Carnegie Mellon

15-826: Multimedia Databases and Data Mining

Bonus Lecture : Anomaly detection in large graphs

Christos Faloutsos

Carnegie Mellon

Repeat: BONUS LECTURE

I.e., you may skip it, and still get 100/100 in the course

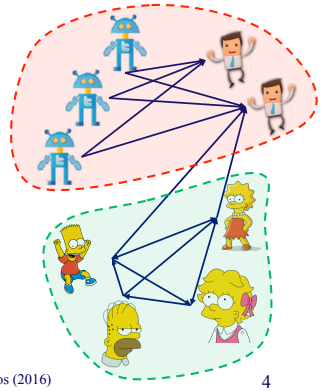
Carnegie Mellon

Roadmap

- Introduction – Motivation
- Unsupervised
  - static graphs
  - Time-evolving graphs
- Supervised: BP
- Conclusions

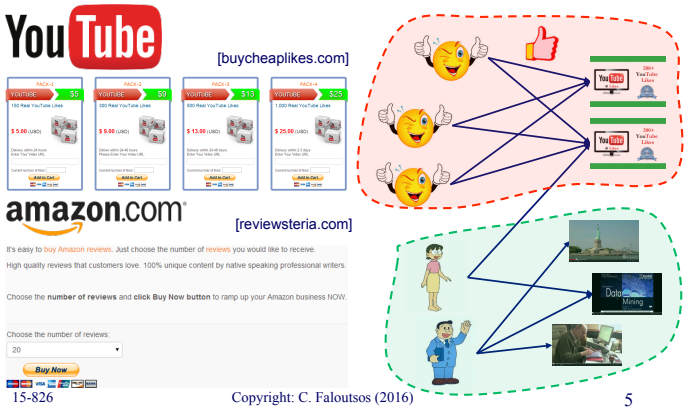
Carnegie Mellon

Fraud Detection: Graph Analysis Problem



Carnegie Mellon

Fraud Detection: Graph Analysis Problem



Carnegie Mellon

Other applications:

- Cyber-security (DDoS attacks – distributed denial of service)
- Abnormal individuals / social groups
  - dysfunctional departments in a company
  - Onset of depression/hijacking wrt a FB user
  - …

Carnegie Mellon

Roadmap

- Introduction – Motivation
- Unsupervised
  - static graphs
    - Eigenspokes, and extensions
    - fbox
  - Time-evolving graphs
- Supervised: BP
- Conclusions

Carnegie Mellon

EigenSpokes

B. Aditya Prakash, Mukund Seshadri, Ashwin Sridharan, Sridhar Machiraju and Christos Faloutsos: *EigenSpokes: Surprising Patterns and Scalable Community Chipping in Large Graphs,* PAKDD 2010, Hyderabad, India, 21-24 June 2010.

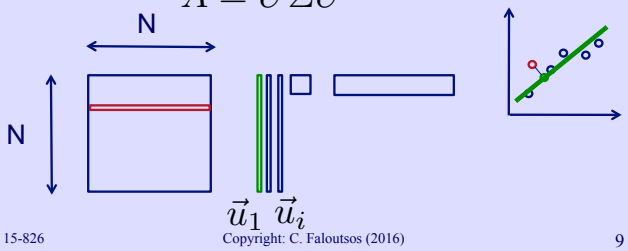
Carnegie Mellon

details

EigenSpokes

- Eigenvectors of adjacency matrix
  - equivalent to singular vectors (symmetric, undirected graph)

$$A = U\Sigma U^T$$

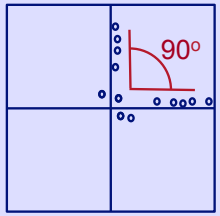


 Copyright: C. Faloutsos (2016)

Carnegie Mellon

EigenSpokes

- EE plot:
- Scatter plot of scores of u1 vs u2
- One would expect
  - Many points @ origin
  - A fe[illegible]ttered ~r[illegible]ly



 Copyright: C. Faloutsos (2016)

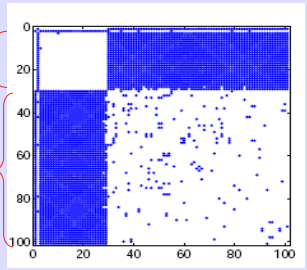
Carnegie Mellon

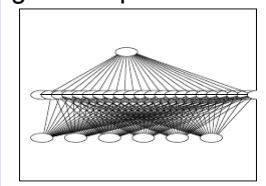
Bipartite Communities!

