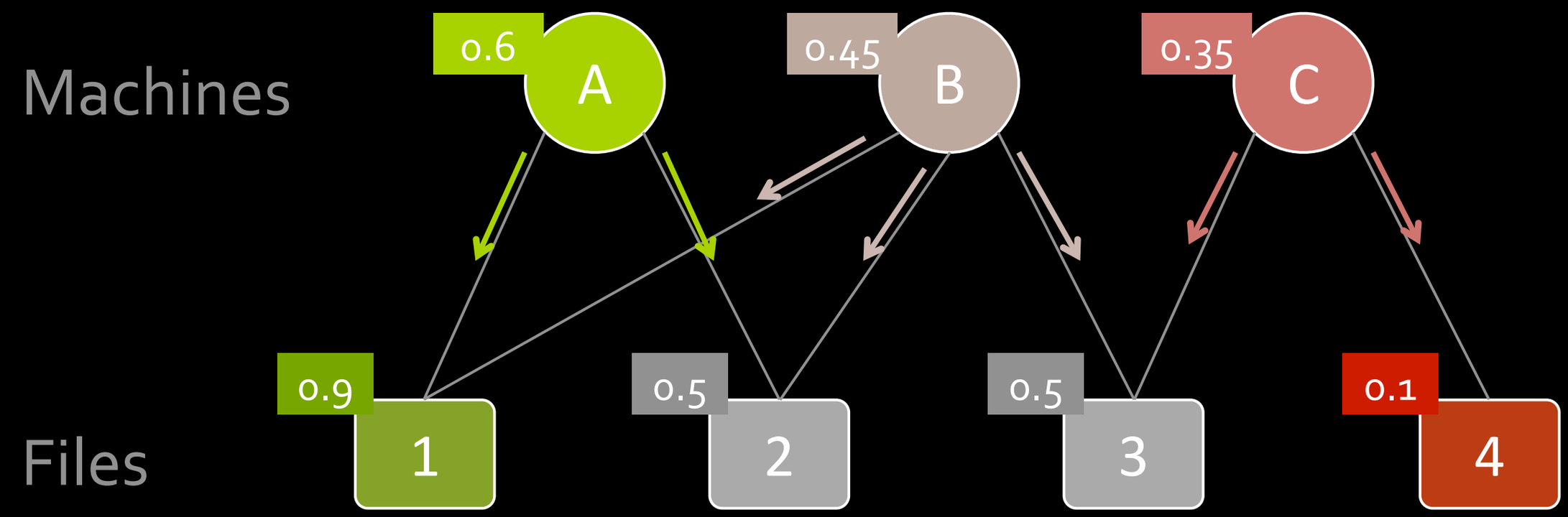
File \ Machine	Good	Bad
Good	0.9	0.1
Bad	0.1	0.9



Example

Propagating Reputation

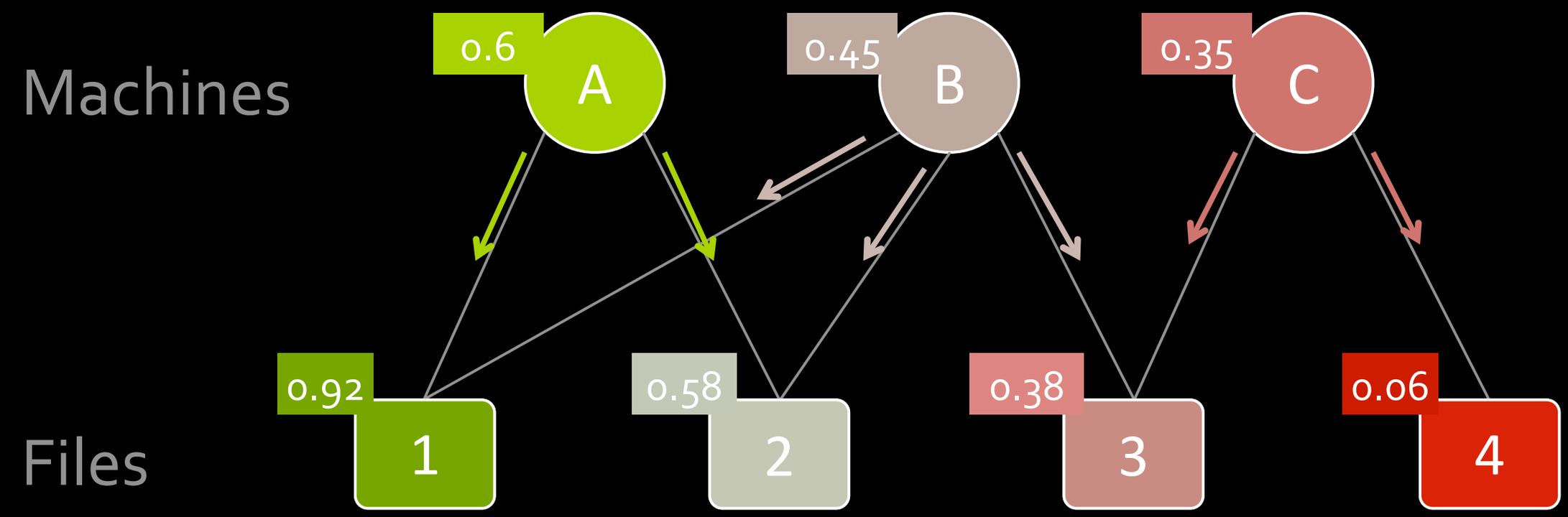
File \ Machine	Good	Bad
Good	0.9	0.1
Bad	0.1	0.9



Example

Propagating Reputation

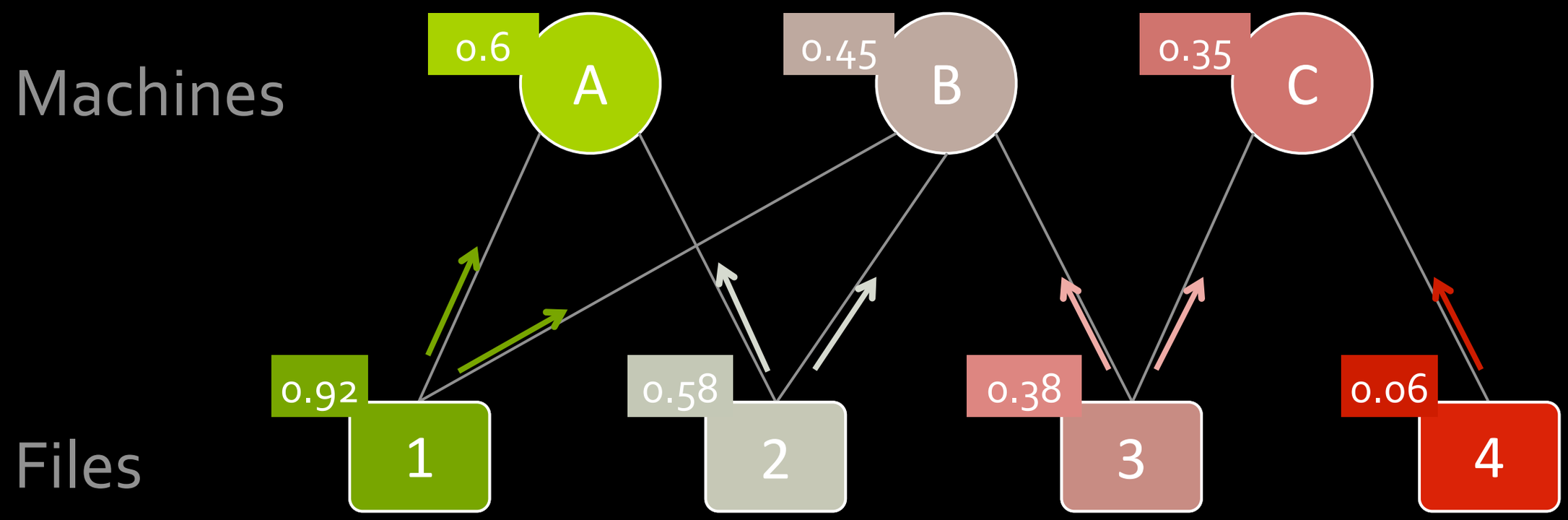
File \ Machine	Good	Bad
Good	0.9	0.1
Bad	0.1	0.9



Example

Propagating Reputation

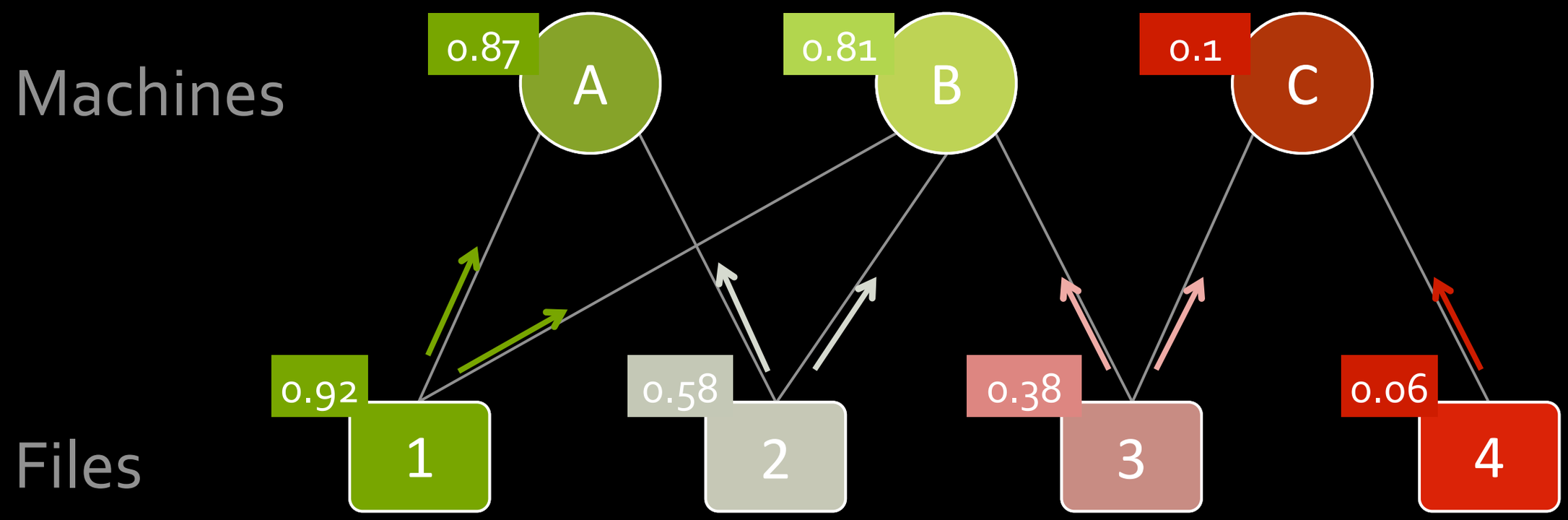
File \ Machine	Good	Bad
Good	0.9	0.1
Bad	0.1	0.9



Example

Propagating Reputation

File \ Machine	Good	Bad
Good	0.9	0.1
Bad	0.1	0.9



Two Equations in Belief Propagation

Details

$$b_i(x_i) = k\phi(x_i) \prod_{x_j \in N(i)} m_{ji}(x_i)$$

$$m_{ij}(x_j) \leftarrow \sum_{x_i \in X} \psi_{ij}(x_i, x_j)\, \phi(x_i) \prod_{k \in N(i) \setminus j} m_{ki}(x_i)$$

Computing Node Belief (Reputation)

Details

$$b_i(x_i) = k\phi(x_i) \prod_{x_j \in N(i)} m_{ji}(x_i)$$

Computing Node Belief (Reputation)

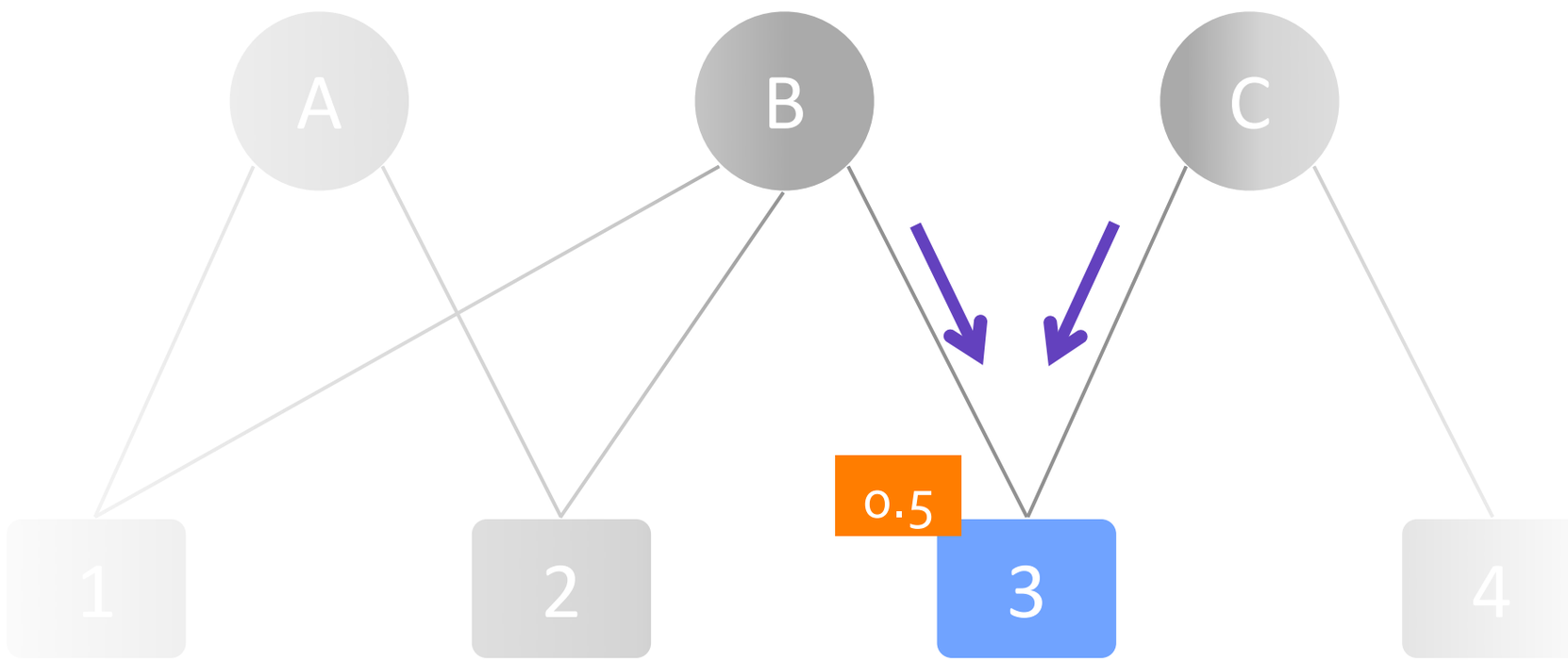
Details

$$b_i(x_i) = k\phi(x_i) \prod_{x_j \in N(i)} m_{ji}(x_i)$$

Belief Prior belief Neighbors' opinions

Computing Node Belief (Reputation)

Details



$$b_i(x_i) = k\phi(x_i) \prod_{x_j \in N(i)} m_{ji}(x_i)$$

Belief Prior belief Neighbors' opinions

Creating Message for Neighbor

Details

$$m_{ij}\left(x_j\right) \leftarrow \sum_{x_i \in X} \psi_{ij}\left(x_i, x_j\right)\ \phi\left(x_i\right) \prod_{k \in N(i) \backslash j} m_{ki}\left(x_i\right)$$

Creating Message for Neighbor

Details

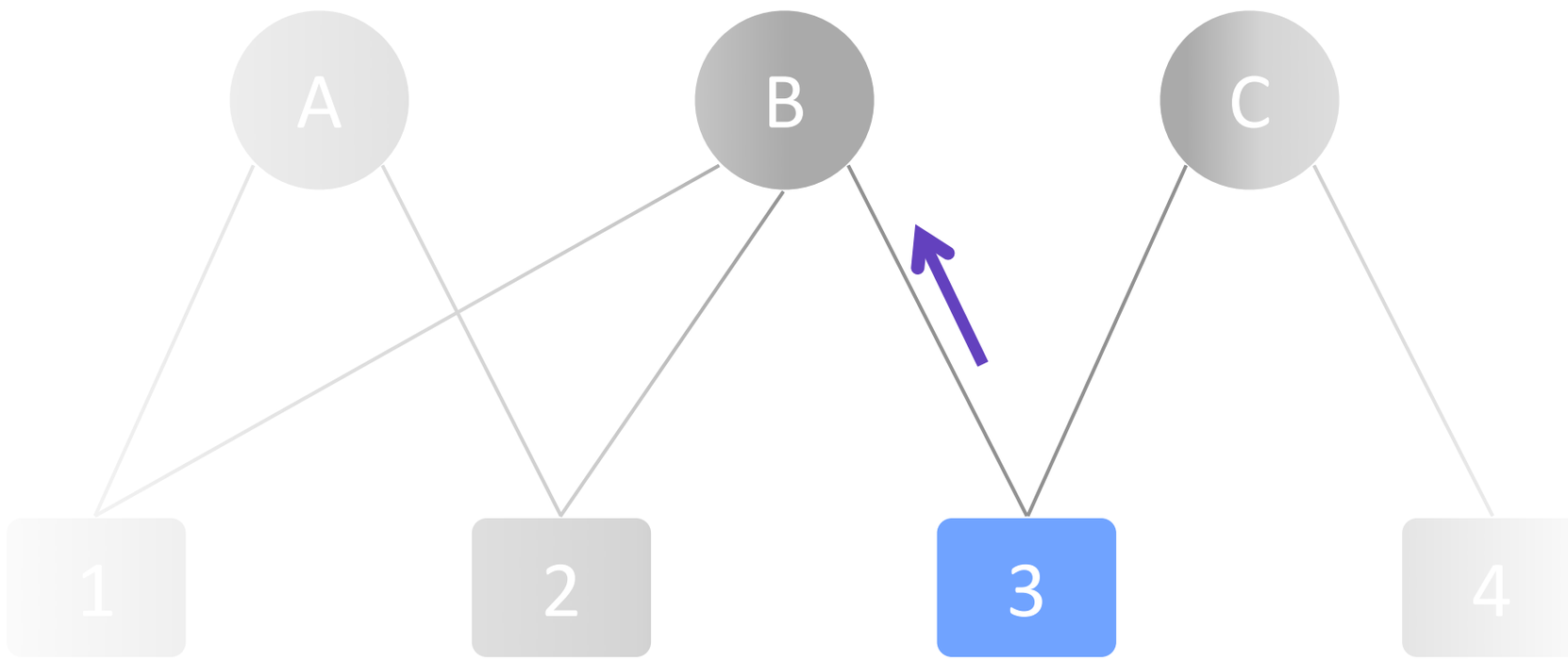
$$m_{ij}(x_j) \leftarrow \sum_{x_i \in X} \psi_{ij}(x_i, x_j) \, \phi(x_i) \prod_{k \in N(i) \setminus j} m_{ki}(x_i)$$

