

15-826: Multimedia (Databases) and Data Mining

Lecture #28: Graph mining -
Belief propagation
Christos Faloutsos

ALSO NOT in the final exam

Sit back and enjoy the show 😊

Final exam

Must-read material

- Danai Koutra, Tai-You Ke, U. Kang, Duen Horng Chau, Hsing-Kuo Kenneth Pao, and Christos Faloutsos. *Unifying guilt-by-association approaches: theorems and fast algorithms.* ECML/PKDD'11, Athens, Greece.



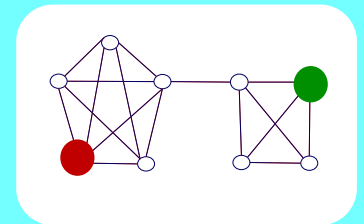
Main outline

- Introduction
- Indexing
- Mining
 - Graphs – patterns
 - ➔ – Graphs – generators and tools
 - ...



Detailed outline - graphs

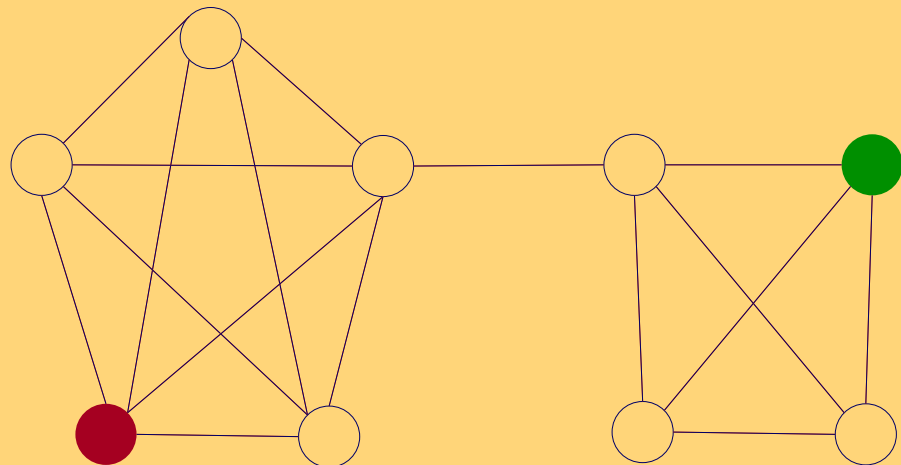
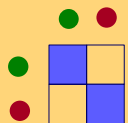
- Graphs – patterns
- Graphs – generators
- Graphs – tools
 - Community detection / graph partitioning
 - ➔ – ‘Belief Propagation’ & fraud detection
 - Motivation - Basics
 - Fast, linear approximation (FaBP)
 - Latest: zooBP
 - Success stories





Problem

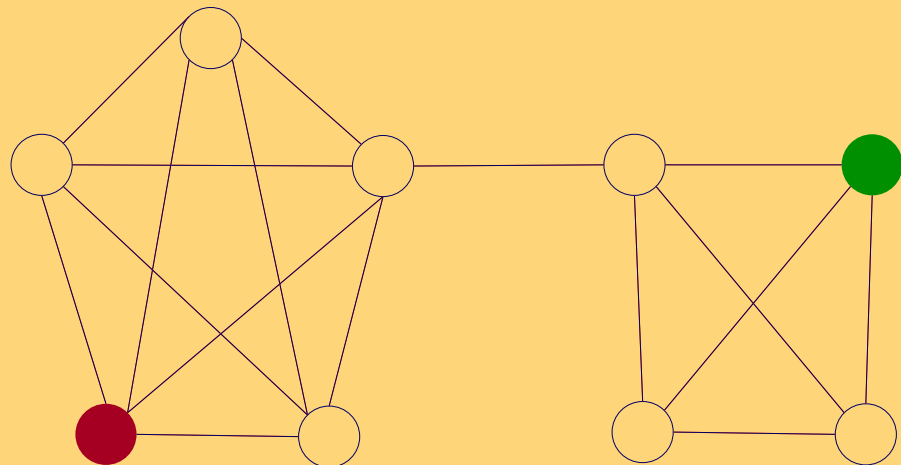
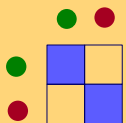
- What color, for the rest?
 - Given homophily (/heterophily etc)?



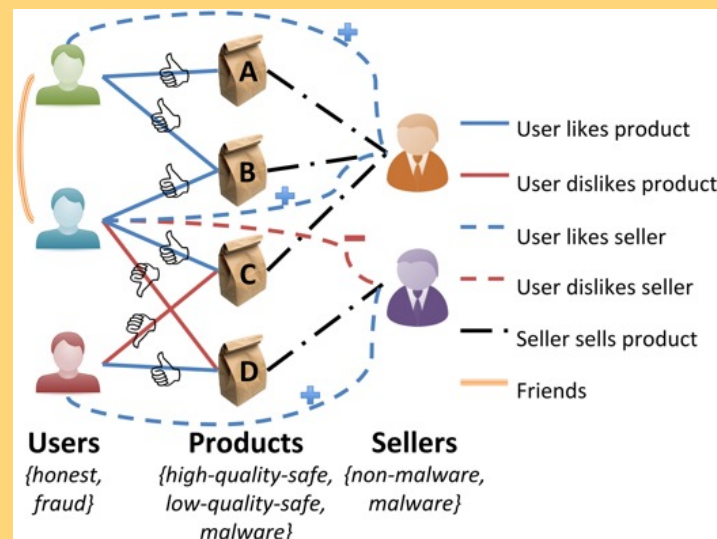


Short answer:

- What color, for the rest?
- A: Belief Propagation ('zooBP')