patents from same inventor(s)

`cut-and-paste' bibliography!

magnified bipartite community





 Copyright: C. Faloutsos (2016)

Carnegie Mellon

Roadmap

- Introduction – Motivation
- Unsupervised
  - static graphs
    - Eigenspokes, and extensions
    - fbox
  - Time-evolving graphs
- Supervised: BP
- Conclusions

 Copyright: C. Faloutsos (2016)

Carnegie Mellon

Inferring Strange Behavior from Connectivity Pattern in Social Networks

Meng Jiang, Peng Cui, Shiqiang Yang
(Tsinghua, Beijing)
Alex Beutel, Christos Faloutsos (CMU)

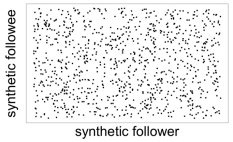
清華大學 Tsinghua University Carnegie Mellon University

Carnegie Mellon

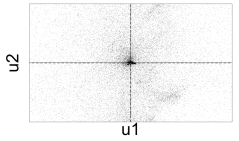
Lockstep and *Spectral Subspace Plot*

- Case #0: No lockstep behavior in random power law graph of 1M nodes, 3M edges
- Random ⟷ "Scatter"

Adjacency Matrix



Spectral Subspace Plot

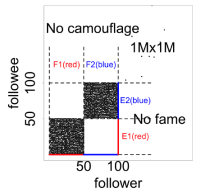


Carnegie Mellon

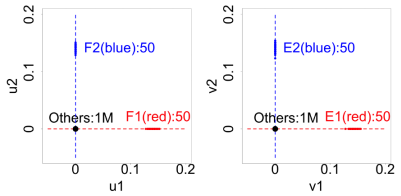
Lockstep and *Spectral Subspace Plot*

- Case #1: non-overlapping lockstep
- "Blocks" ⟷ "Rays"

Adjacency Matrix



Spectral Subspace Plot



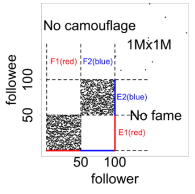
Rule 1 (short "rays"): two blocks, high density (90%), no "camouflage", no "fame"

Carnegie Mellon

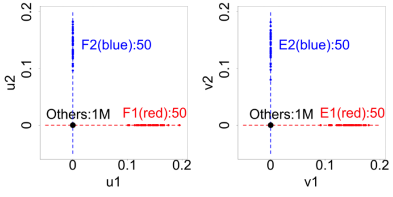
Lockstep and *Spectral Subspace Plot*

- Case #2: non-overlapping lockstep
- "Blocks; low density" ⟷ Elongation

Adjacency Matrix



Spectral Subspace Plot



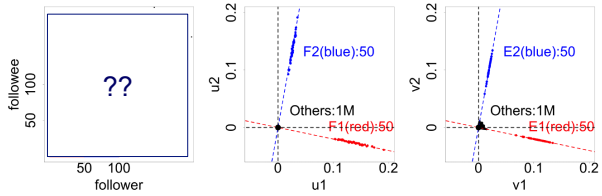
Rule 2 (long "rays"): two blocks, low density (50%), no "camouflage", no "fame"

Carnegie Mellon

Lockstep and *Spectral Subspace Plot*

- Case #3: non-overlapping lockstep

Adjacency Matrix Spectral Subspace Plot

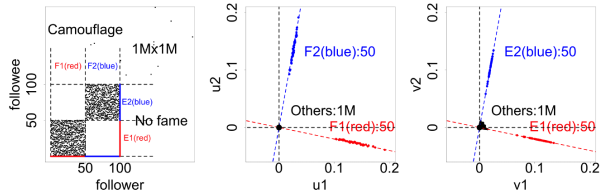


Carnegie Mellon

Lockstep and *Spectral Subspace Plot*

- Case #3: non-overlapping lockstep
- "**Camouflage**" (or "Fame") ⟷ Tilting "Rays"

Adjacency Matrix Spectral Subspace Plot



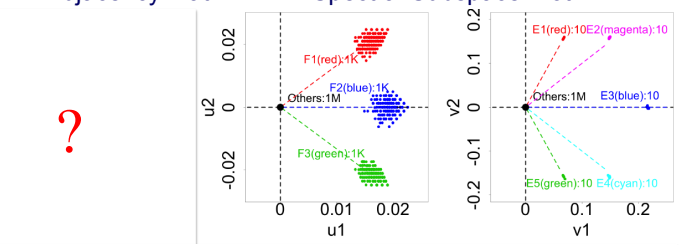
Rule 3 (tilting "rays"): two blocks, with "camouflage", no "fame"