Opinion for neighbor Edge potential Belief

	Good	Bad
Good	0.9	0.1
Bad	0.1	0.9

Creating Message for Neighbor

Details



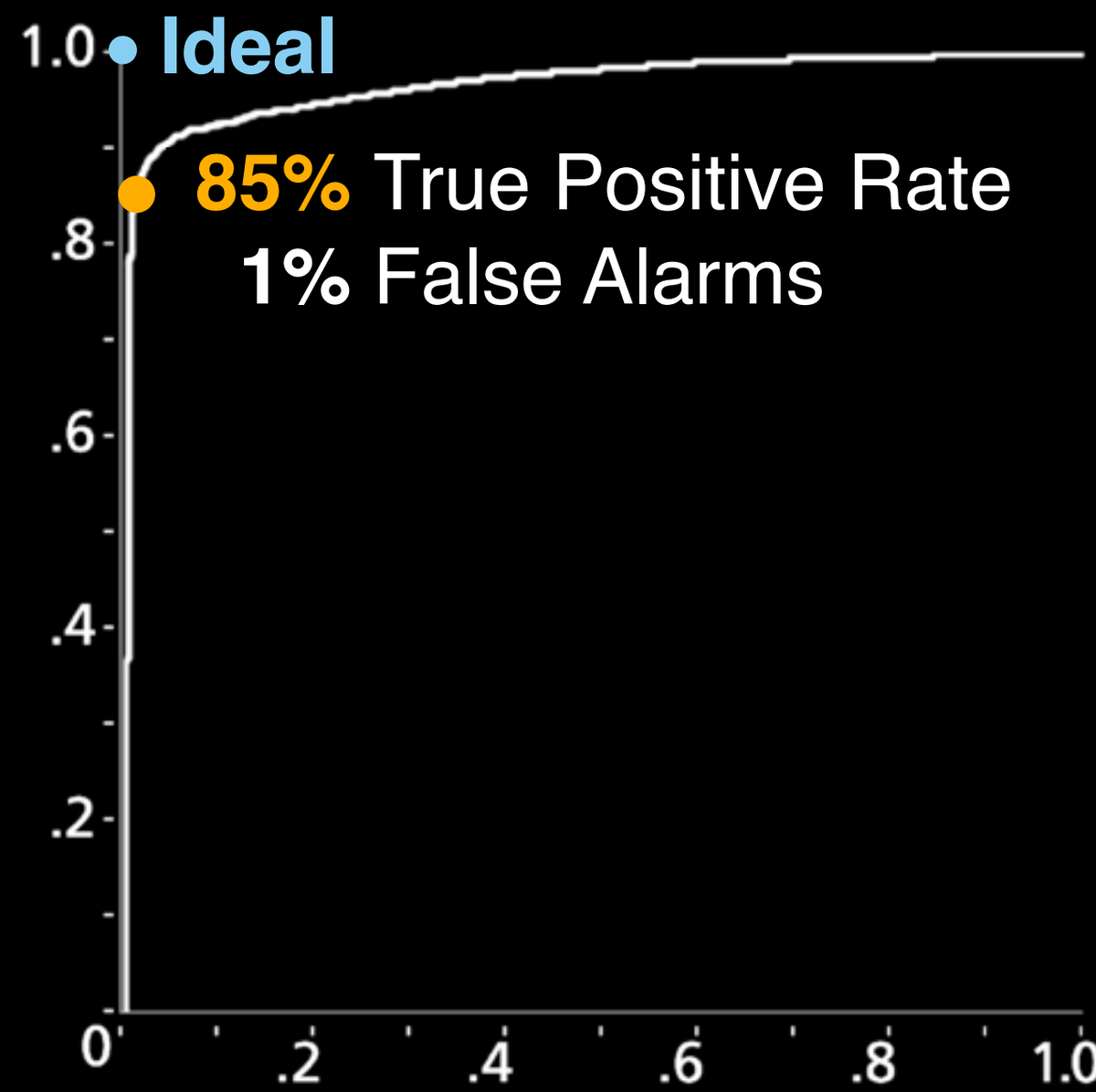
$$m_{ij}(x_j) \leftarrow \sum_{x_i \in X} \psi_{ij}(x_i, x_j) \, \phi(x_i) \prod_{k \in N(i) \setminus j} m_{ki}(x_i)$$

Opinion for neighbor Edge potential Belief

	Good	Bad
Good	0.9	0.1
Bad	0.1	0.9

Evaluation

Using **millions** of ground truth files,10-fold cross validation



True Positive Rate

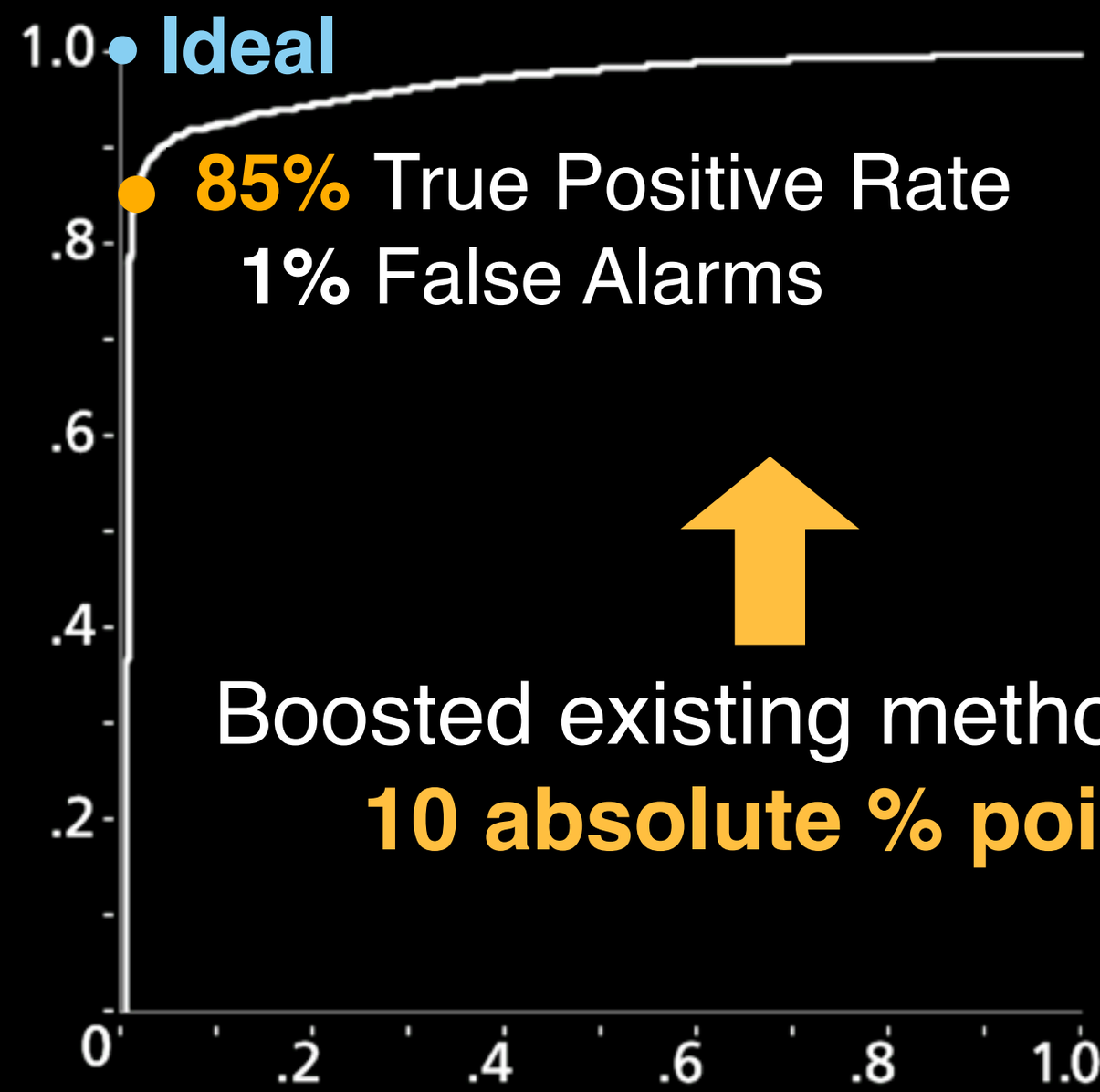
% of **bad** correctly labeled

False Positive Rate (False Alarms)

% of **good** labeled as **bad**

Evaluation

Using **millions** of ground truth files,10-fold cross validation



True Positive Rate
% of **bad** correctly labeled

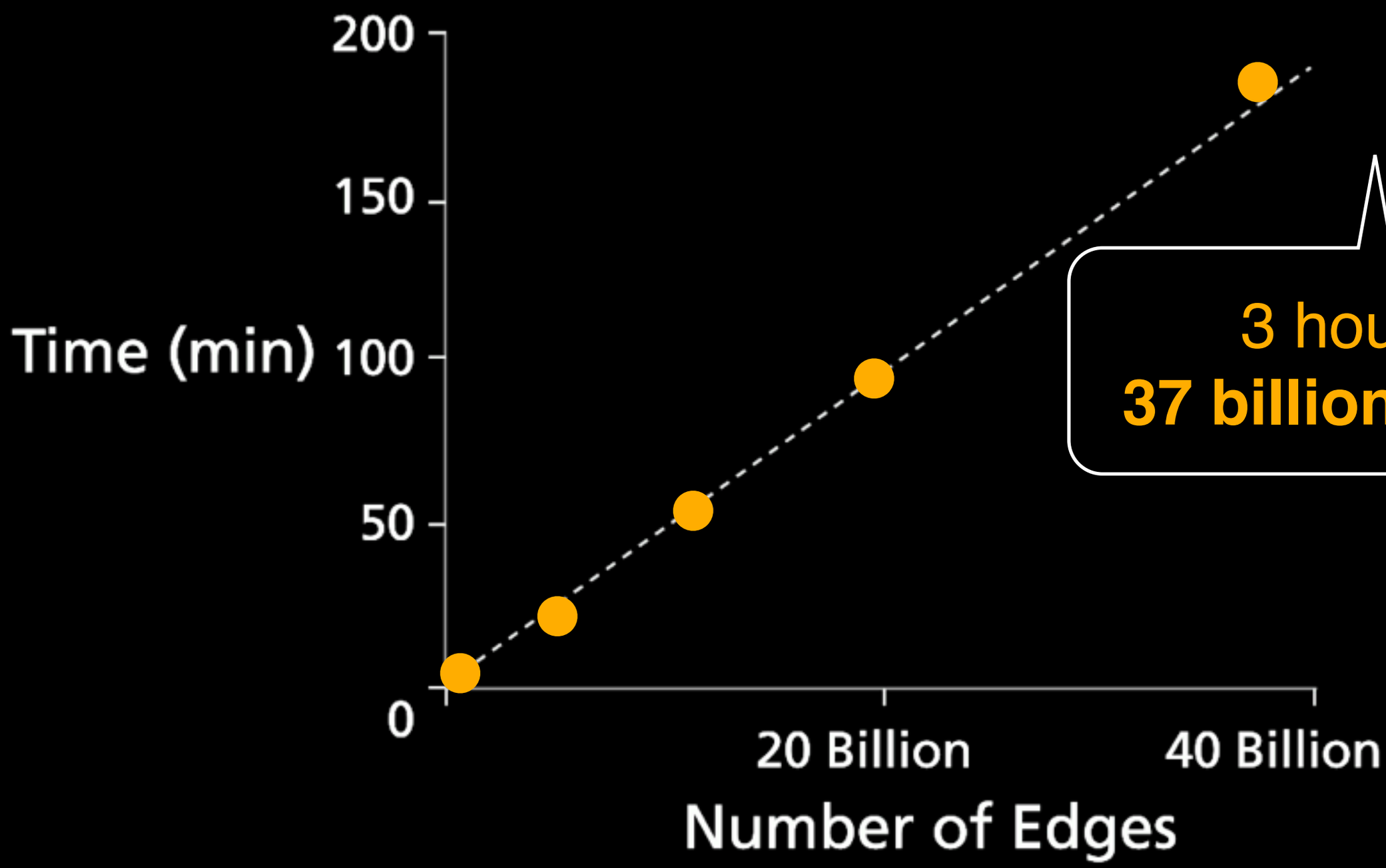
Boosted existing methods by
10 absolute % point

False Positive Rate (False Alarms)
% of **good** labeled as **bad**

Scalability

Running Time Per Iteration

Linux
16-core Opteron
256GB RAM



3 hours,
37 billion edges

Details

Scalability

How Did I Scale Up BP?

1. Early termination (after 6 iterations) → Faster

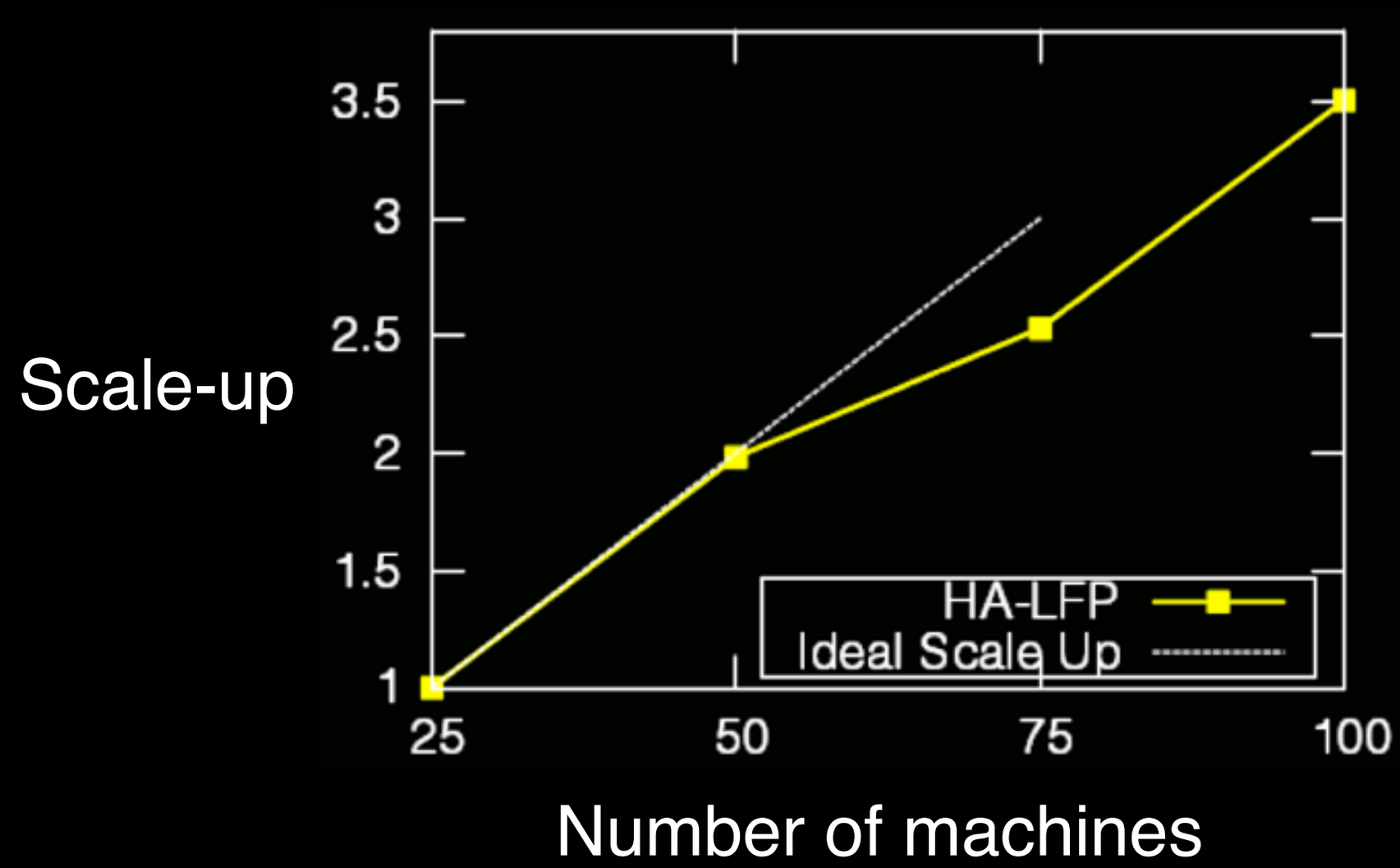
2. Keep edges on disk → Saves 200GB of RAM

3. Computes half of the messages → Twice as fast

Further Scale Up Belief Propagation

Use **Hadoop** if graph doesn't fit in memory [ICDE'11]

Speed scales up **linearly** with number of machines

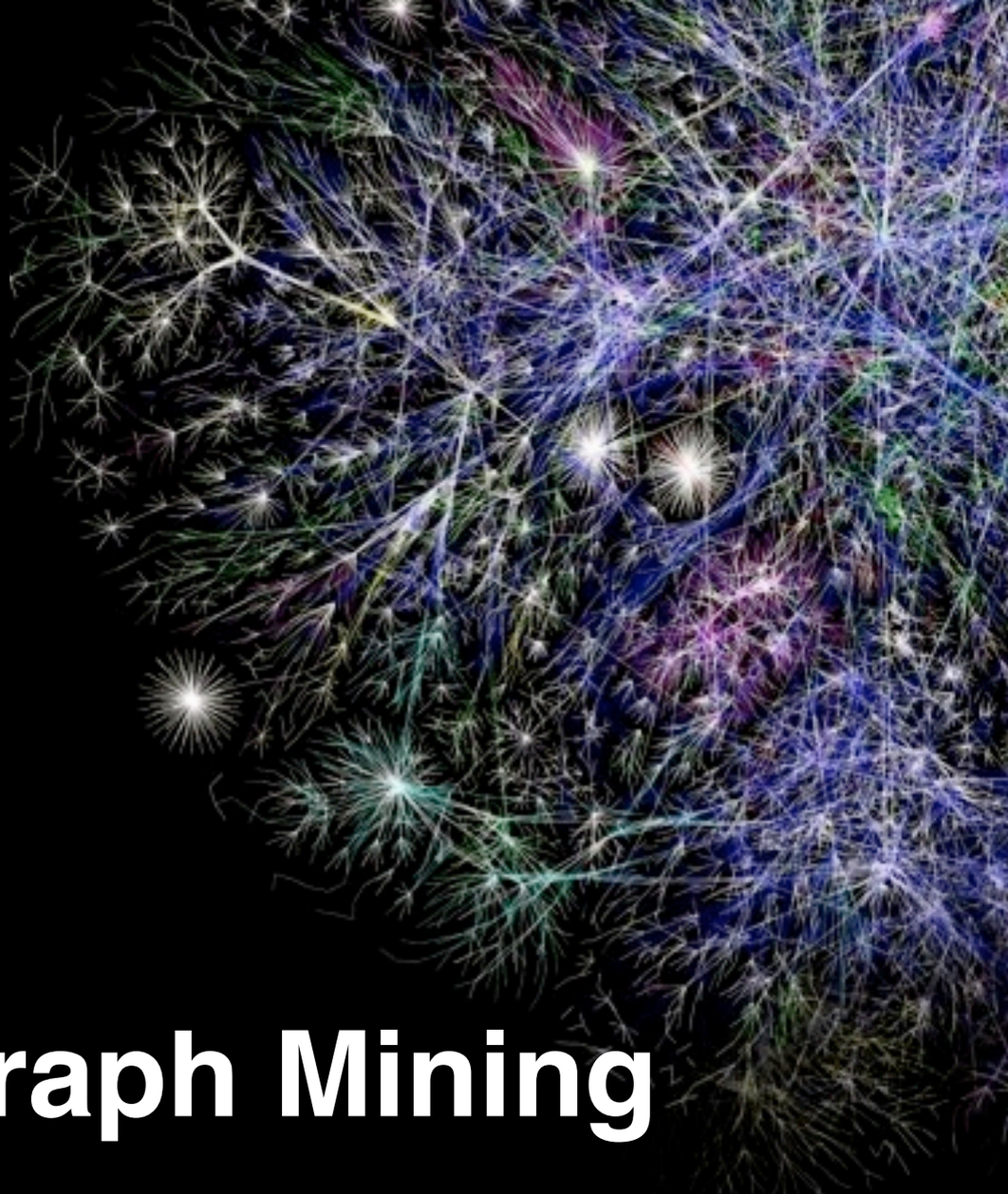


Yahoo! M45 cluster

480 machines

1.5 PB storage

3.5TB machine



Attention Routing

Polonium – Find Malware

NetProbe – Find Fraudsters

Human-In-The-Loop Graph Mining

Apolo – Mixed-Initiative Graph Sensemaking (ML + Vis)