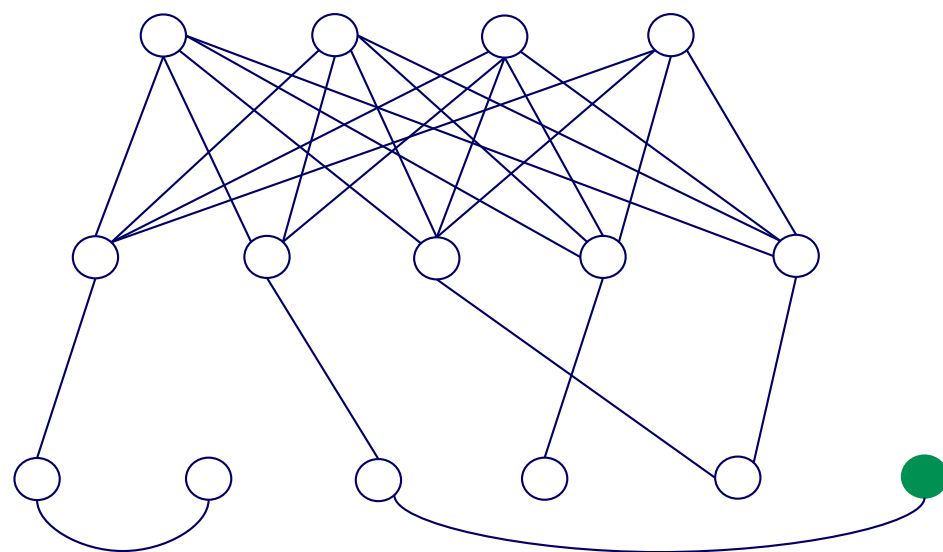
www.cs.cmu.edu/~deswaran/code/zoobp.zip



E-bay Fraud detection

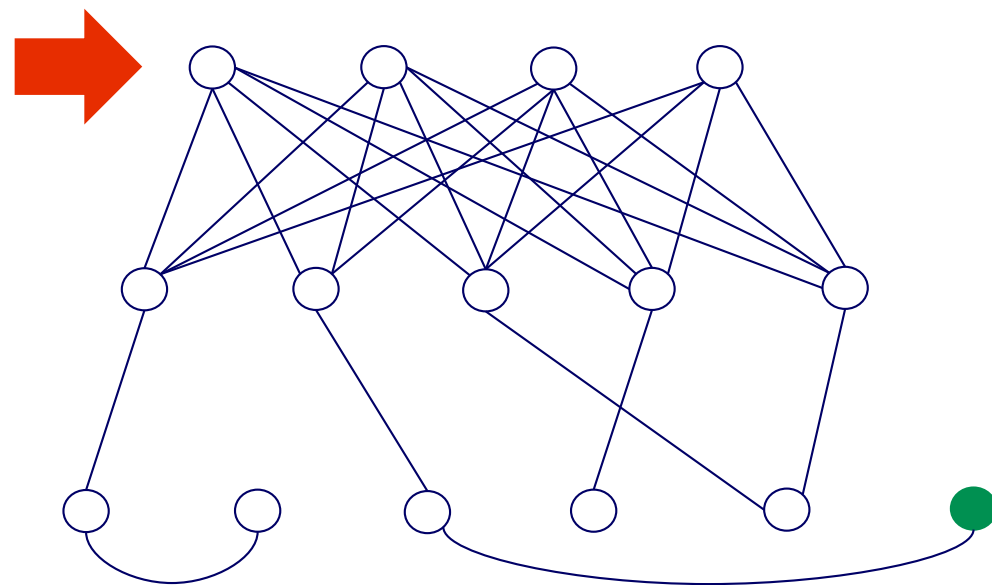


w/ Polo Chau &
Shashank Pandit, CMU
[PKDD'06][WWW'07]

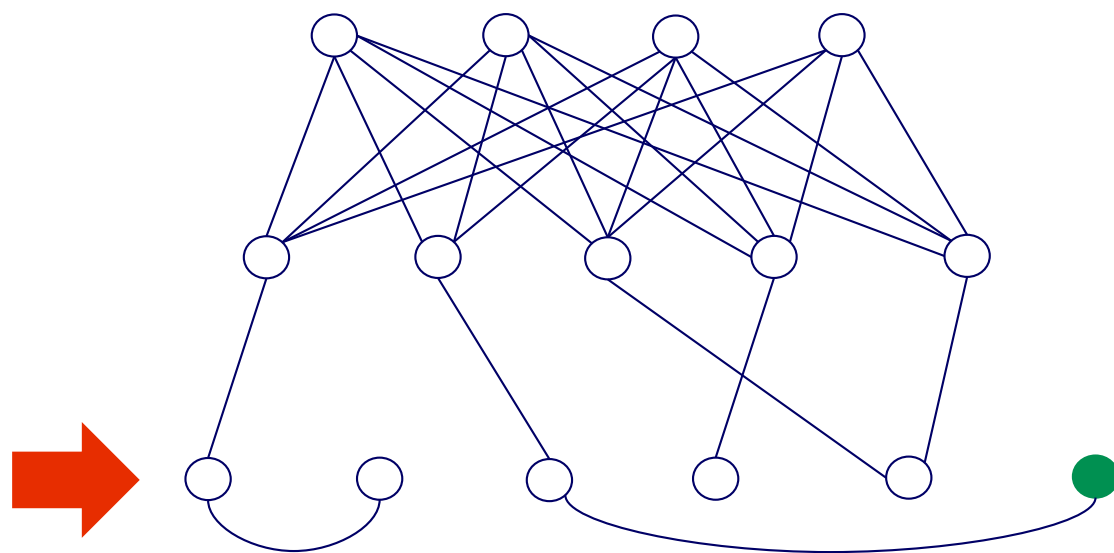


Detecting Fraudulent Personalities in Networks of Online Auctioneers. *Duen Horng (Polo) Chau, Shashank Pandit, and Christos Faloutsos. (PKDD) 2006*