Carnegie Mellon

Lockstep and *Spectral Subspace Plot*

- Case #4: ? lockstep
- "?" ⟷ "Pearls"

Adjacency Matrix Spectral Subspace Plot

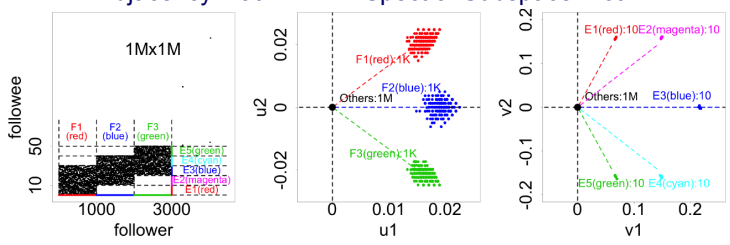


Carnegie Mellon

Lockstep and *Spectral Subspace Plot*

- Case #4: **overlapping** lockstep
- "**Staircase**" ⟷ "Pearls"

Adjacency Matrix Spectral Subspace Plot



Rule 4 ("pearls"): a "staircase" of three partially overlapping blocks.

Carnegie Mellon

Roadmap

- Introduction – Motivation
- Unsupervised
  - static graphs
    - Eigenspokes, and extensions
    - fbox
  - Time-evolving graphs
- Supervised: BP
- Conclusions

Carnegie Mellon

Issue: Flying below the radar

- Plotting the top k eigenspokes, will miss small groups
- Q: Can we spot them, too?
- A: 'fBox'

Carnegie Mellon

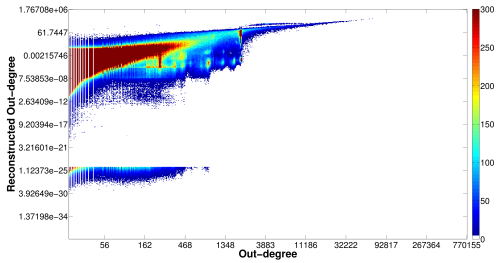
Problem: Social Network Link Fraud

Target: find "stealthy" attackers missed by other algorithms

Clique

41.7M nodes
1.5B edges

Bipartite core



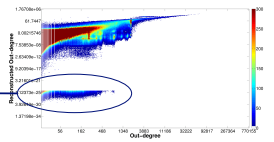
Carnegie Mellon

Problem: Social Network Link Fraud

Target: find "stealthy" attackers missed by other algorithms

Lekan Olawole Lowe @loweinc 26 Jul 09
Sign up free and Get 400 followers a day using http://tweeteradder.com

Lekan Olawole Lowe @loweinc 26 Jul 09
Get 400 followers a day using http://www.tweeterfollow.com



Takeaway: use *reconstruction error* between true/latent representation!

Neil Shah, Alex Beutel, Brian Gallagher and Christos Faloutsos. *Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective.* ICDM 2014, Shenzhen, China.

Carnegie Mellon

Roadmap

- Introduction – Motivation
- Unsupervised
  - – static graphs
    - Eigenspokes, and extensions
    - fBox
    - Spot strange, individual nodes: oddBall
  - – Time-evolving graphs
- Supervised: BP
- Conclusions

Carnegie Mellon

OddBall: Spotting Anomalies in Weighted Graphs

Leman Akoglu, Mary McGlohon, Christos Faloutsos

Carnegie Mellon University

School of Computer Science

PAKDD 2010, Hyderabad, India

Carnegie Mellon

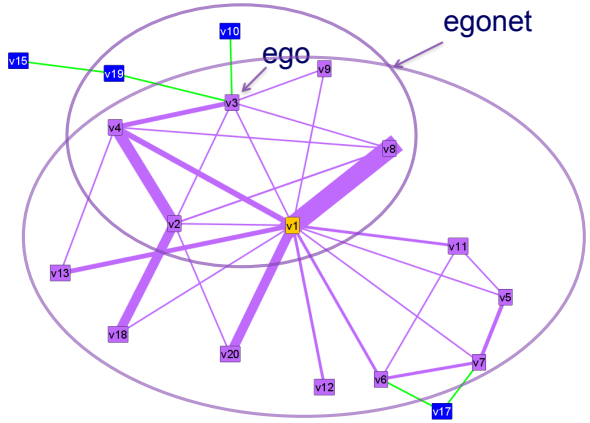
Main idea

For each node,

- extract 'ego-net' (=1-step-away neighbors)
- Extract features (#edges, total weight, etc etc)
- Compare with the rest of the population