NetProbe:
Fraud Detection in Online Auction

WWW 2007

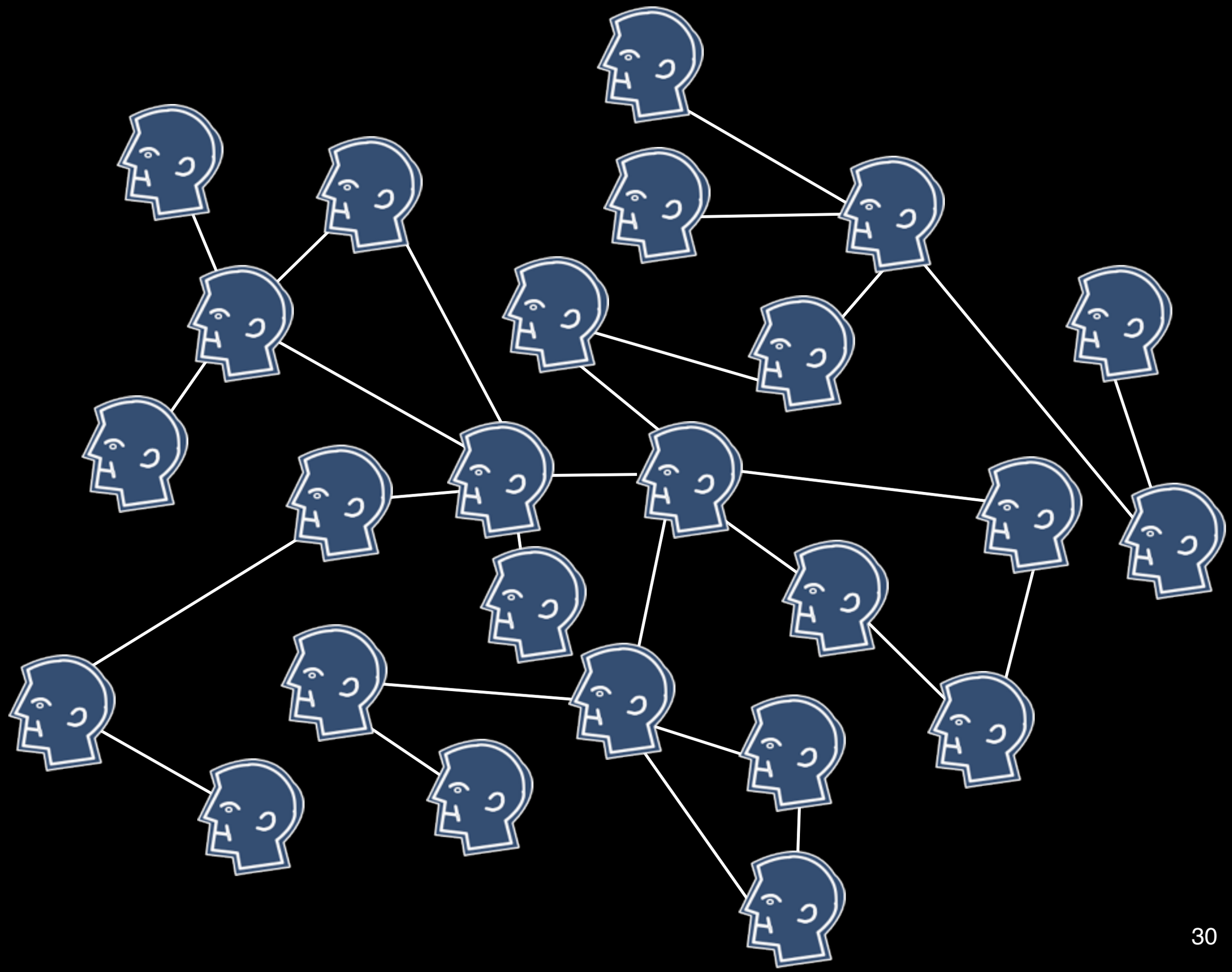
NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks

NetProbe: The Problem

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

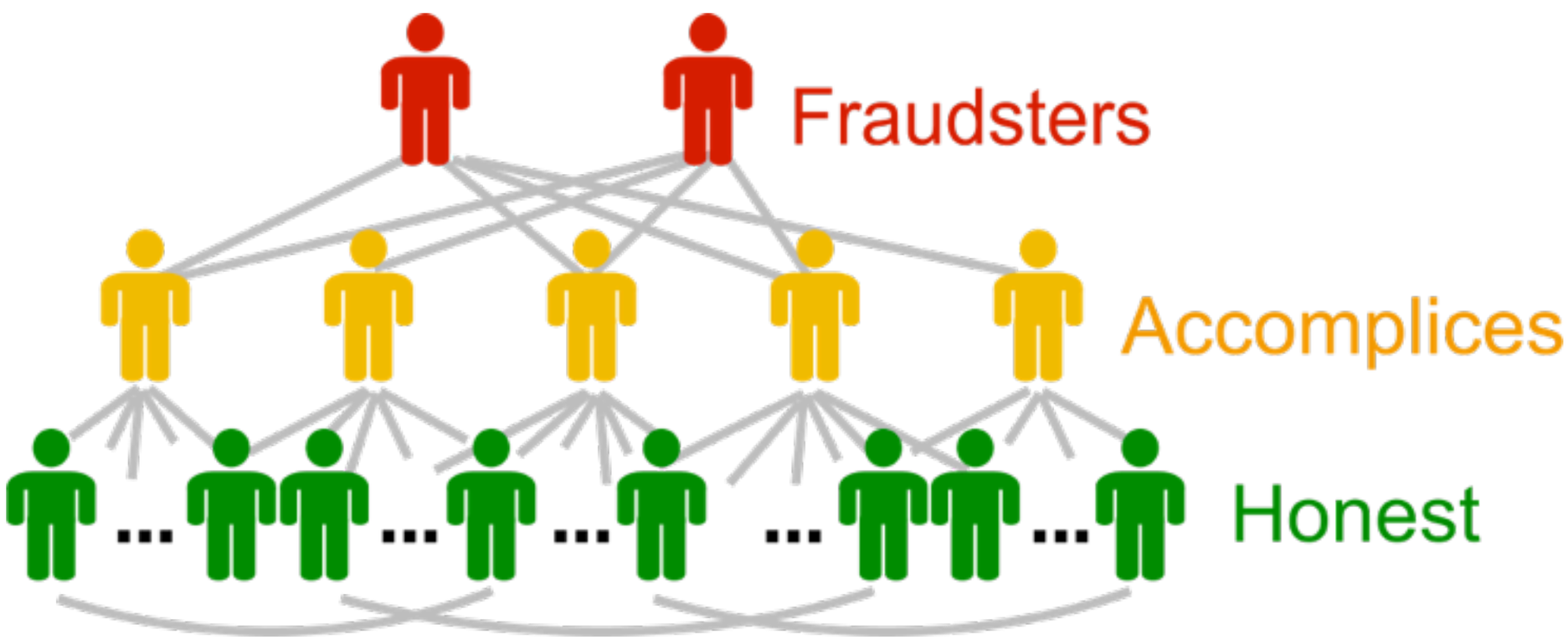
Auction fraud is **#3** online crime in 2010

source: www.ic3.gov



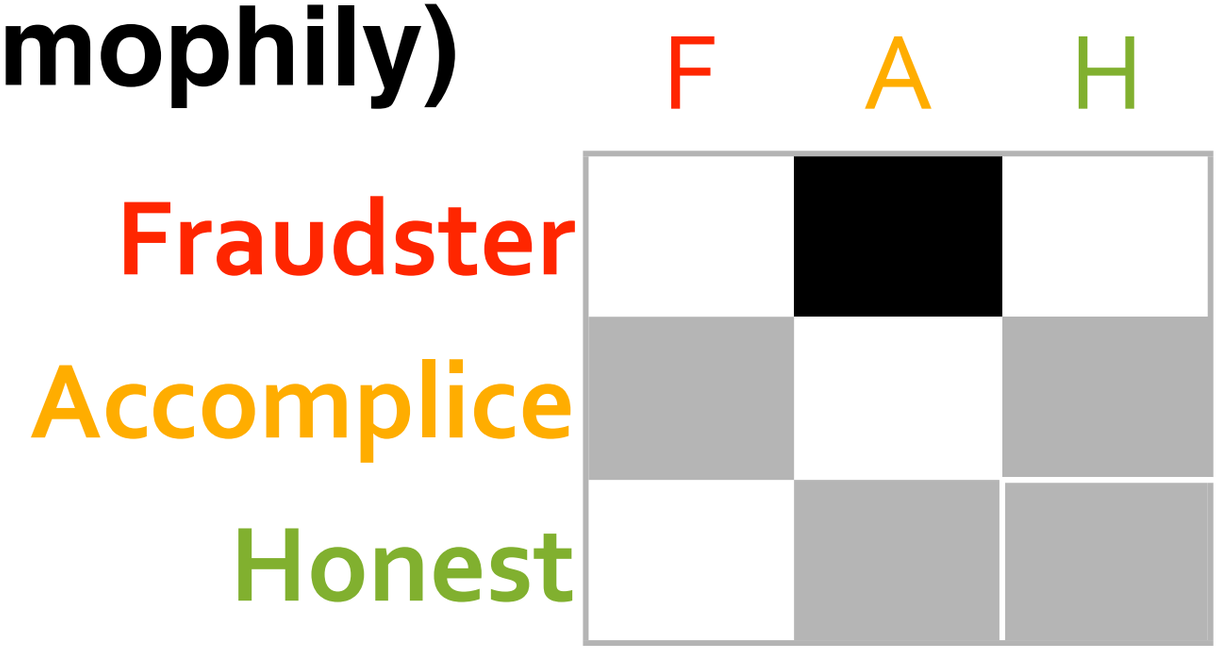
NetProbe: Key Ideas

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

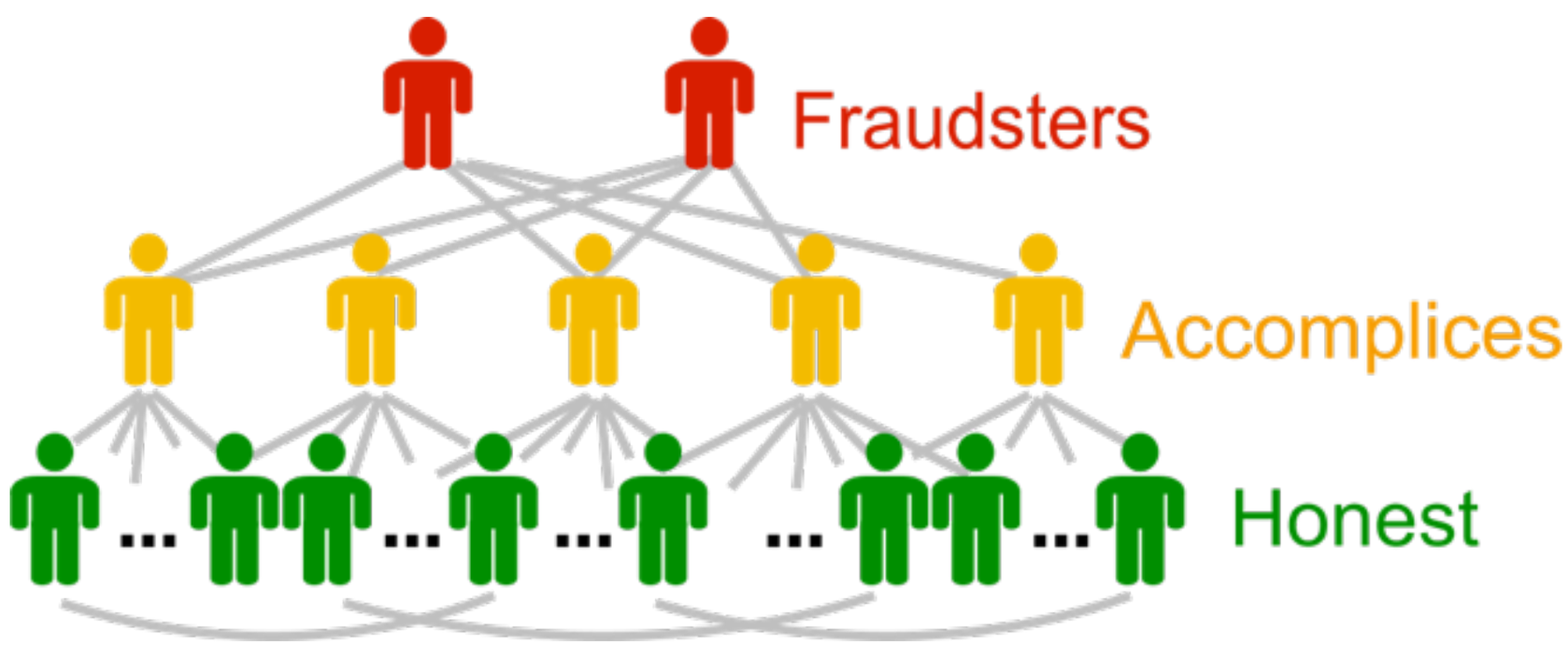


NetProbe: Key Ideas

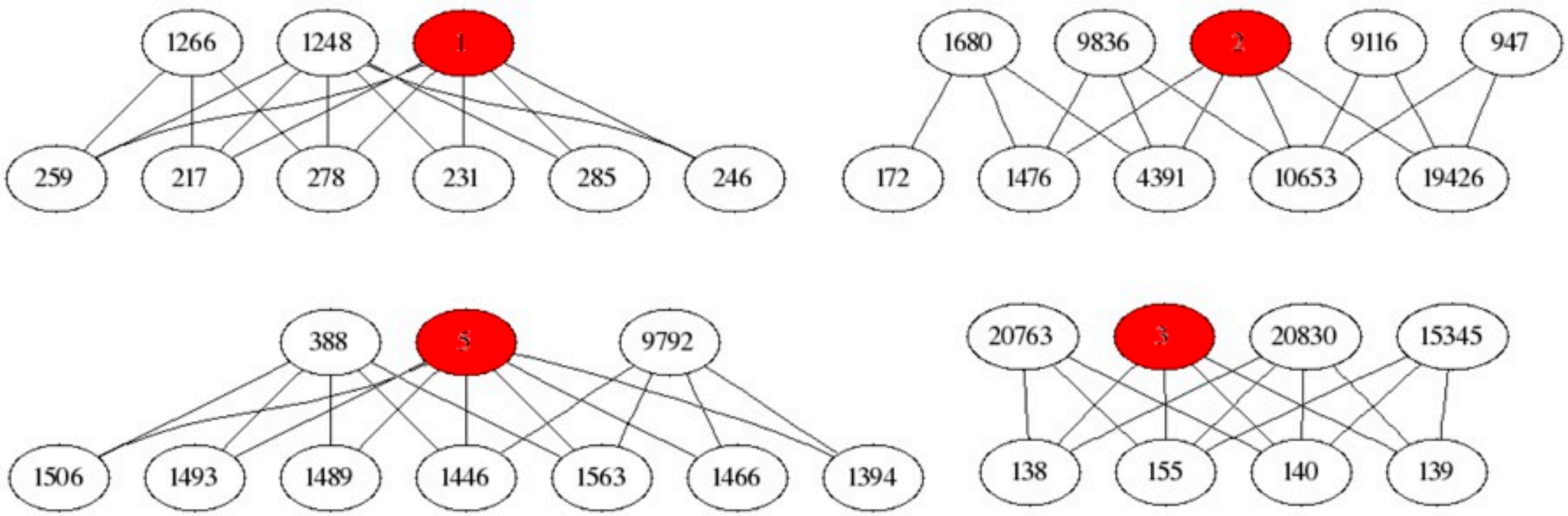
Use **Belief Propagation** and **heterophily (+ homophily)**



Darker means more likely



NetProbe: Main Results




ebay

THE WALL STREET JOURNAL.

PITTSBURGH TRIBUNE-REVIEW

Symantec

ebay

USA TODAY

CNNMoney.com
A Service of CNN, Fortune & Money

THE WALL STREET JOURNAL.