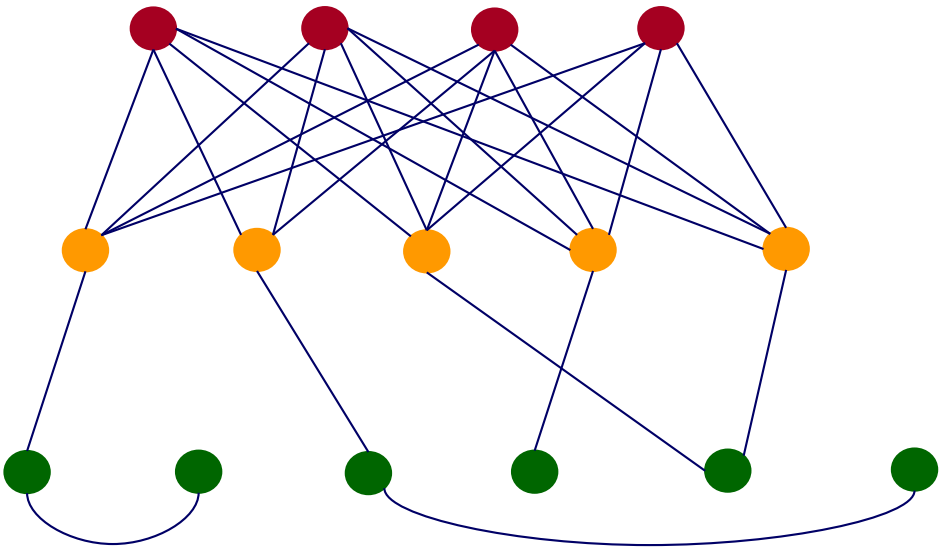
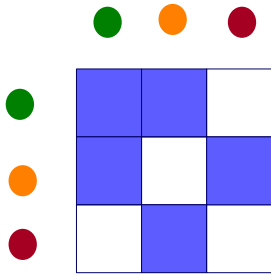
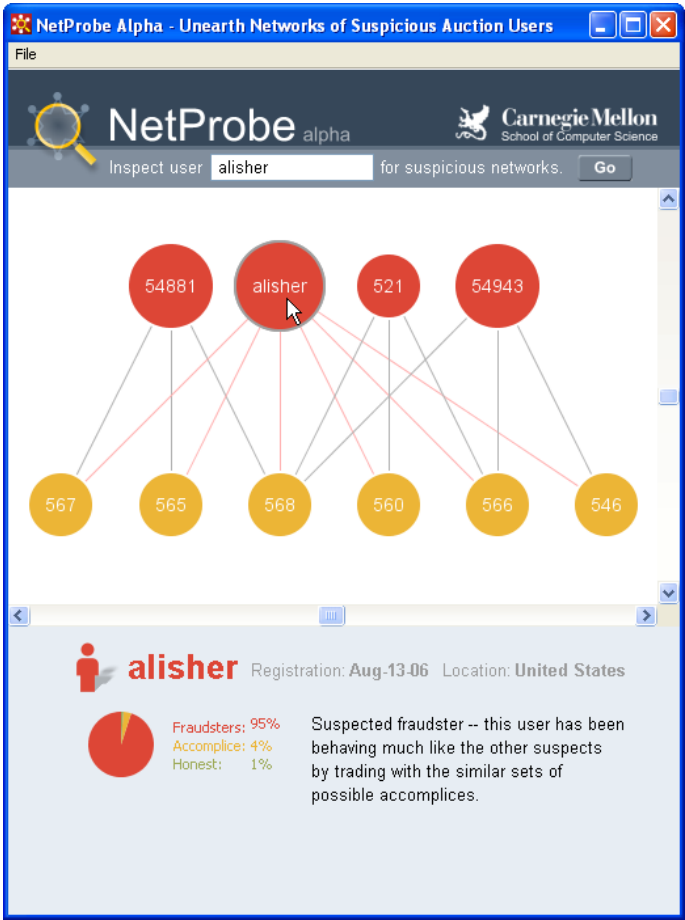
E-bay Fraud detection

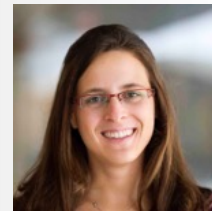


E-bay Fraud detection



E-bay Fraud detection - NetProbe





Prof. Danai Koutra
U. Michigan

Background

Belief Propagation

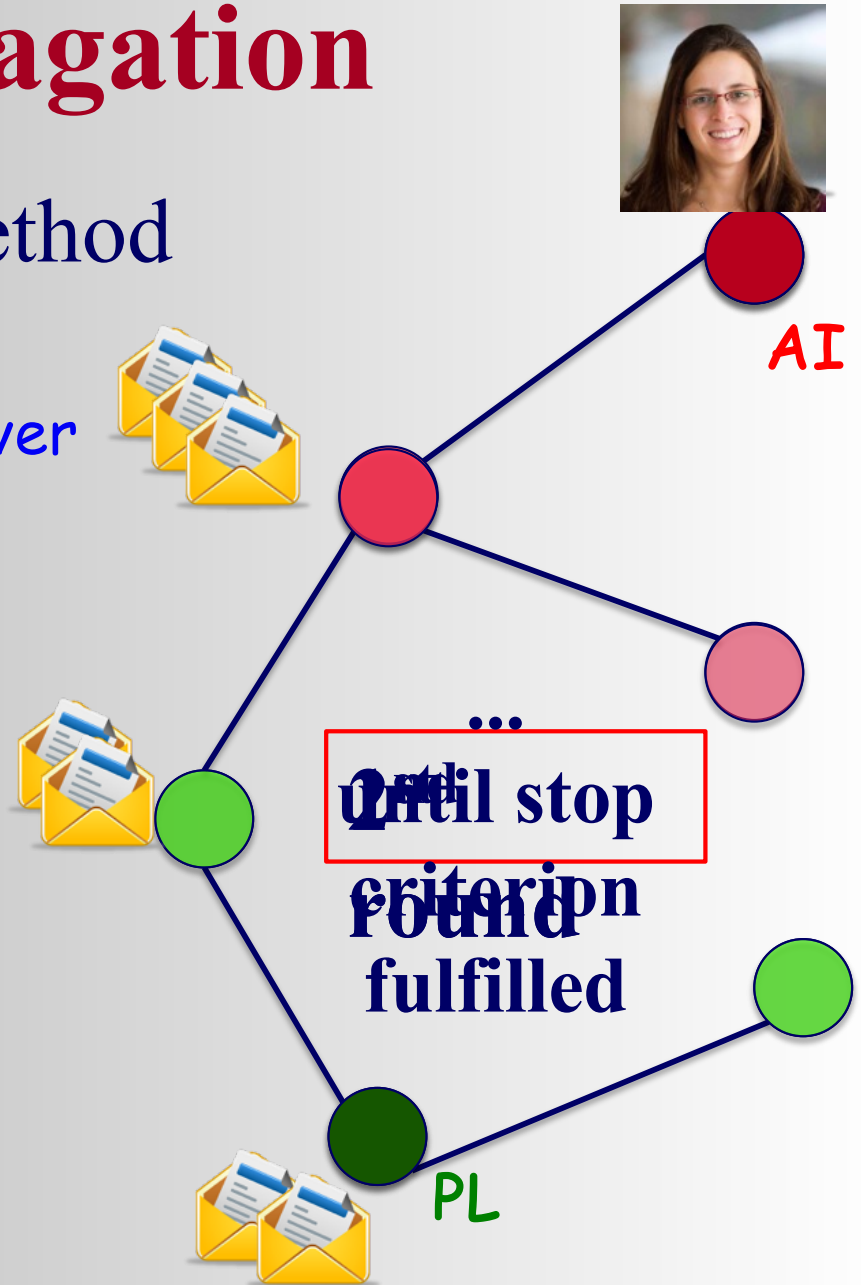
- Iterative message-based method
- “Propagation matrix”:

✧ Homophily

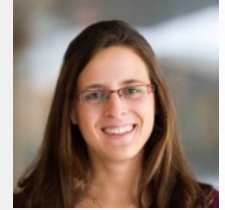
class of receiver

class of sender

		
	0.9	0.1
	0.1	0.9



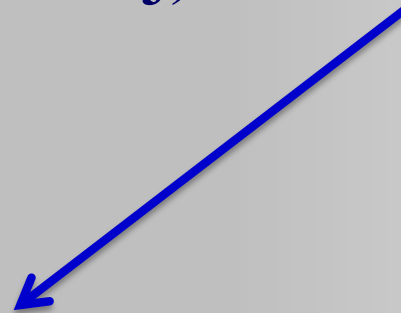
[Pearl '82][Yedidia+ '02] ... [Gonzalez+ '09][Chechetka+ '10]



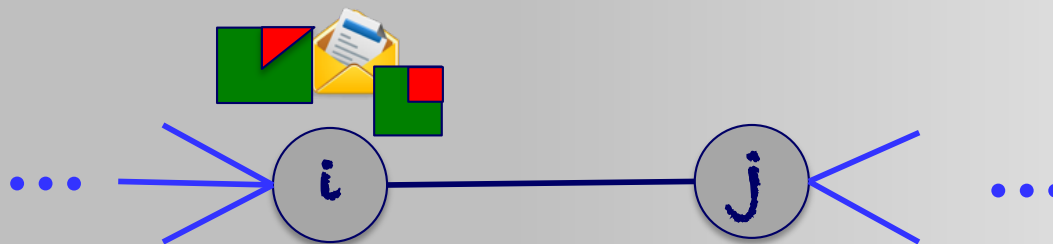
Belief Propagation Equations

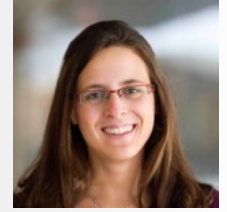


$\text{message}(i \rightarrow j) \approx \text{belief}(i) \square \text{homophily strength}$



		
	0.9	0.1
	0.2	0.8



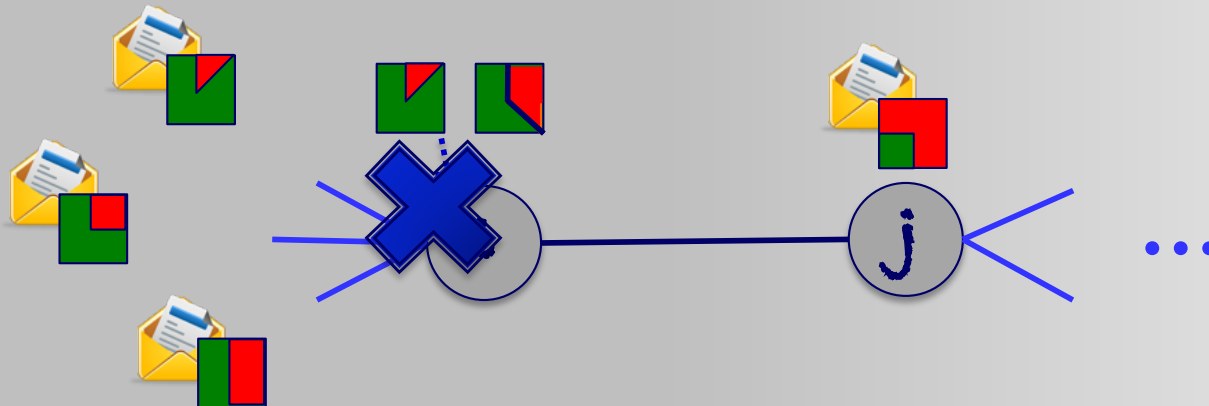


Belief Propagation Equations


 .. belief of i

$$b_i(x_i) \propto \phi_i(x_i) \cdot \prod_{j \in N(i)} m_{ij}(x_i)$$

prior belief
 messages from neighbors



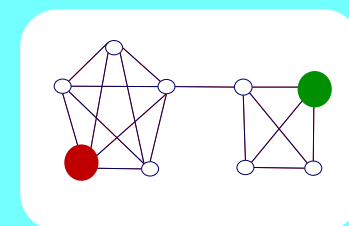
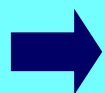


Background

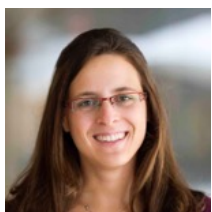


Detailed outline - graphs

- Graphs – patterns
- Graphs – generators
- Graphs – tools
 - Community detection / graph partitioning
 - ‘Belief Propagation’ & fraud detection
 - Motivation - Basics
 - Fast, linear approximation (FaBP)
 - Latest: zooBP
 - Success stories