Carnegie Mellon

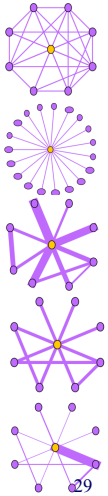
What is an egonet?



Carnegie Mellon

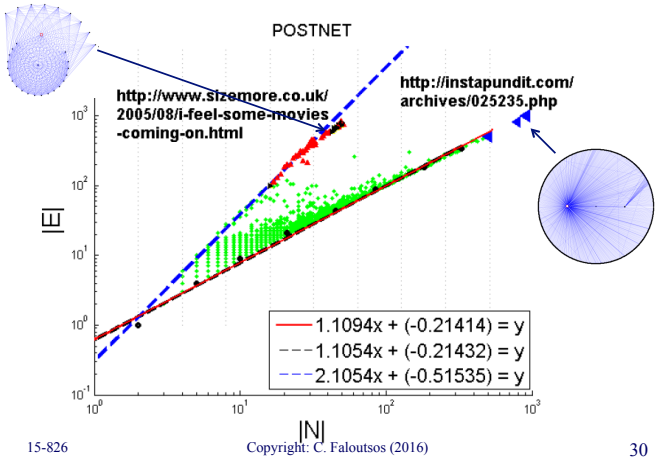
Selected Features

- N_i: number of neighbors (degree) of ego i
- E_i: number of edges in egonet i
- W_i: total weight of egonet i
- $\lambda_{w,i}$: principal eigenvalue of the weighted adjacency matrix of egonet I



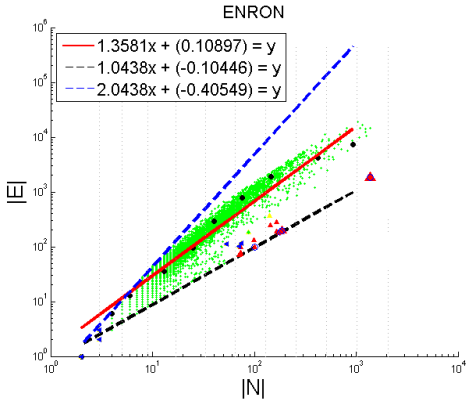
Carnegie Mellon

Near-Clique/Star



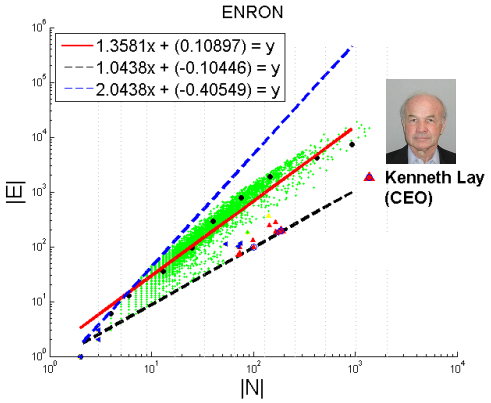
Carnegie Mellon

Near-Clique/Star



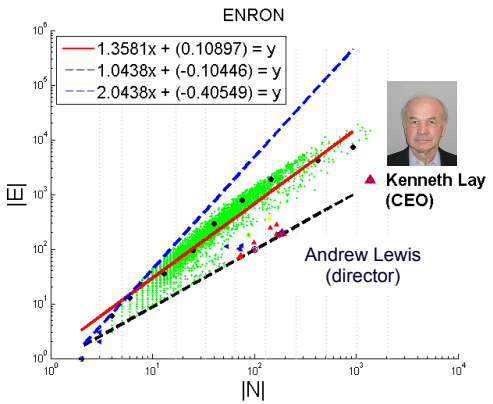
Carnegie Mellon

Near-Clique/Star



Carnegie Mellon

Near-Clique/Star



Carnegie Mellon

Roadmap

- Introduction – Motivation
- Unsupervised
  - static graphs
  - Time-evolving graphs
    - 'copyCatch' in FaceBook
    - Patterns of IAT
    - [tensors]
- Supervised: BP
- Conclusions