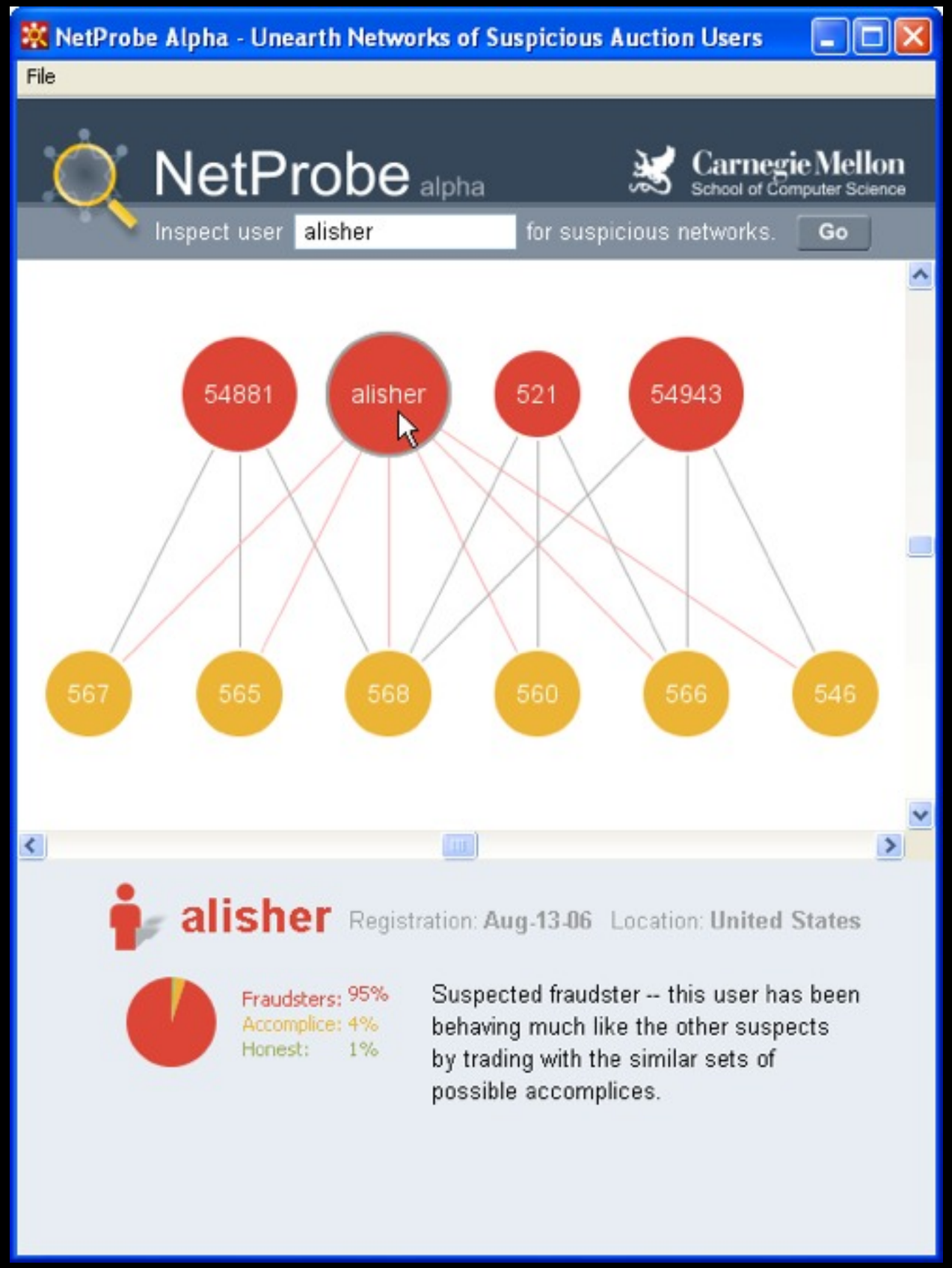
MSNBC

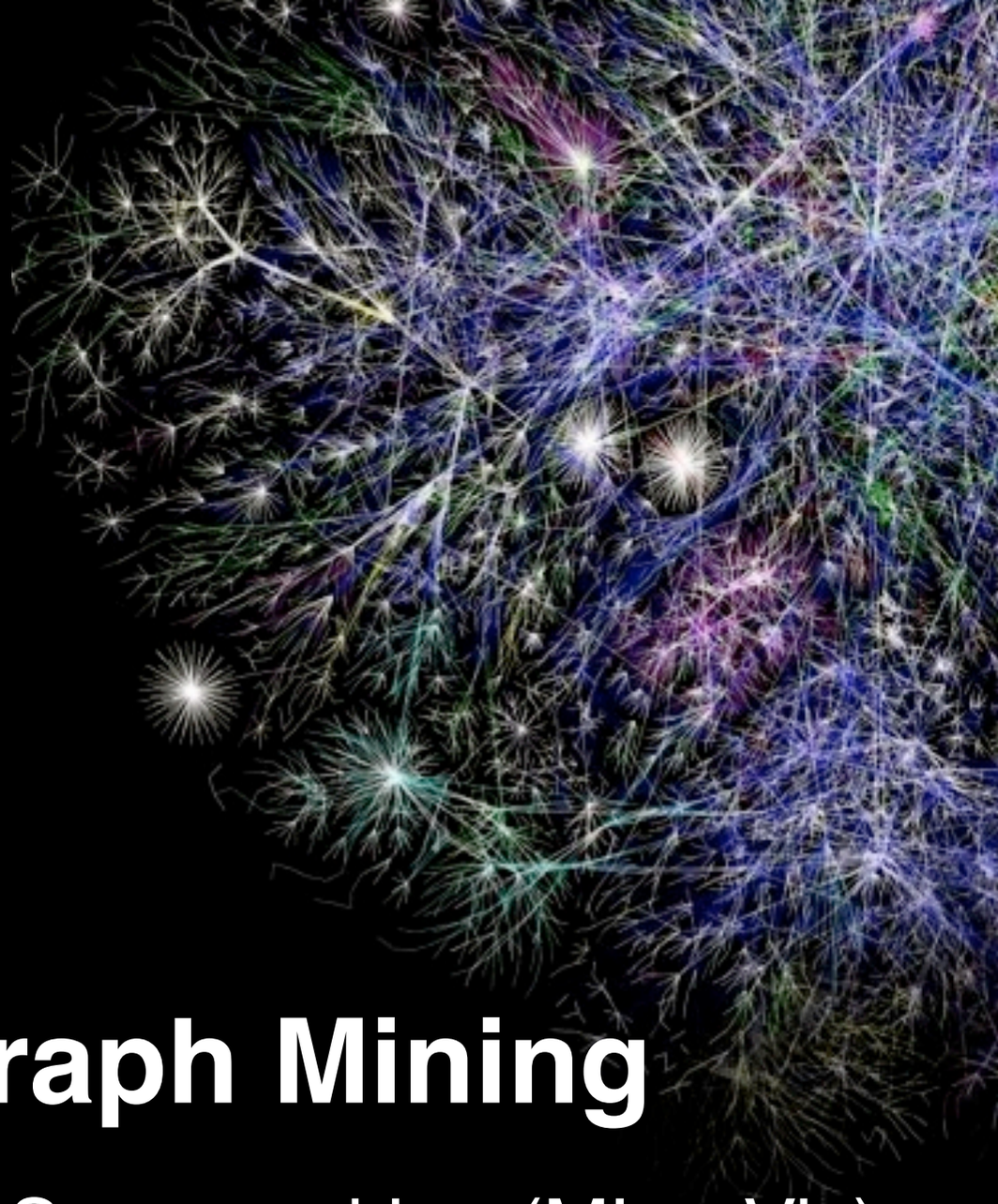
KDKA 2

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.



Attention Routing

Polonium – Find Malware

NetProbe – Find Fraudsters

Human-In-The-Loop Graph Mining

Apolo – Mixed-Initiative Graph Sensemaking (ML + Vis)

Human-In-The-Loop Graph Mining

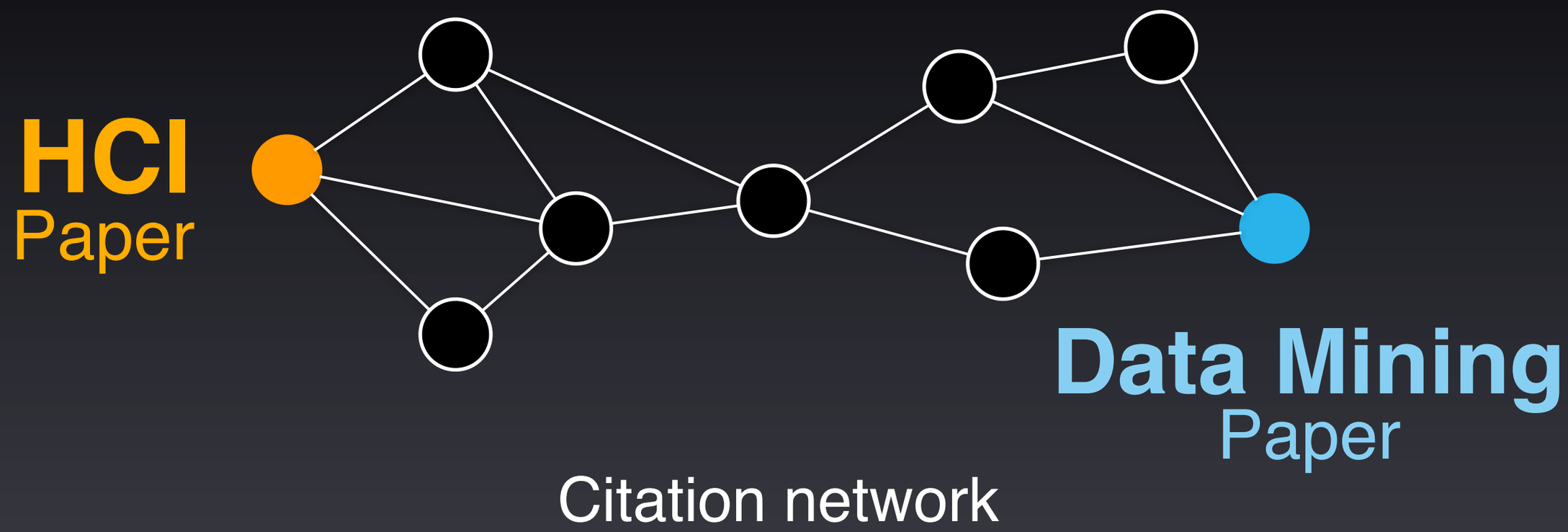
Apolo:

Machine Learning + Visualization

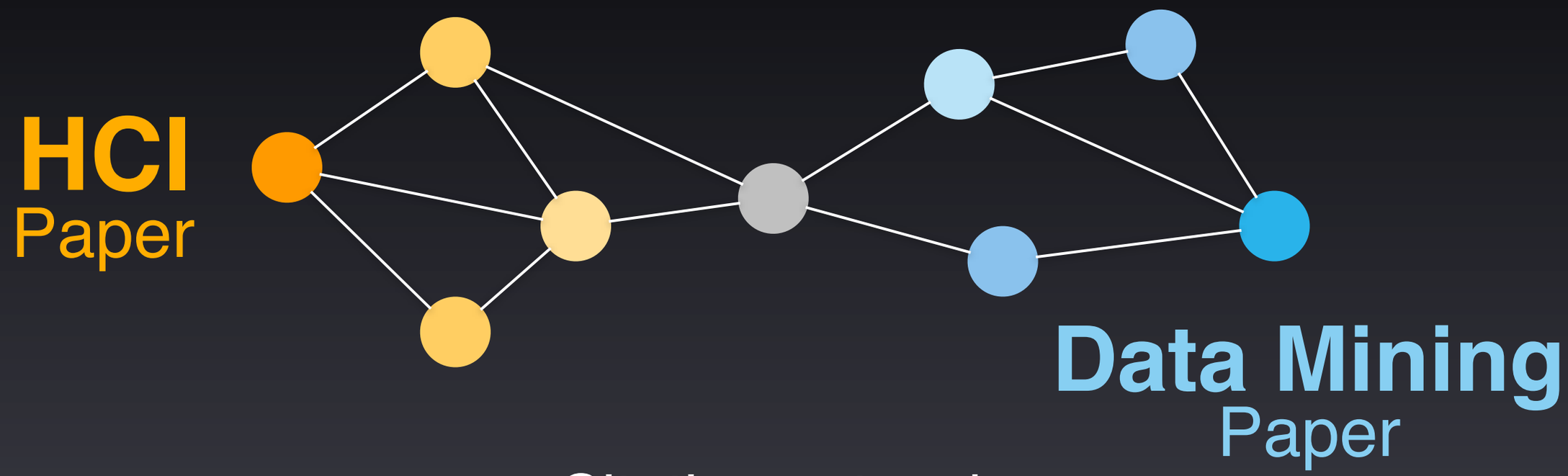
CHI 2011

Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning

Finding More Relevant Nodes

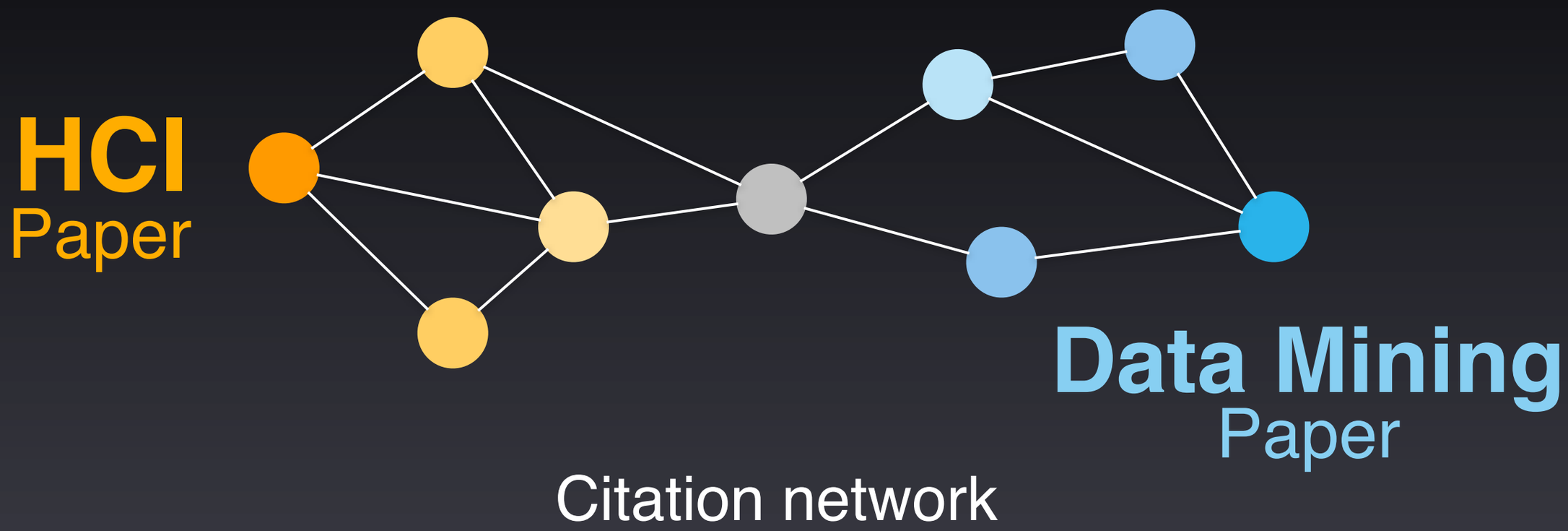


Finding More Relevant Nodes



Citation network

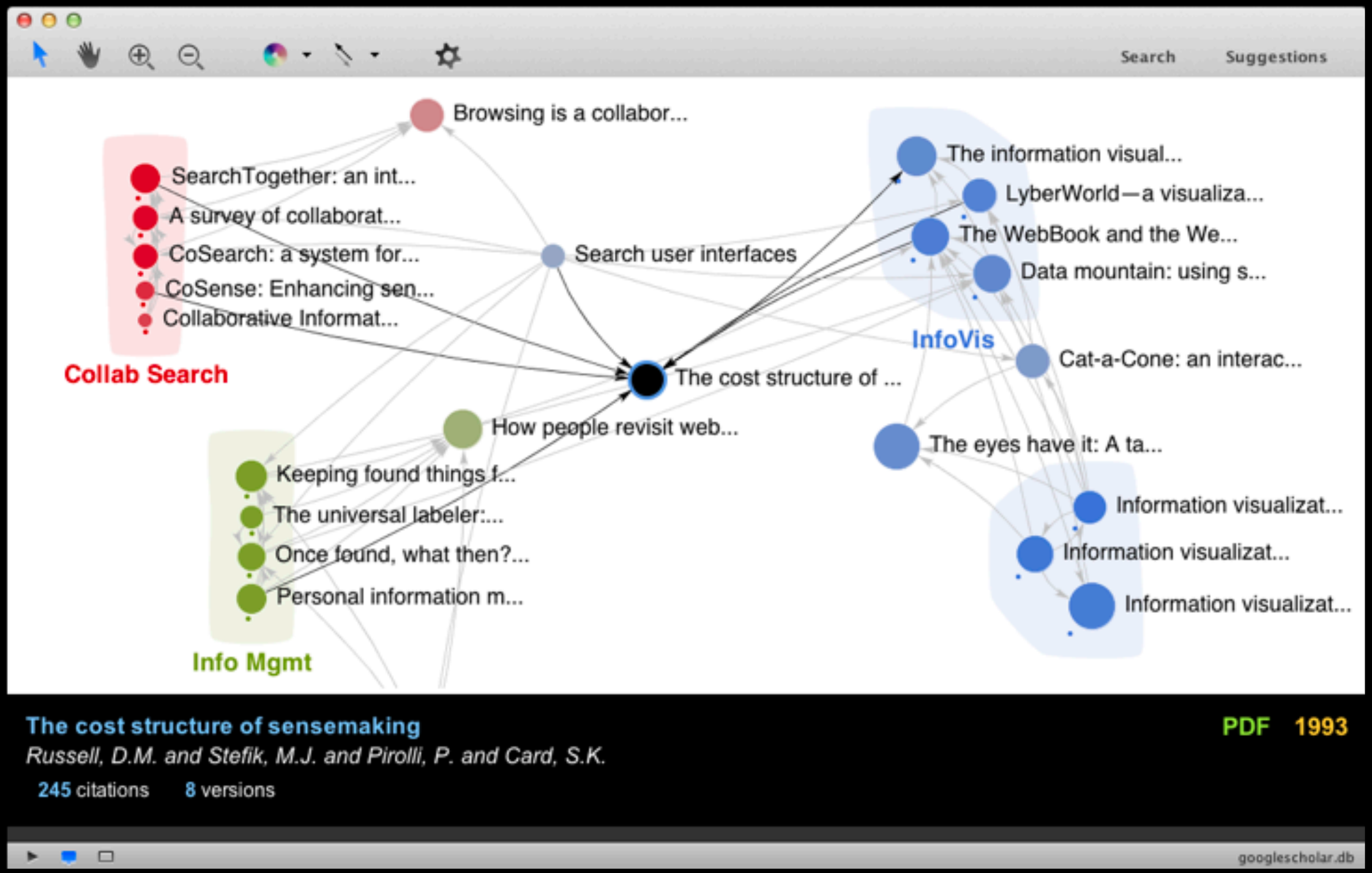
Finding More Relevant Nodes



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

Demo: Mapping the Sensemaking Literature

Nodes: 80k papers from Google Scholar (node size: #citation)
Edges: 150k citations



Search | Suggestions

For **The cost structure of sensemaking**

The information visualizer, an inf...	Card, S.K. and Robertson, G.G. and Macki...	1991 532
The WebBook and the Web Forag...	Card, S.K. and Robertson, G.G. and York, W.	1996 403
LyberWorld—a visualization user...	Hemmje, M. and Kunkel, C. and Willett, A.	1994 223
The structure of the information...	Card, S.K. and Mackinlay, J.	1997 198
Information visualization	Card, S. and Mackinlay, JD and Shneiderm...	2009 180
"I"ll get that off the audio": a cas...	Moran, T.P. and Palen, L. and Harrison, S....	1997 143
An organic user interface for sear...	Mackinlay, J.D. and Rao, R. and Card, S.K.	1995 123
Using a landscape metaphor to re...	Chalmers, M.	1993 122
Personal information management	Jones, W.P. and Teevan, J.	2007 109
SearchTogether: an interface for c...	Morris, M.R. and Horvitz, E.	2007 108
Information foraging theory: Ada...	Pirolli, P.	2007 107
Investigating behavioral variabilit...	White, R.W. and Drucker, S.M.	2007 79
Jigsaw: Supporting investigative...	Stasko, J. and Görg, C. and Liu, Z.	2008 71
The cost-of-knowledge character...	Card, S.K. and Pirolli, P. and Mackinlay, J.D.	1994 54
Collaborative conceptual design:...	Potts, C. and Catledge, L.	1996 45

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.