Unifying Guilt-by-Association Approaches: Theorems and Fast Algorithms



Danai Koutra

U Kang

Hsing-Kuo Kenneth Pao

Tai-You Ke

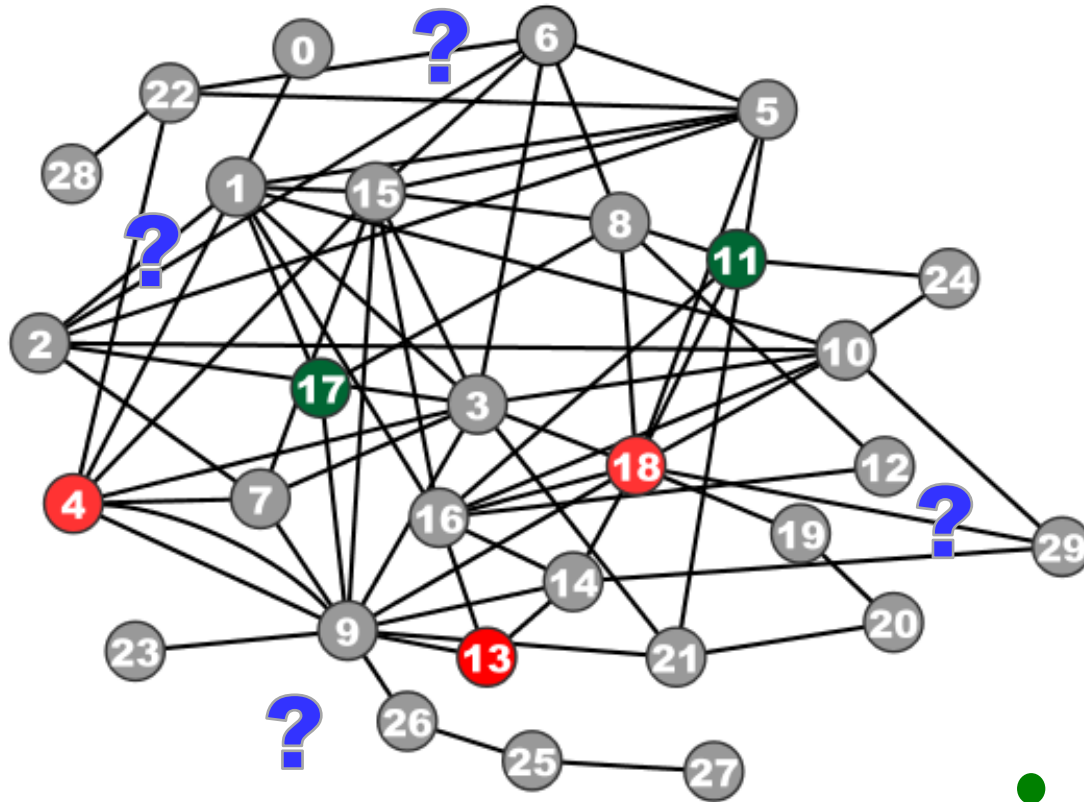
Duen Horng (Polo) Chau

Christos Faloutsos

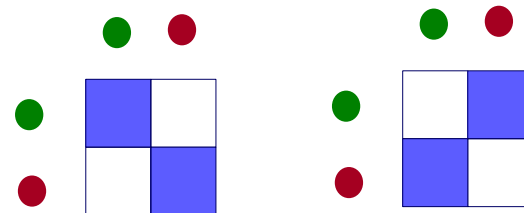
ECML PKDD, 5-9 September 2011, Athens, Greece

Problem Definition:

GBA techniques

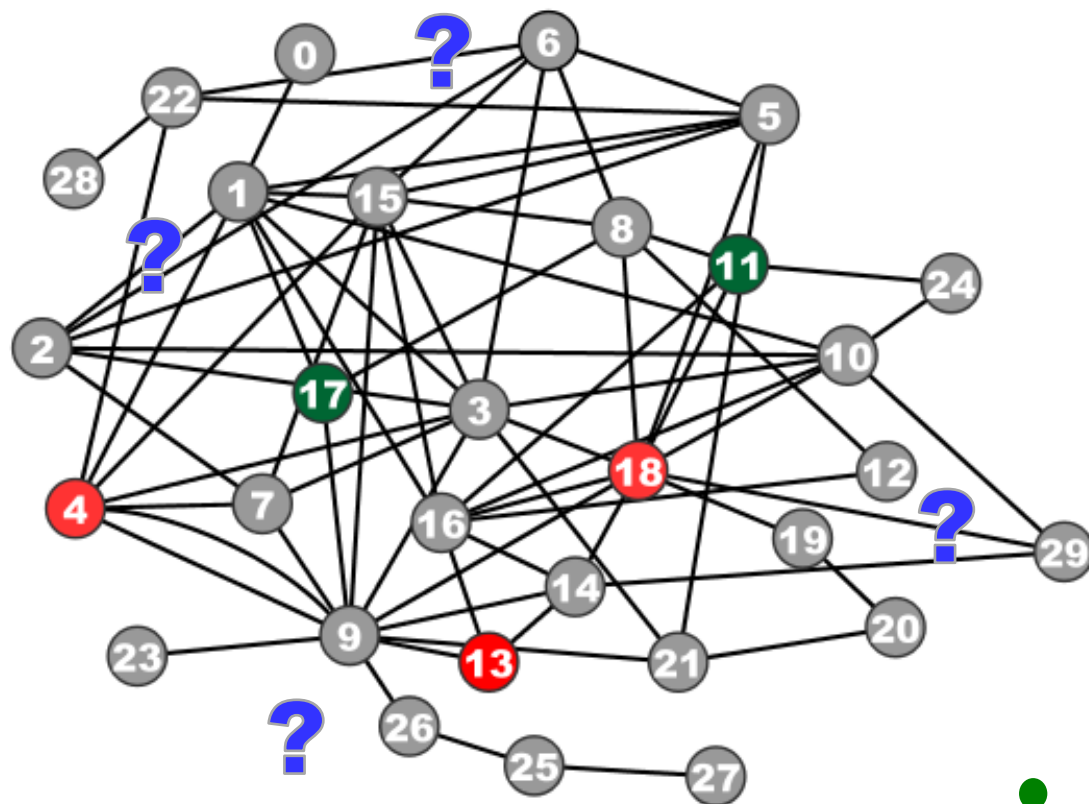


Given: Graph; &
few labeled nodes
Find: labels of rest
(assuming network
effects)

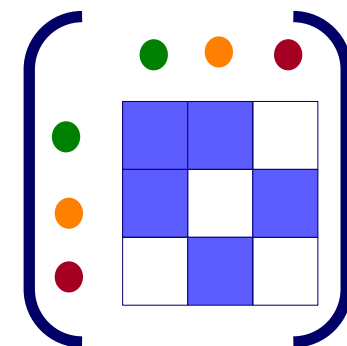
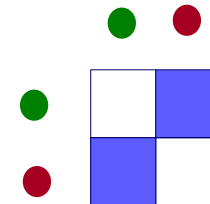
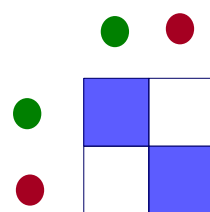


Problem Definition:

GBA techniques



Given: Graph; &
few labeled nodes
Find: labels of rest
(assuming network
effects)



BP vs. Linearized BP

DETAILS

Original [Yedidia+]:

Belief Propagation



$$m_{ij}(x_j) \leftarrow \sum_{x_i} \phi_i(x_i) \cdot \psi_{ij}(x_i, x_j) \cdot \prod_{n \in N(i) \setminus j} m_{ni}(x_i)$$



$$b_i(x_i) \leftarrow \phi_i(x_i) \cdot \prod_{j \in N(i)} m_{ij}(x_j)$$

non-linear

- Closed-form formula?
- Convergence?

BP vs. Linearized BP

DETAILS

Original [Yedidia+]:

Our proposal:

Belief Propagation



$$m_{ij}(x_j) \leftarrow \sum_{x_i} \phi_i(x_i) \cdot \psi_{ij}(x_i, x_j) \cdot \prod_{n \in N(i) \setminus j} m_{ni}(x_i)$$



$$b_i(x_i) \leftarrow \phi_i(x_i) \cdot \prod_{j \in N(i)} m_{ij}(x_j)$$

non-linear

- ✓ Closed-form formula?
- ✓ Convergence?

Linearized BP

BP is approximated by

$$[\mathbf{I} + a\mathbf{D} - c'\mathbf{A}] \mathbf{b}_h = \phi_h$$

$\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}$	$\begin{bmatrix} d1 & & \\ & d2 & \\ & & d3 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	$?$	$\begin{bmatrix} 0 \\ -10^{-2} \\ 10^{-2} \end{bmatrix}$
---	--	---	-----	--

linear

Are they related?