Carnegie Mellon

Fraud

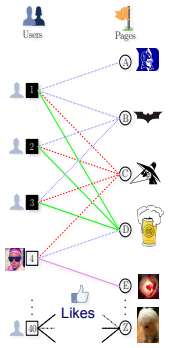
- Given
  - Who 'likes' what page, and when
- Find
  - Suspicious users and suspicious products

CopyCatch: Stopping Group Attacks by Spotting Lockstep Behavior in Social Networks, Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, Christos Faloutsos *WWW, 2013.*

Carnegie Mellon

Fraud

- Given
  - Who 'likes' what page, and when
- Find
  - Suspicious users and suspicious products



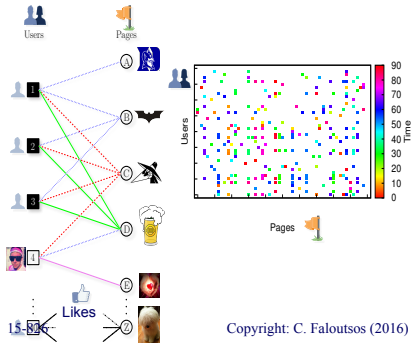
CopyCatch: Stopping Group Attacks by Spotting Lockstep Behavior in Social Networks, Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, Christos Faloutsos *WWW, 2013.*

Carnegie Mellon

Graph Patterns and Lockstep Behavior

Our intuition

- Lockstep behavior: Same Likes, same time



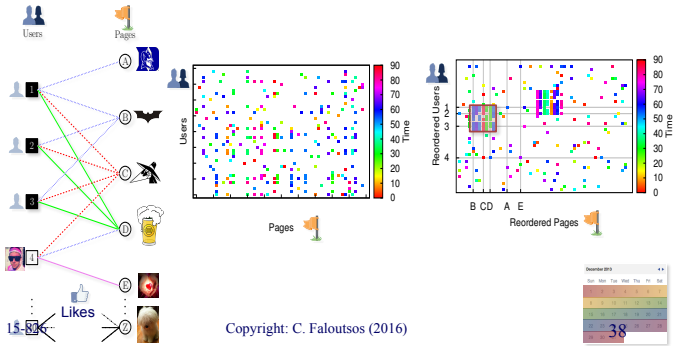
15-826 Copyright: C. Faloutsos (2016) 37

Carnegie Mellon

Graph Patterns and Lockstep Behavior

Our intuition

- Lockstep behavior: Same Likes, same time



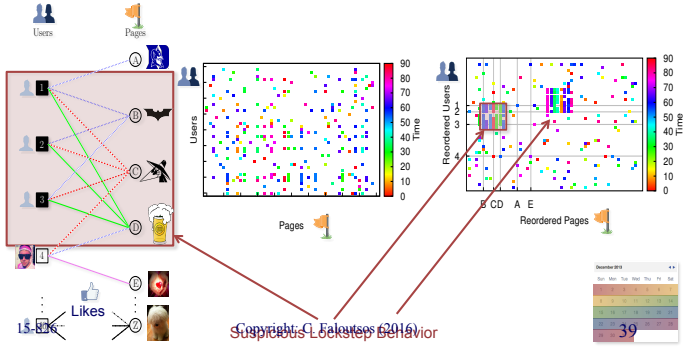
15-826 Copyright: C. Faloutsos (2016) 38

Carnegie Mellon

Graph Patterns and Lockstep Behavior

Our intuition

- Lockstep behavior: Same Likes, same time



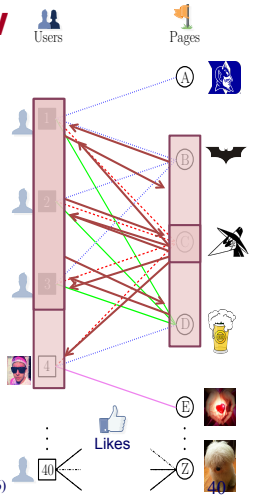
15-826 Copyright: C. Faloutsos (2016) 39

Suspicious Lockstep Behavior

Carnegie Mellon

MapReduce Overview

- Use Hadoop to search for many clusters in parallel:
  1. Start with randomly seed
  2. Update set of Pages and center Like times for each cluster
  3. Repeat until convergence