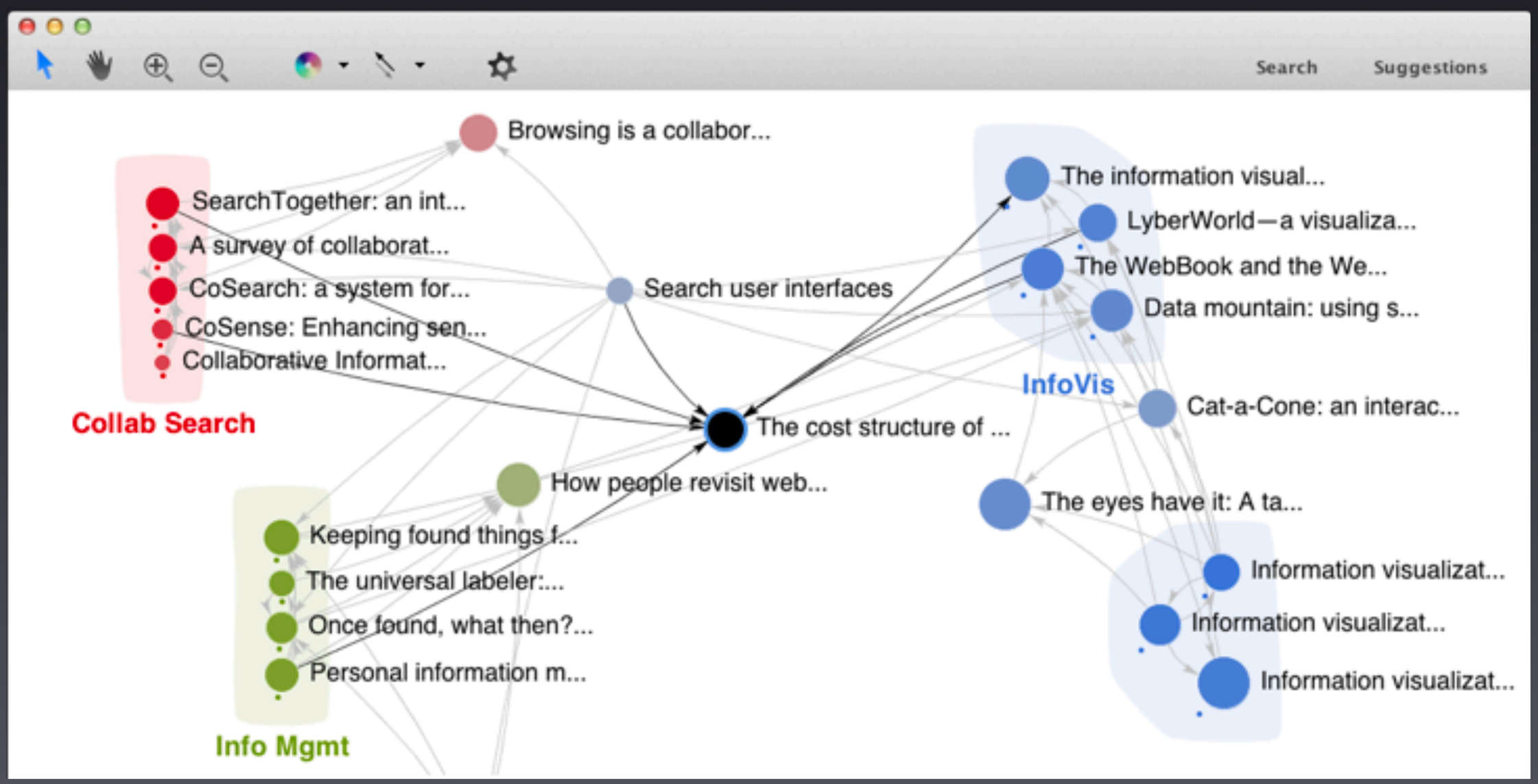
245 citations 8 versions

googlescholar.db

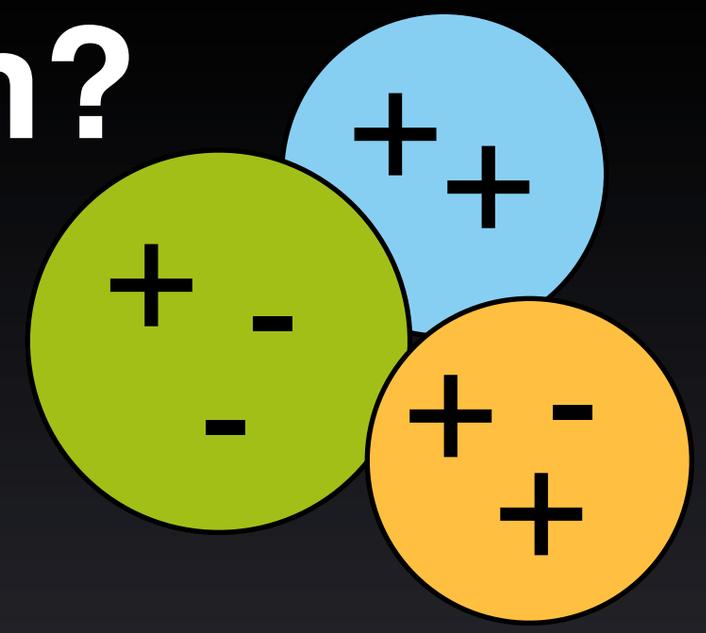
Key Ideas (Recap)

Specify exemplars

Find other relevant nodes (BP)



Why Belief Propagation?



- Multiple groups
- Multiple examples in each group
- Positive and negative examples
- Fast: running time linear in #edges
- We're first to adapt BP for sensemaking
- Other methods considered: Personalized PageRank (= Random Walk with Restart = RWR), etc.
  - Our PKDD'11 paper shows BP>>RWR

Apolo’s Contributions

1 Human + Machine

It was like having a **partnership** with the machine.

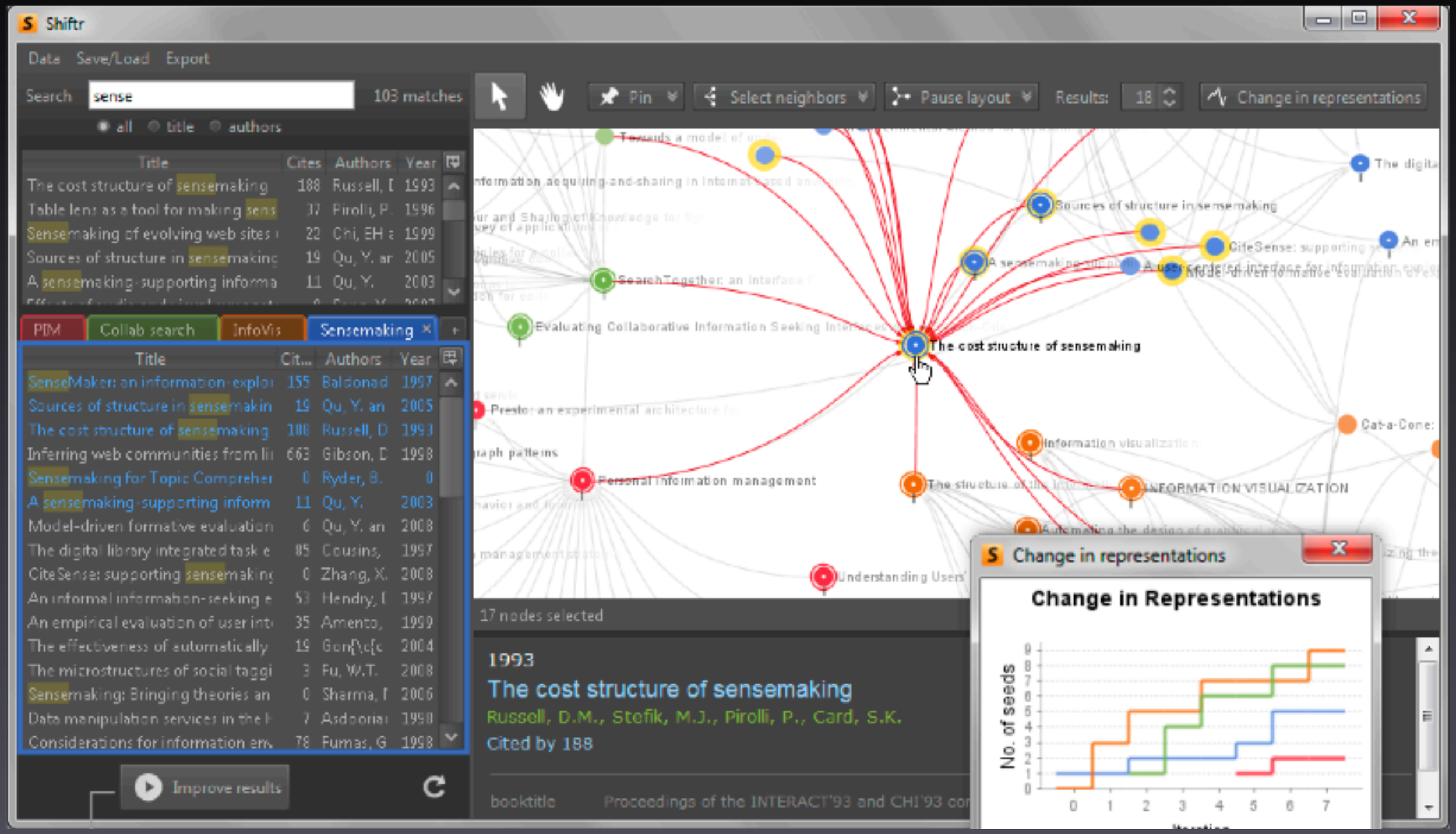
Apolo User

2 Personalized Landscape

Apolo 2009

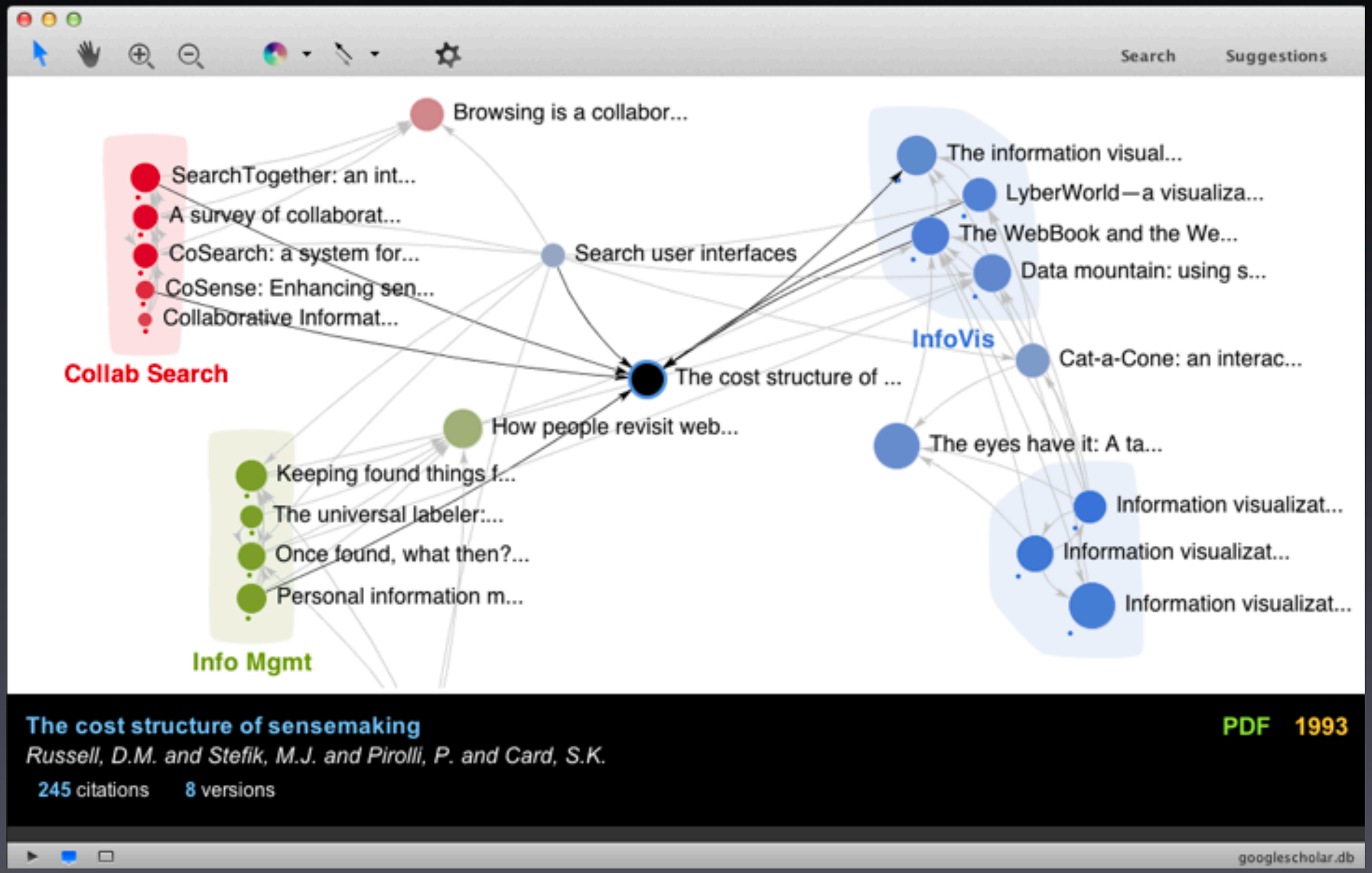

Cluster Data
Add Group
Recommendations:
Not Interested
Automatically generating user inte...
Decision-Theoretic User Interface ...
Daniel S. Weld
Krzysztof Z. Gajos
Automatically generating
Exploring the design spac
Predictability and accurac
Planning-Based Control o
End User Programming
End users creating effective softw...
End user software engineering: chi...
Invited research overview: end-us...
Brad A. Myers
Margaret M. Burnett
Mary Beth Rosson
Andrew Jensen Ko
Alan F. Blackwell
Show: All
Brad
Brad A. Myers
The garnet user interface developm...
Using HCI Techniques to Design a M...
Creating charts by demonstration.
The Amulet User Interface Developm...
Easily Adding Animations to Interfac...
Simplifying video editng using metad...
SILVER: simplifying video editing wit...
Text Entry
In-stroke word completion.
Integrating isometric joysticks into...
Eyes on the road, hands on the whe...
An alternative to push, press, and t...
Maximizing the guessability of symb...
Few-key text entry revisited: mnem...
Text entry from power wheelchairs: ...
Joystick text entry with date stamp, ...
Show: Papers
Interface Generation
Huddle: automatically generating i...
UNIFORM: automatically generatin...
Demonstrating the viability of auto...
Jeffrey Nichols
Brandon Rothrock
Duen Horng Chau

Apolo 2010


Shiftr
Data Save/Load Export
Search sense 103 matches
all title authors
Pin
Select neighbors
Pause layout
Results: 18
Change in representations
Title Cites Authors Year
The cost structure of sensemaking 188 Russell, D 1993
Table lens as a tool for making sens 37 Pirolli, P. 1996
Sensemaking of evolving web sites 22 Chi, EH 1999
Sources of structure in sensemaking 19 Qu, Y. 2005
A sensemaking-supporting informa 11 Qu, Y. 2003
PIM
Collab search
InfoVis
Sensemaking
Title Cit... Authors Year
SenseMaker: an information-explor 155 Baldonad 1997
Sources of structure in sensemakin 19 Qu, Y. 2005
The cost structure of sensemaking 188 Russell, D 1993
Inferring web communities from li 663 Gibson, D 1998
Sensemaking for Topic Compreher 0 Ryder, B. 0
A sensemaking-supporting inform 11 Qu, Y. 2003
Model-driven formative evaluation 6 Qu, Y. 2008
The digital library integrated task e 85 Cousins, 1997
CiteSense: supporting sensemaking 0 Zhang, X. 2008
An informal information-seeking e 53 Hendry, 1997
An empirical evaluation of user int 35 Amento, 1999
The effectiveness of automatically 19 2004
The microstructures of social taggi 3 Fu, W.T. 2008
Sensemaking: Bringing theories an 0 Sharma, 2006
Data manipulation services in the 7 Asdoorian 1998
Considerations for information en 78 Furnas, G 1998
Improve results
The cost structure of sensemaking
Sources of structure in sensemaking
CiteSense: supporting
Evaluating Collaborative Information Seeking
Presto: an experimental architecture
Personal information management
Information visualization
INFORMATION VISUALIZATION
Understanding Users'
Cat-a-Cone:
17 nodes selected
1993
The cost structure of sensemaking
Russell, D.M., Stefik, M.J., Pirolli, P., Card, S.K.
Cited by 188
booktitle Proceedings of the INTERACT'93 and CHI'93 con
Change in representations
Change in Representations
No. of seeds
0 1 2 3 4 5 6 7

Apolo 2011

22,000 lines of code. Java 1.6. Swing.
Uses SQLite3 to store graph on disk

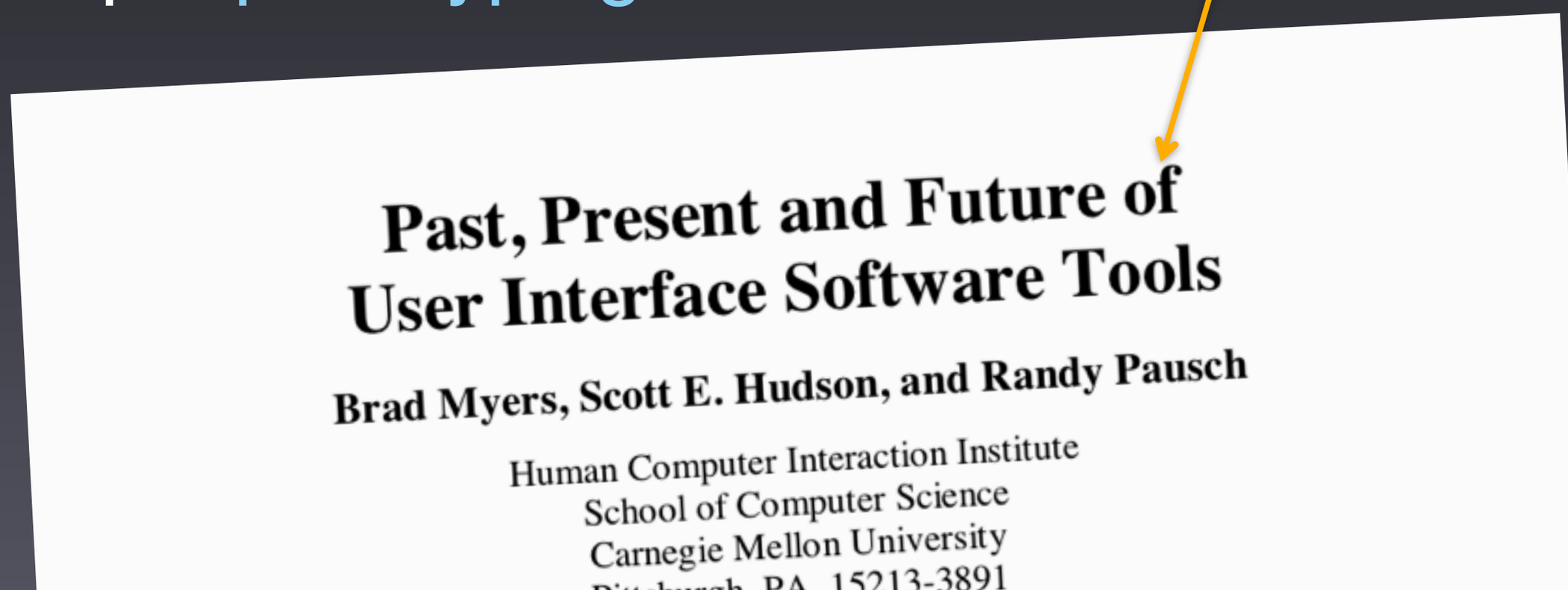


User Study

Used citation network

<u>Task</u>: Find related papers for 2 sections in a survey paper on *user interface*