- RWR (Random Walk with Restarts)
 - google's pageRank (*'if my parents are important, I'm important, too'*)
- SSL (Semi-supervised learning)
 - minimize the differences among neighbors
- BP (Belief propagation)
 - send messages to neighbors, on what you believe about them

Are they related? **YES!**



- RWR (Random Walk with Restarts)
 - google's pageRank (*'if my parents are important, I'm important, too'*)
- SSL (Semi-supervised learning)
 - minimize the differences among neighbors
- BP (Belief propagation)
 - send messages to neighbors, on what you believe about them

Correspondence of Methods

$$\mathbf{p} = (1-c)/n \quad [\mathbf{I} - c \mathbf{B}]^{-1} \vec{e}$$

Method	Matrix	Unknown	known
RWR	$[\mathbf{I} - c \mathbf{A} \mathbf{D}^{-1}] \times$	$\mathbf{x}$	$(1-c)\mathbf{y}$
SSL	$[\mathbf{I} + a(\mathbf{D} - \mathbf{A})] \times$	$\mathbf{x}$	$\mathbf{y}$
FABP	$[\mathbf{I} + a \mathbf{D} - c' \mathbf{A}] \times$	$\mathbf{b}_h$	ϕ_h

$$\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}$$

$$\begin{bmatrix} d1 & & \\ & d2 & \\ & & d3 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} ? \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

adjacency
matrix

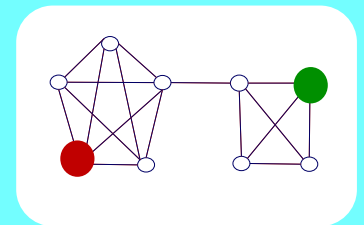
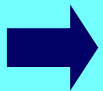
final
labels/
beliefs

prior
labels/
beliefs



Detailed outline - graphs

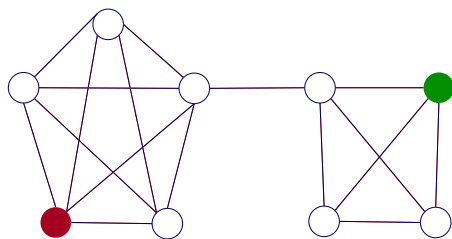
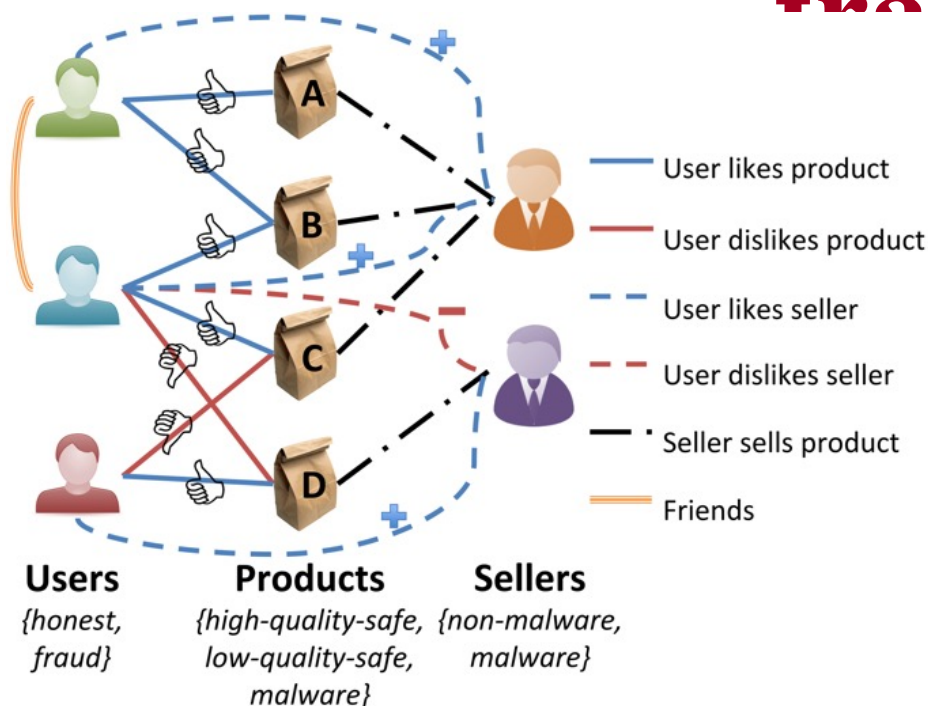
- Graphs – patterns
- Graphs – generators
- Graphs – tools
 - Community detection / graph partitioning
 - ‘Belief Propagation’ & fraud detection
 - Motivation - Basics
 - Fast, linear approximation (FaBP)
 - Latest: zooBP
 - Success stories



Problem: e-commerce ratings

fraud

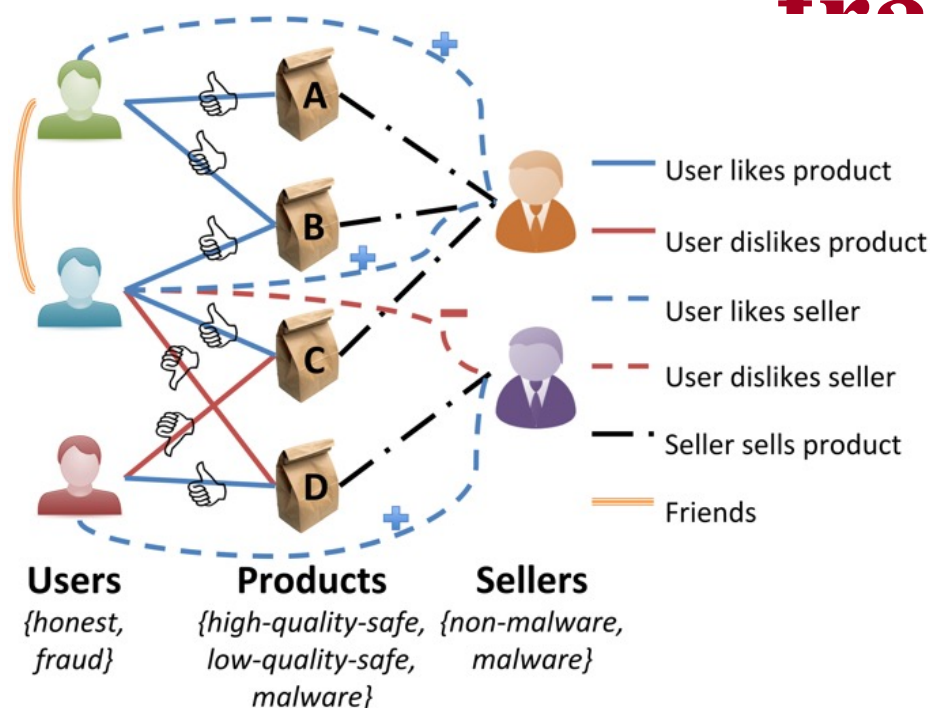
- Given a heterogeneous graph on users, products, sellers and positive/negative ratings with “seed labels”
- Find the top k most fraudulent users, products and sellers



Problem: e-commerce ratings

fraud

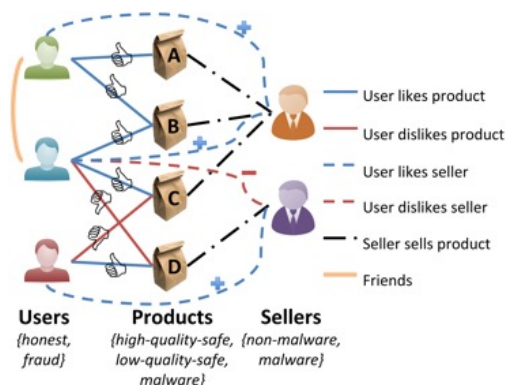
- Given a heterogeneous graph on users, products, sellers and positive/negative ratings with “seed labels”
- Find the top k most fraudulent users, products and sellers



Dhivya Eswaran, Stephan Günnemann, Christos Faloutsos,
 “ZooBP: Belief Propagation for Heterogeneous Networks”,
 VLDB 2017

Problem: e-commerce rating fraud

DETAILS



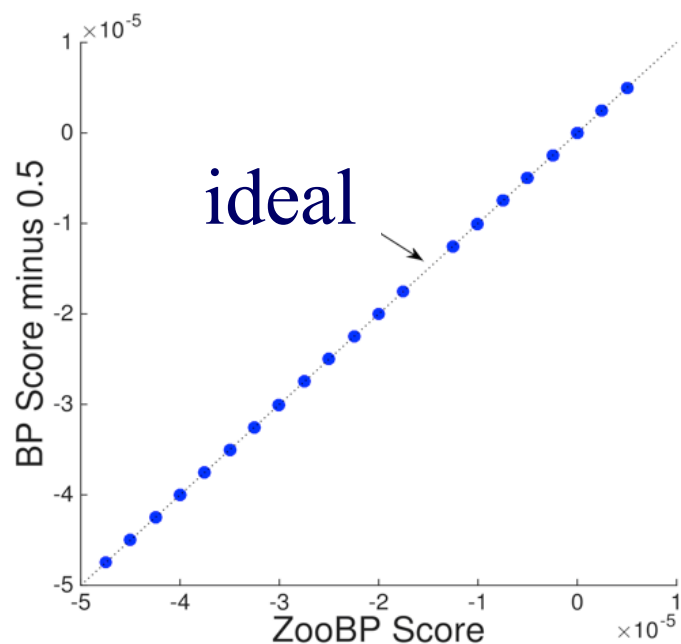
Theorem 1 (ZooBP). *If $\mathbf{b}, \mathbf{e}, \mathbf{P}, \mathbf{Q}$ are constructed as described above, the linear equation system approximating the final node beliefs given by BP is:*

$$\mathbf{b} = \mathbf{e} + (\mathbf{P} - \mathbf{Q})\mathbf{b} \quad (\text{ZooBP}) \quad (10)$$

Dhivya Eswaran, Stephan Günnemann, Christos Faloutsos,
 “ZooBP: Belief Propagation for Heterogeneous Networks”,
 VLDB 2017

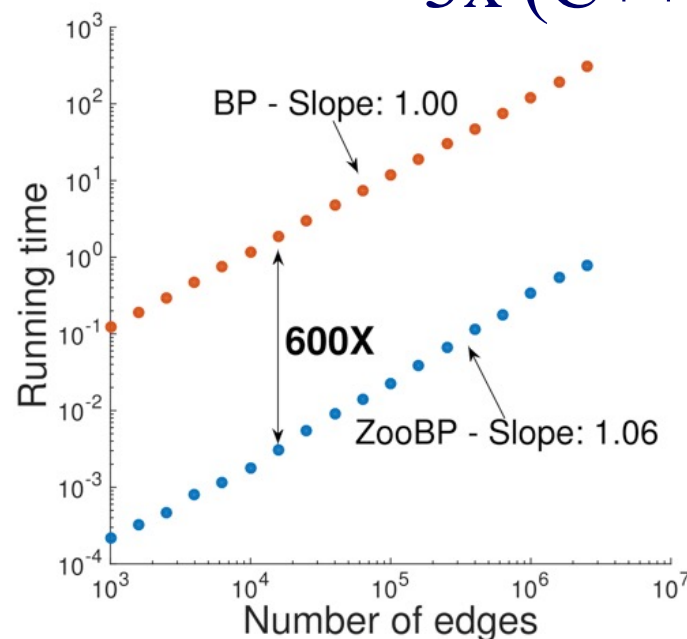
ZooBP: features

Fast; convergence guarantees.



Near-perfect accuracy

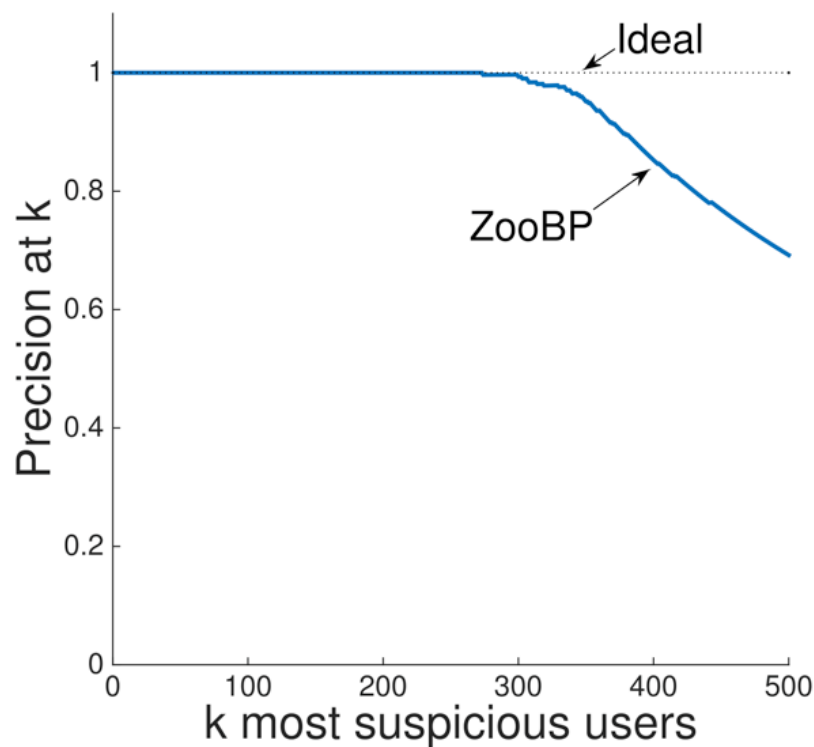
600x (matlab)
3x (C++)



linear in graph size

Dhivya Eswaran, Stephan Günnemann, Christos Faloutsos,
“ZooBP: Belief Propagation for Heterogeneous Networks”,
VLDB 2017