1. Model-based generation of UI
2. Rapid prototyping tools



Past, Present and Future of User Interface Software Tools

Brad Myers, Scott E. Hudson, and Randy Pausch

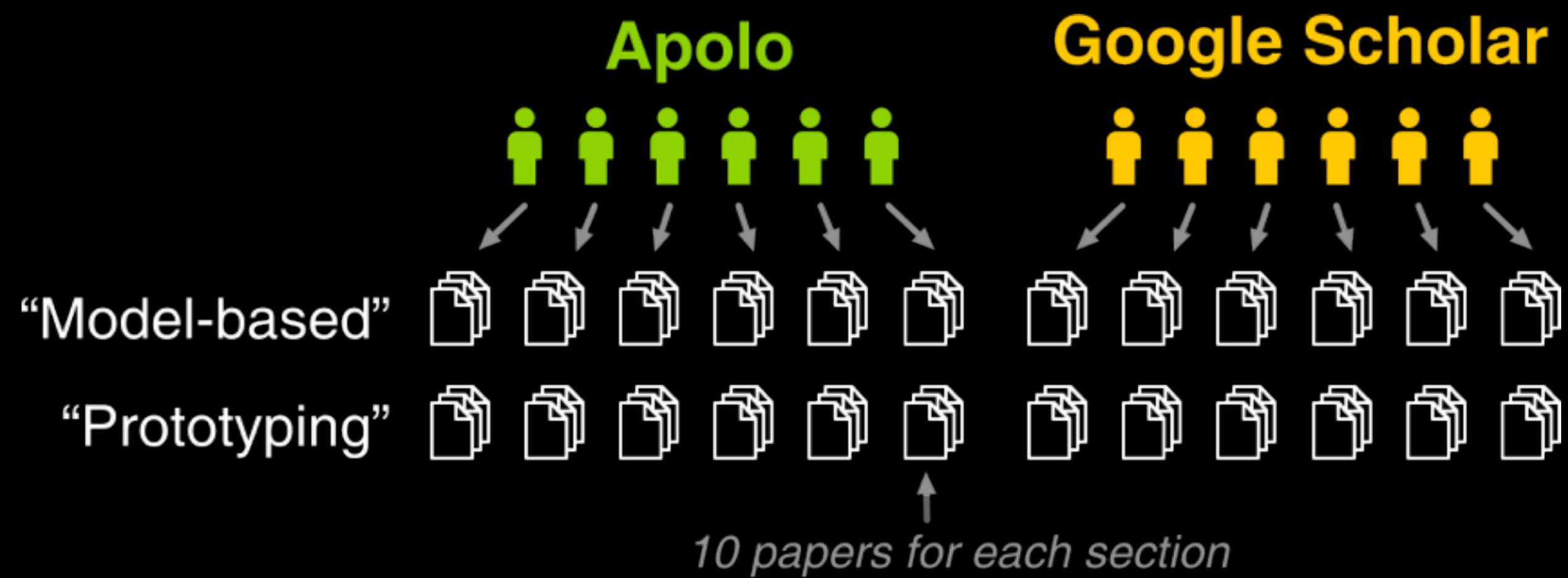
Human Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3891

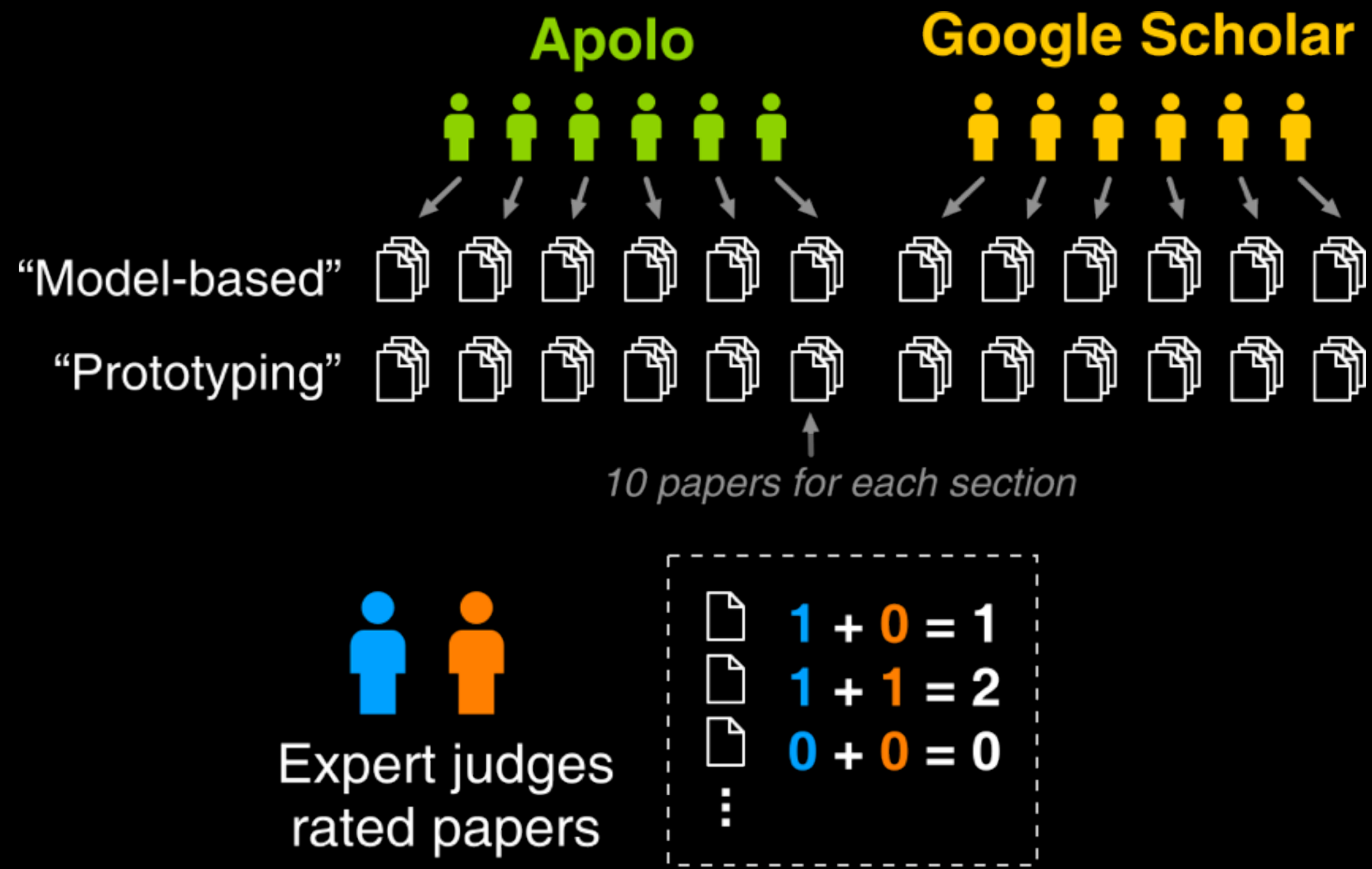
Apolo

Google Scholar

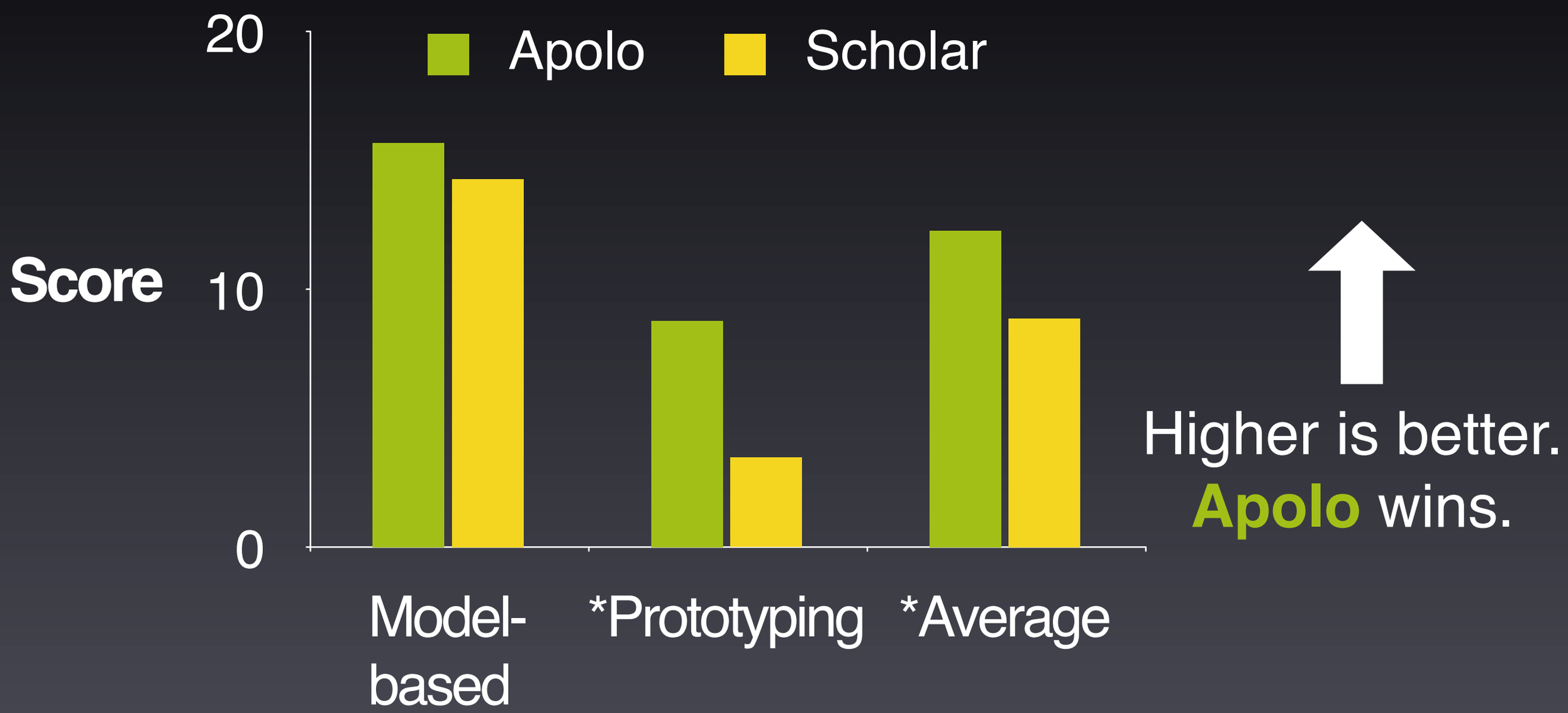
Between subjects design

Participants: **grad student** or **research staff**


Apolo
Google Scholar
"Model-based"
"Prototyping"
10 papers for each section


Apolo
Google Scholar
"Model-based"
"Prototyping"
10 papers for each section
Expert judges rated papers
1 + 0 = 1
1 + 1 = 2
0 + 0 = 0

Judges' Scores

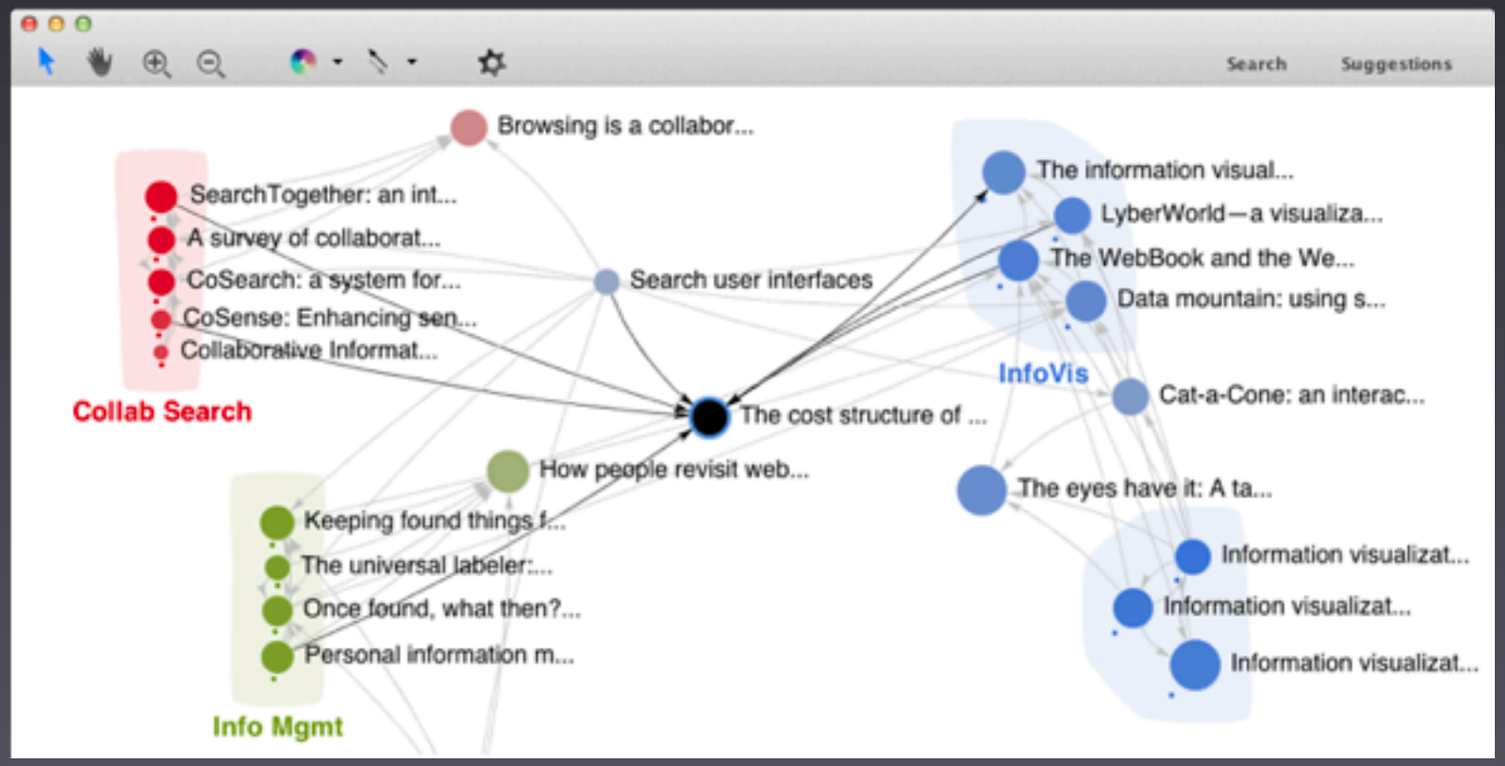


** Statistically significant, by two-tailed t test, $p < 0.05$*

Apolo: Recap

A mixed-initiative approach for exploring and creating personalized landscape for large network data

Apolo = ML + Visualization + Interaction



Feldspar

Finding Information by Association. CHI 2008

Polo Chau, Brad Myers, Andrew Faulring

YouTube: http://www.youtube.com/watch?v=Q0TIV8F_o_E&feature=youtu.be&list=ULQ0TIV8F_o_E
Paper: http://www.cs.cmu.edu/~dchau/feldspar/feldspar-chi08.pdf

Feldspar

Feldspar

A system that helps people find things on their computers when typical search or browsing tools don't work

Feldspar

A system that helps people find things on their computers when typical search or browsing tools don't work

An example scenario…

“Find the **webpage** mentioned in the **email** from the **person** I met at an **event**“

"Find the **webpage** mentioned in the **email** from the **person** I met at an **event**"

If I can't remember the specifics, such as any text in the webpage, email, etc.

→ Can't search

"Find the **webpage** mentioned in the **email** from the **person** I met at an **event**"

If I can't remember the specifics, such as any text in the webpage, email, etc.
→ Can't search

If I haven't bookmarked the webpage
→ Can't browse

“Find the **webpage** mentioned in the **email** from the **person** I met at an **event**“

"Find the **webpage** mentioned in the **email** from the **person** I met at an **event**"

But I can describe the webpage with a chain of associations.

"Find the **webpage** mentioned in the **email** from the **person** I met at an **event**"

But I can describe the webpage with a chain of associations.

webpage – email – person – event

"Find the **webpage** mentioned in the **email** from the **person** I met at an **event**"

But I can describe the webpage with a chain of associations.

webpage – email – person – event

The psychology literature has shown that people often remember things exactly like this.

Natural question:

Can I find things by associations?

Natural question:

Can I find things by associations?

Can I find the webpage by specifying its associated information (email, person, and event)?

Natural question:

Can I find things by associations?

Can I find the webpage by specifying its associated information (email, person, and event)?

We created Feldspar, which supports this associative retrieval of information.

Feldspar stands for….

Feldspar stands for….