ZooBP in the real world



- Near 100% precision on top 300 users (Flipkart)
- Flagged users: suspicious
 - 400 ratings in 1 sec
 - 5000 good ratings and no bad ratings

Dhivya Eswaran, Stephan Günnemann, Christos Faloutsos,
“ZooBP: Belief Propagation for Heterogeneous Networks”,
VLDB 2017

ZooBP: code etc

<http://www.cs.cmu.edu/~deswaran/code/zoobp.zip>

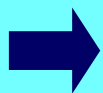
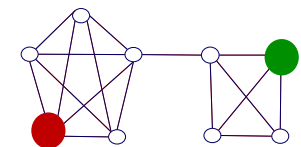


Dhivya Eswaran, Stephan Günnemann, Christos Faloutsos,
“*ZooBP*: Belief Propagation for Heterogeneous Networks”,
VLDB 2017



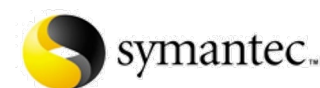
Detailed outline - graphs

- Graphs – patterns
- Graphs – generators
- Graphs – tools
 - Community detection / graph partitioning
 - ‘Belief Propagation’ & fraud detection
 - Motivation - Basics
 - Fast, linear approximation (FaBP)
 - Latest: zooBP
 - Success stories



Other ‘success stories’?

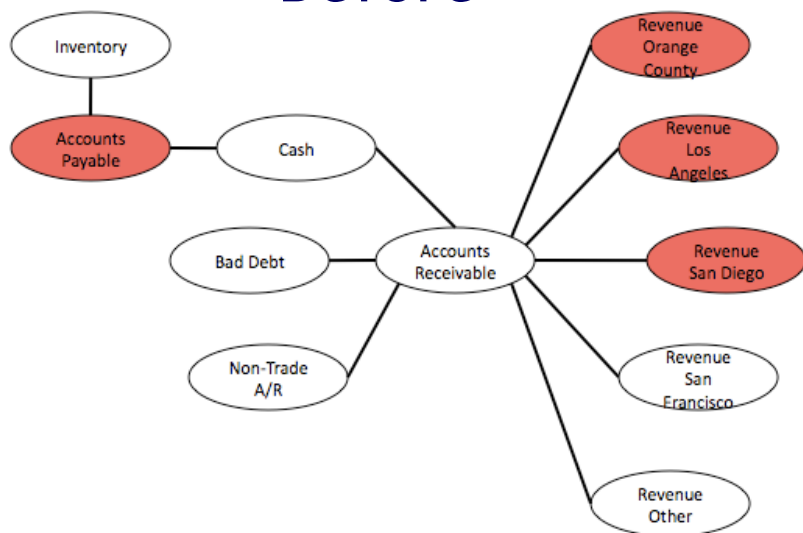
- Accounting fraud
- Malware detection



Network Effect Tools: SNARE

- Some accounts are sort-of-suspicious – how to combine weak signals?

Before



Mary McGlohon, Stephen Bay, Markus G. Anderle, David M. Steier, Christos Faloutsos: *SNARE: a link analytic system for graph labeling and risk detection*. KDD 2009: 1265-1274

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

SDM 2011, Mesa, Arizona



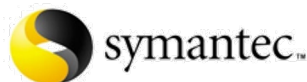
Polo Chau

Machine Learning Dept



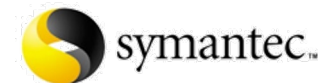
Carey Nachenberg

Vice President & Fellow



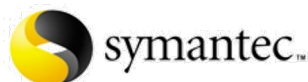
Jeffrey Wilhelm

Principal Software Engineer



Adam Wright

Software Engineer



Prof. Christos Faloutsos

Computer Science Dept

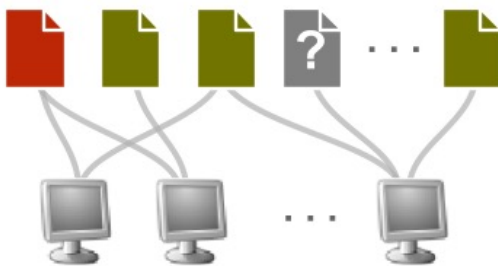
Polonium: The Data



60+ terabytes of data *anonymously*
contributed by participants of worldwide
Norton Community Watch program

50+ million machines

900+ million executable files



Constructed a machine-file bipartite
graph (0.2 TB+)

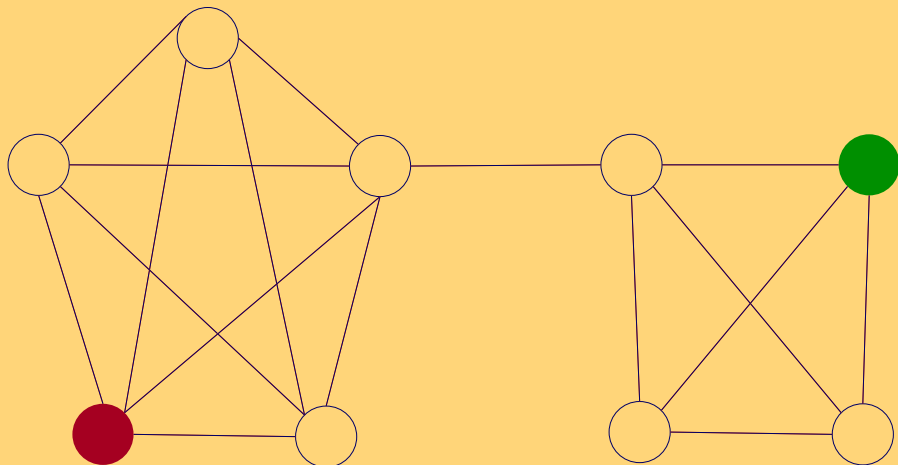
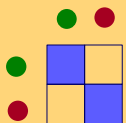
1 billion nodes (machines and files)

37 billion edges



Short answer:

- What color, for the rest?
- A: Belief Propagation ('zooBP')



www.cs.cmu.edu/~deswaran/code/zoobp.zip