Finding
Elements by
Leveraging
Diverse
Sources of
Pertinent
Associative
Recollection

Implementation: Overview

Create a graph database to store the associations among items on the computer

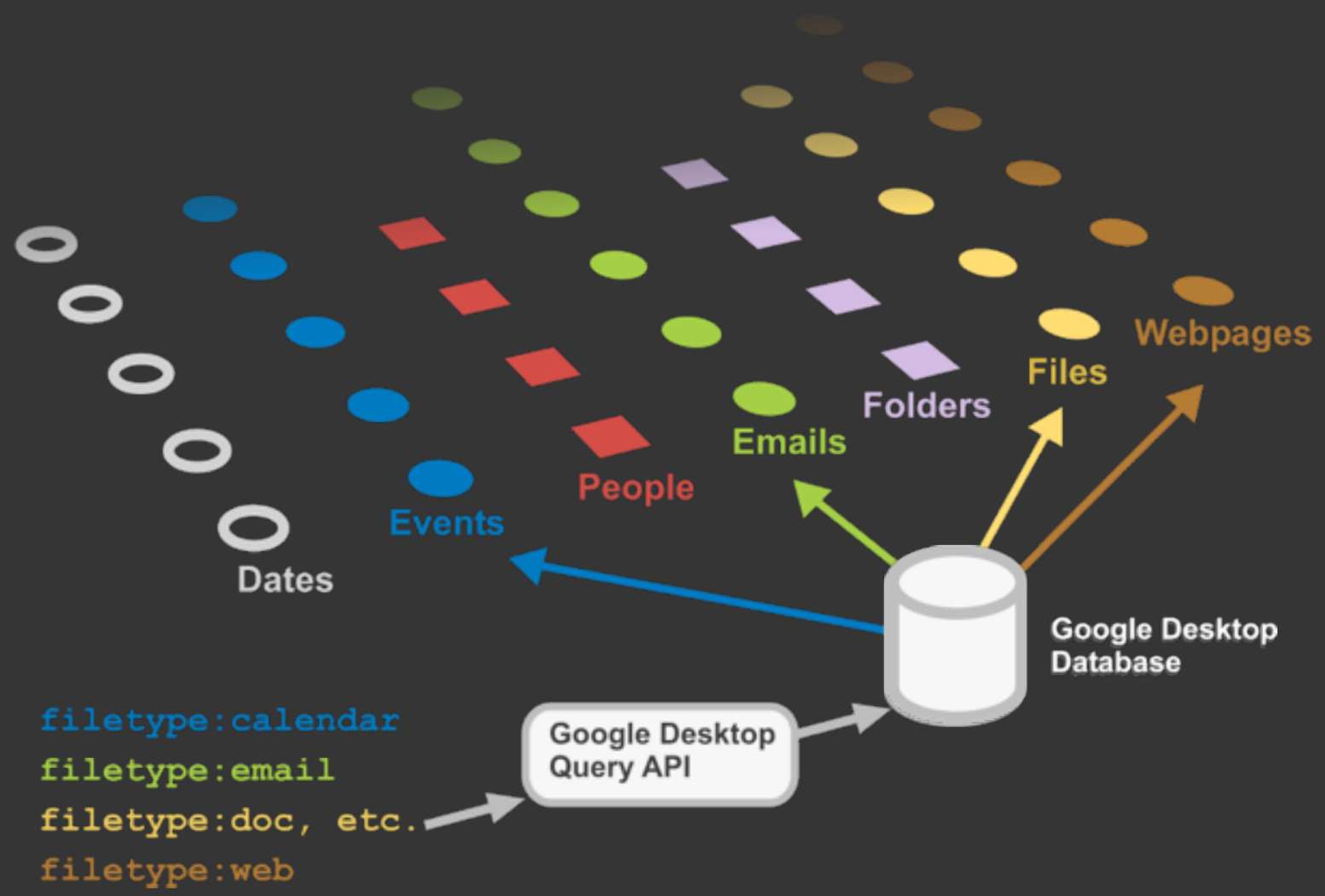
Develop an algorithm that processes the query and returns results

Creating an Association Database (a graph)

Google Desktop Database

Install **Google Desktop** and let it index all the items on the computer

Creating an Association Database (a graph)

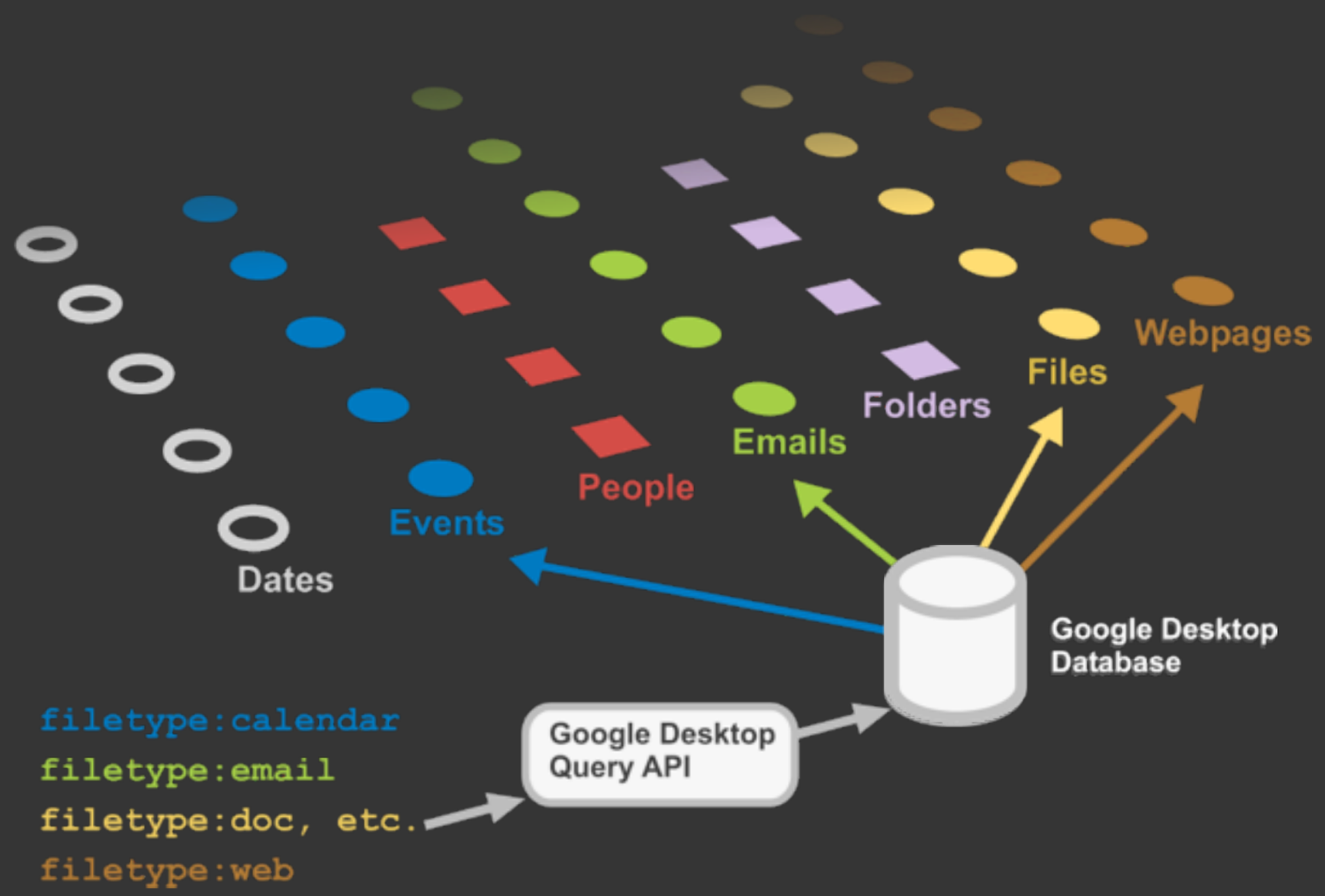


Focus on **7** types

Install **Google Desktop** and let it index all the items on the computer

Creating an Association Database (a graph)

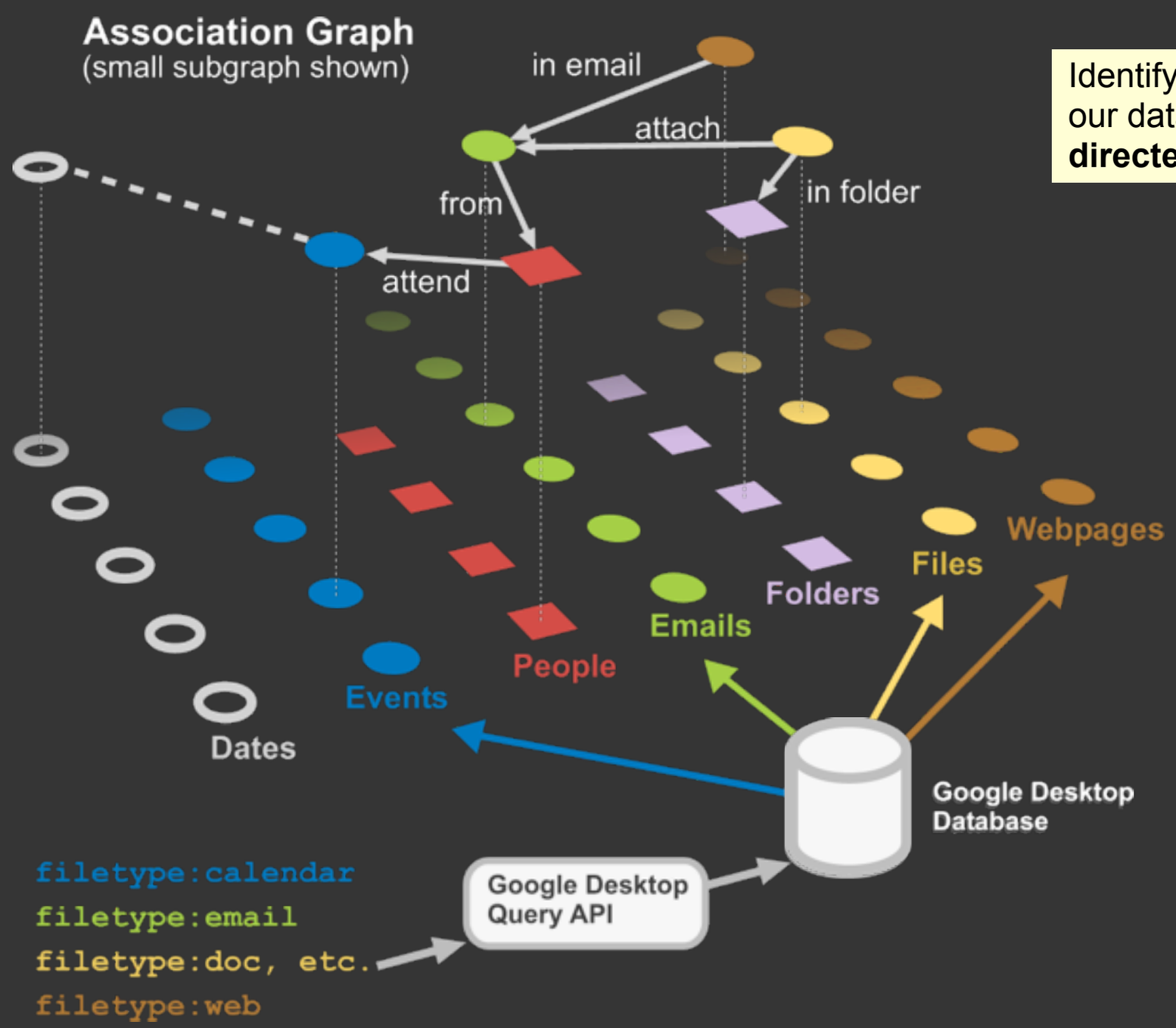
Identify **associations** and build our database, which is a **directed graph**



Focus on **7** types

Install **Google Desktop** and let it index all the items on the computer

Creating an Association Database (a graph)

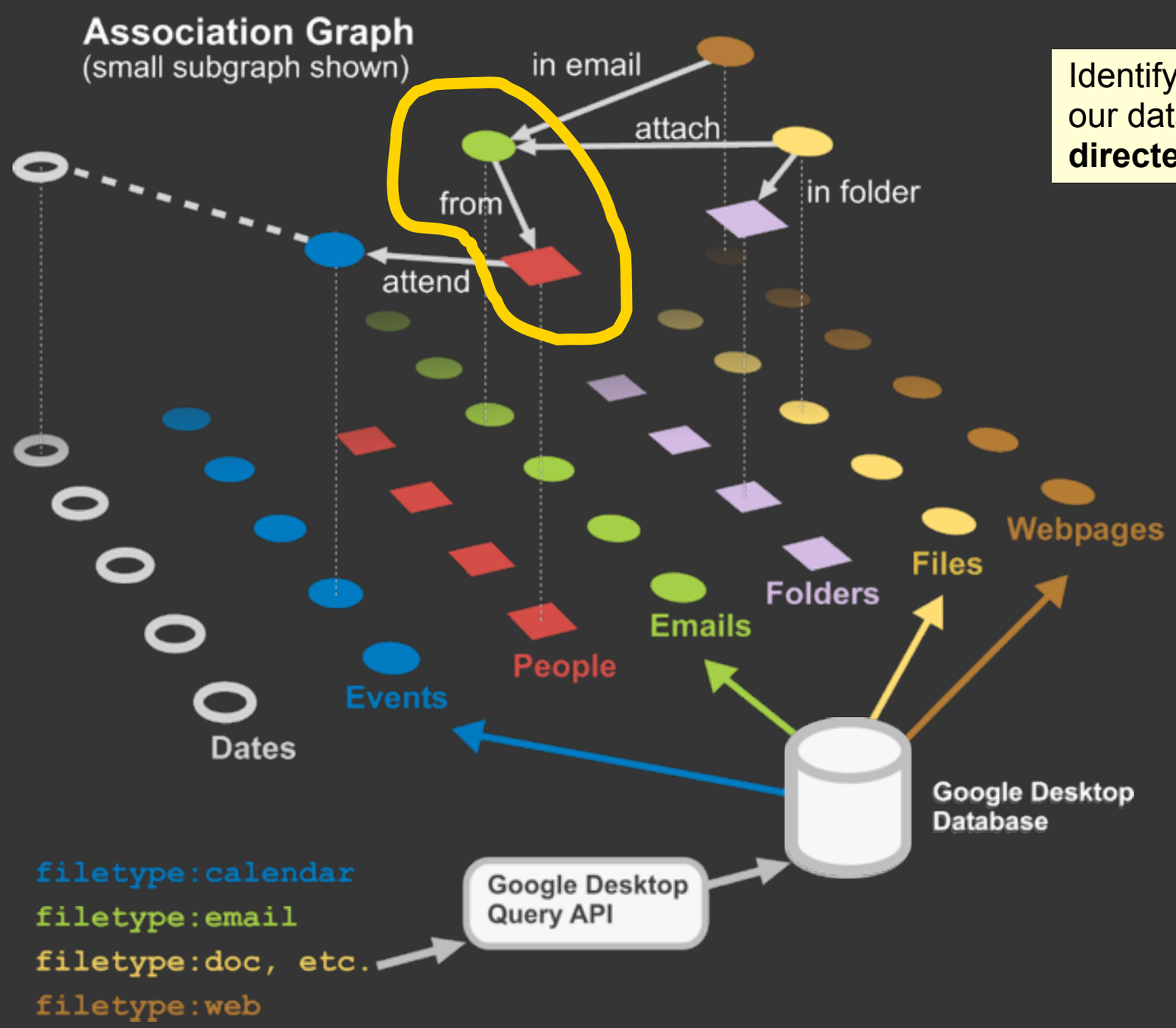


Identify **associations** and build our database, which is a **directed graph**

Focus on **7** types

Install **Google Desktop** and let it index all the items on the computer

Creating an Association Database (a graph)



Identify **associations** and build our database, which is a **directed graph**

Focus on **7** types

Install **Google Desktop** and let it index all the items on the computer

Algorithm that processes the query

Algorithm that processes the query

Feldspar

Feldspar « Back Forward » Saved Queries... Save

Where's the

Email	Email	Email	Email
Person	Person	Person	Person
Folder	Folder	Folder	Folder
File	File	File	File
Webpage	Webpage	Webpage	Webpage
Date	Date	Date	Date
Event	Event	Event	Event ?
Other	Other	Other	Other

related to

related to

related to

... related to...

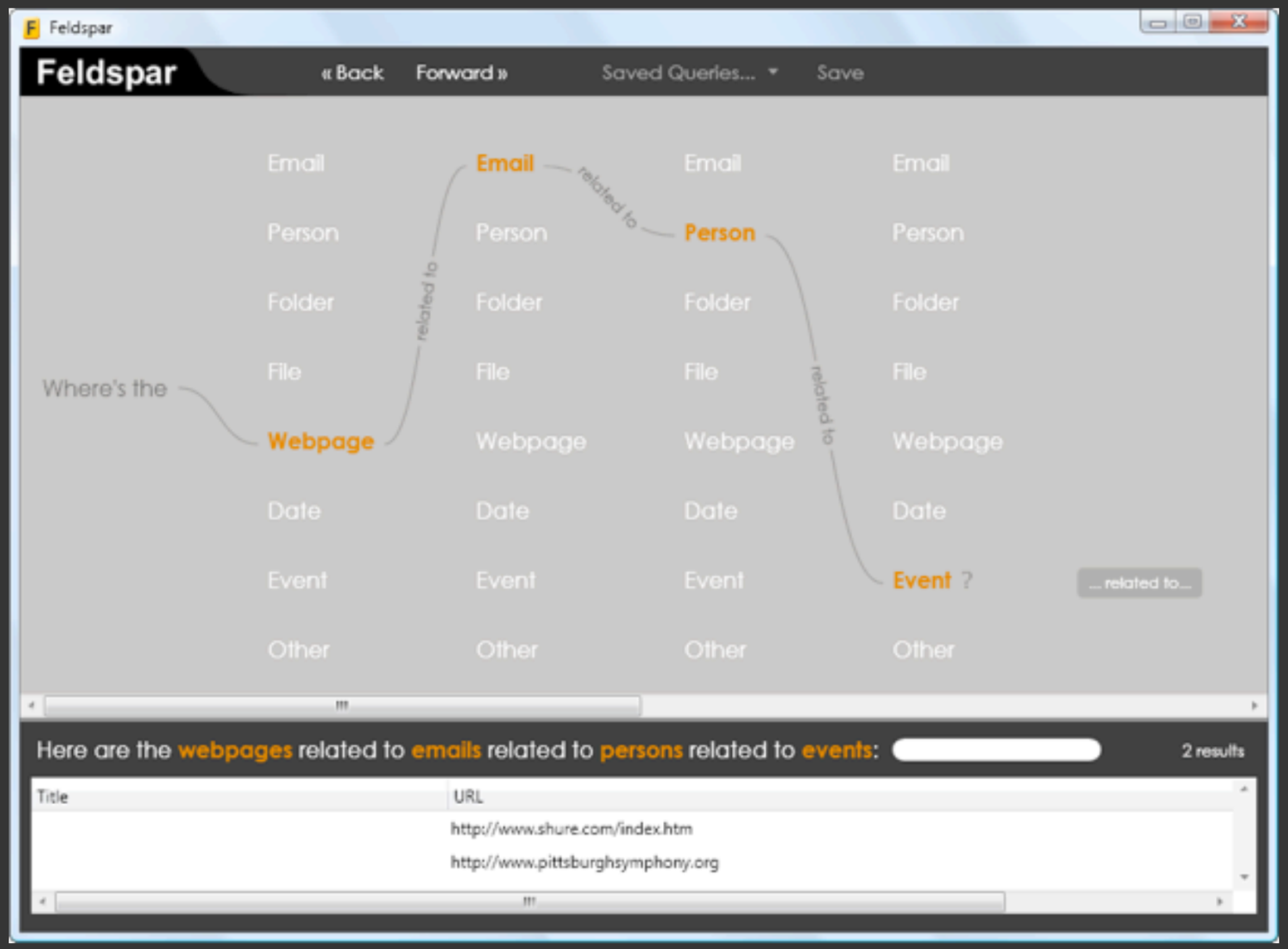
Here are the webpages related to emails related to persons related to events:

2 results

Title	URL
	http://www.shure.com/index.htm
	http://www.pittsburghsymphony.org

Algorithm that processes the query

webpages - emails - persons - events

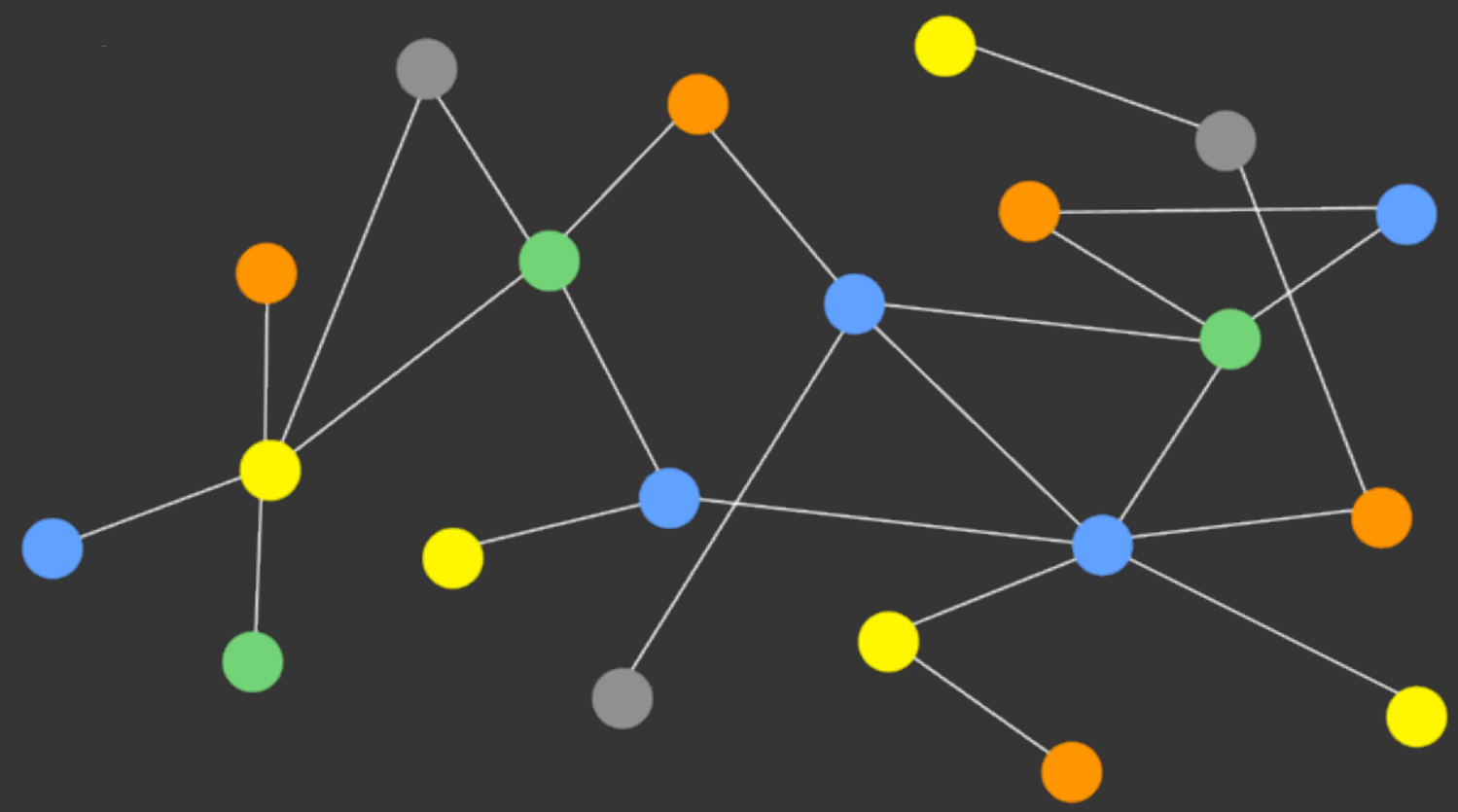


Algorithm that processes the query

webpages - emails - persons - events

Algorithm that processes the query

webpages - emails - persons - events



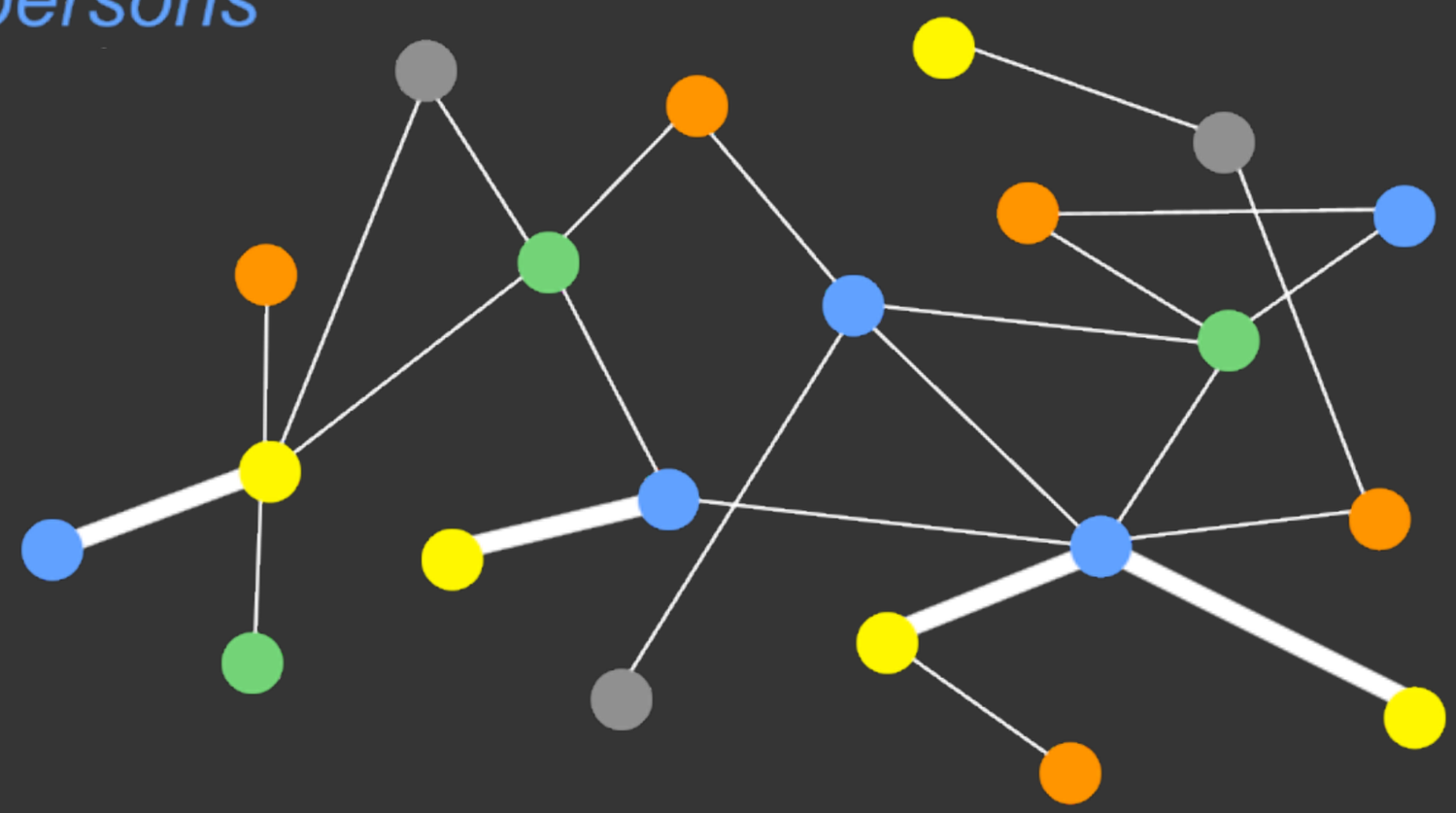
Sample association
graph database

Algorithm that processes the query

webpages - emails - persons - events

Results Generator

webpages - emails - *persons*



Sample association graph database

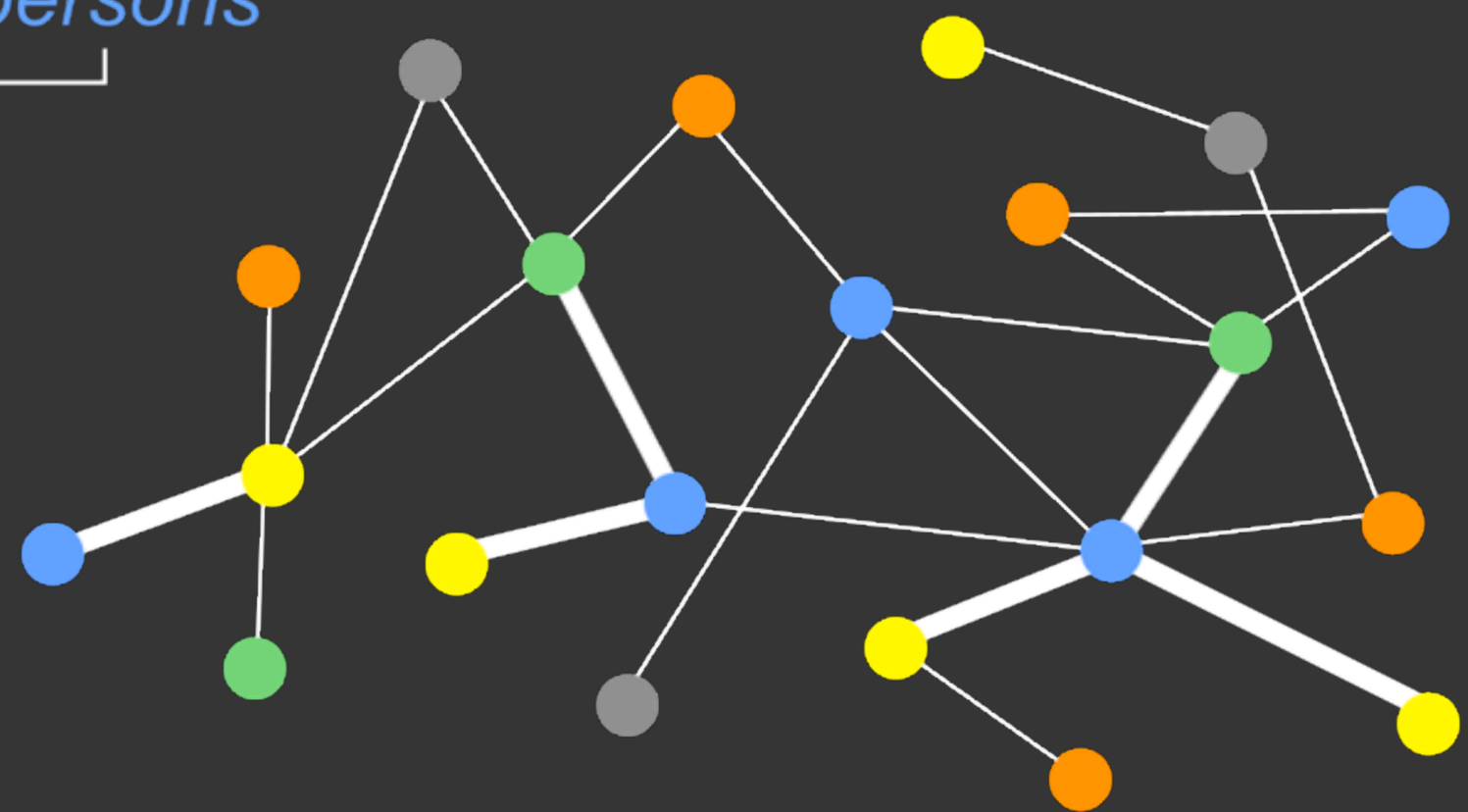
Algorithm that processes the query

webpages - emails - persons - events

Results Generator

webpages - emails - *persons*

webpages - *emails*



Sample association graph database

Algorithm that processes the query

webpages - emails - persons - events

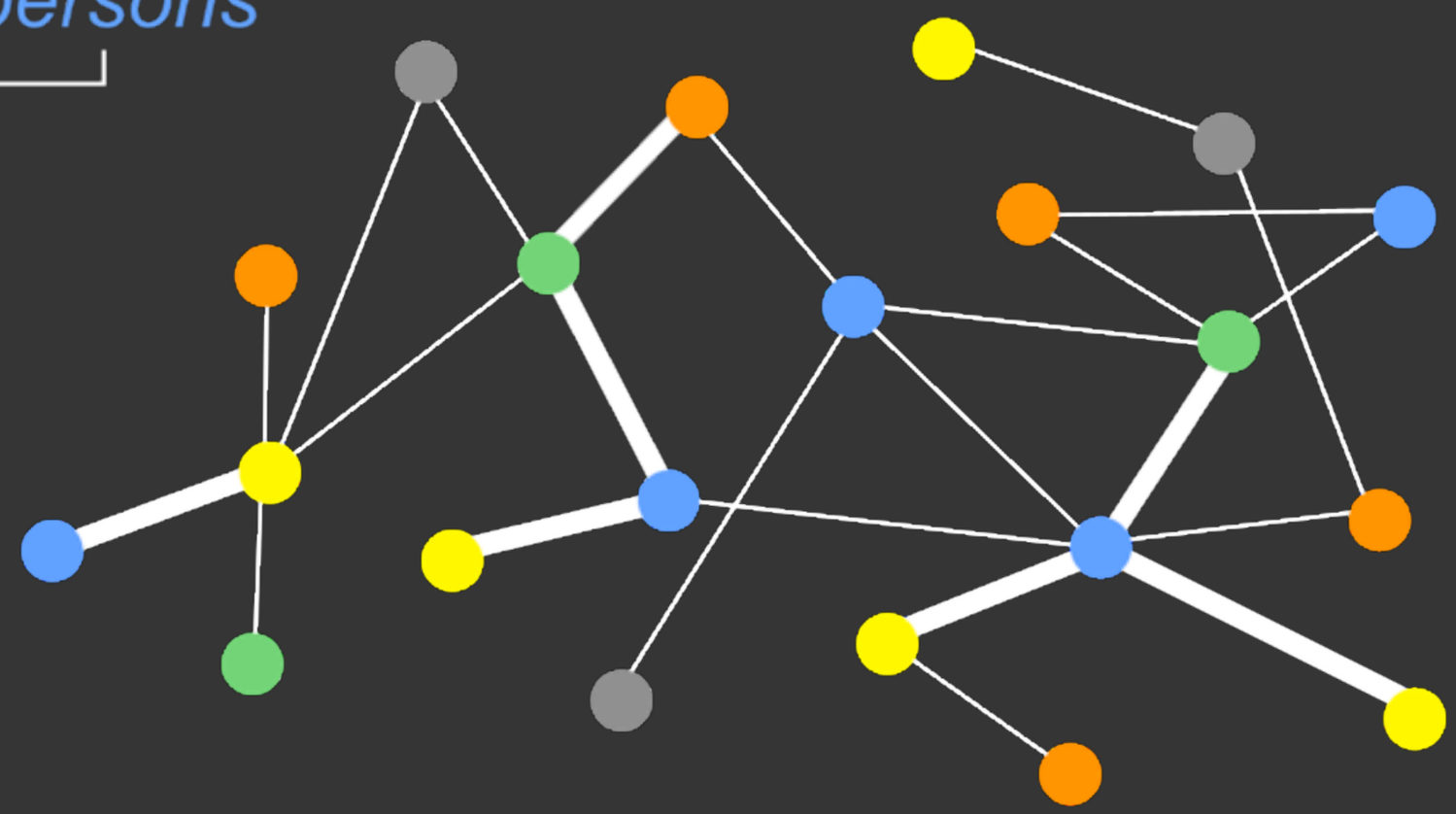
Results Generator

webpages - emails - *persons*

webpages - *emails*

webpages

Sample association graph database



Algorithm that processes the query

webpages - emails - persons - events

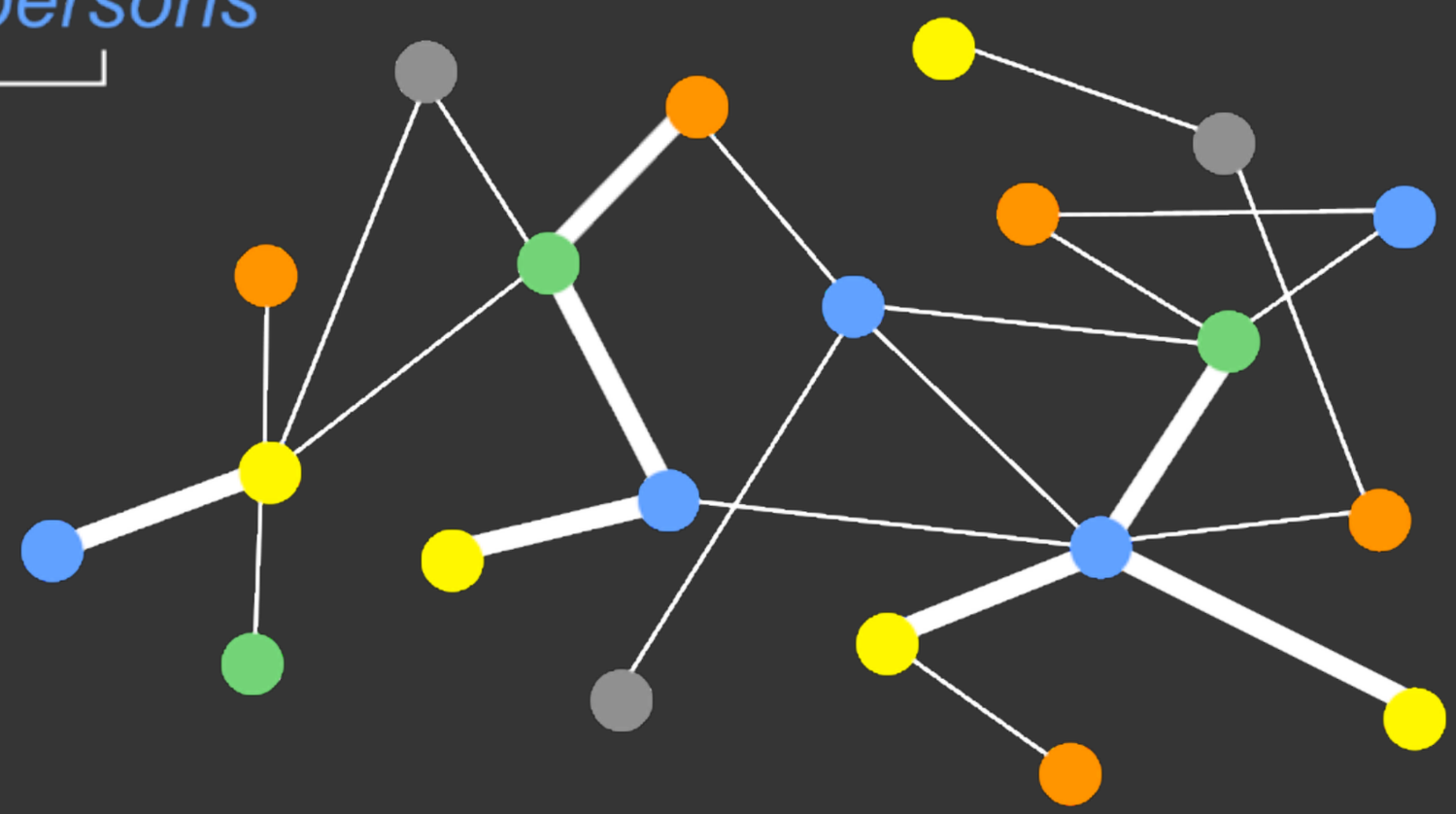
Results Generator

webpages - emails - *persons*

webpages - *emails*

webpages

Sample association graph database



One <u>Results Generator</u> for each pair of associations
For 7 data types → needs 7x7=49 results generators

About the algorithm

About the algorithm

Adding more data types

→Implement more results generators

→Number of generators grows quadratically

About the algorithm

Adding more data types

→Implement more results generators

→Number of generators grows quadratically

But not all combinations make sense

E.g. date related to date doesn't quite make sense

Implemented 38 (out of 49)

About the algorithm

Adding more data types

→Implement more results generators

→Number of generators grows quadratically

But not all combinations make sense

E.g. date related to date doesn't quite make sense

Implemented 38 (out of 49)

Also, there are limited number of data types

Practitioners' guide to building (interactive) applications?

Important that you pick a **good** problem!

- Otherwise, you solve a "non-problem", and nobody cares

Think about scalability early

- e.g., picking a scalable algorithm early on

When building interactive applications, use **iterative** design approach (as in Apolo)

- Why? It's hard to get it right the first time
- Create prototype, evaluate, modify prototype, evaluate, ...
- Quick evaluation helps you identify **important fixes early** (can **save you a lot of time**)

Practitioners' guide to building (interactive) applications?

How to do **iterative** design?

What kinds of **prototypes**?

- Paper prototype, lo-fi prototype, high-fi prototype

What kinds of **evaluation**?

- Ask you friends,
- Lab study (controlled, as in Apolo),
- Longitudinal study (usage over months)
- Deploy it and see the world's reaction!