15-826 Copyright: C. Faloutsos (2016) 40

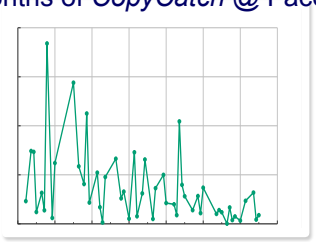
Carnegie Mellon

Deployment at Facebook

- *CopyCatch* runs regularly (along with many other security mechanisms, and a large Site Integrity team)

3 months of *CopyCatch* @ Facebook

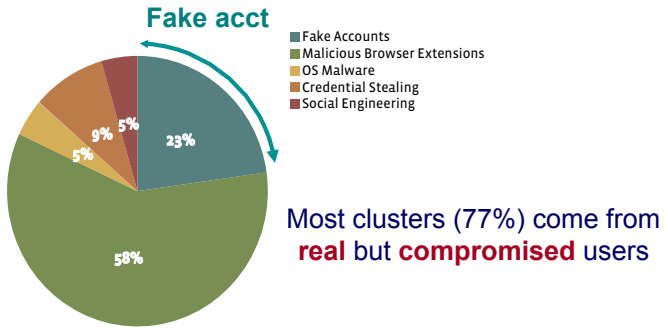
#users caught



time

Carnegie Mellon

Deployment at Facebook



Most clusters (77%) come from **real** but **compromised** users

Manually labeled 22 randomly selected *clusters* from February 2013

Carnegie Mellon

Roadmap

- Introduction – Motivation
- Unsupervised
  - static graphs
  - Time-evolving graphs
    - 'copyCatch' in FaceBook
    - Patterns of IAT
    - [tensors]
- Supervised: BP
- Conclusions

Carnegie Mellon

Universidade de São Paulo

筑波大学 University of Tsukuba

Carnegie Mellon University

KDD 2015 – Sydney, Australia

RSC: Mining and Modeling Temporal Activity in Social Media

Alceu F. Costa* Yuto Yamaguchi Agma J. M. Traina
Caetano Traina Jr. Christos Faloutsos

*alceufc@icmc.usp.br

Carnegie Mellon

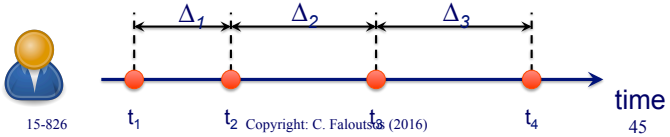
Pattern Mining: Datasets

Reddit Dataset	Twitter Dataset
Time-stamp from comments	Time-stamp from tweets
21,198 users	6,790 users
20 Million time-stamps	16 Million time-stamps

For each user we have:

Sequence of postings time-stamps: $\boldsymbol{T} = (t_1, t_2, t_3, \ldots)$

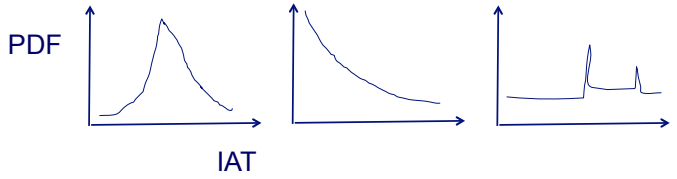
Inter-arrival times (IAT) of postings: $(\Delta_1, \Delta_2, \Delta_3, \ldots)$



Carnegie Mellon

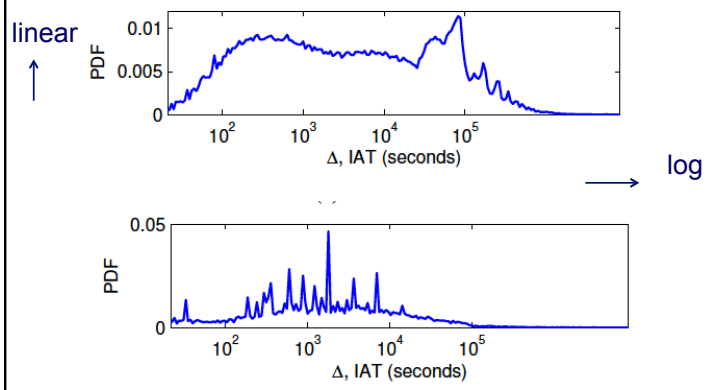
Guess the IAT pdf

- Gaussian (10' +/- a bit)?
- Power law? Slope?
- Log-logistic?
- Periodic (spike for 1day, 1 week)?



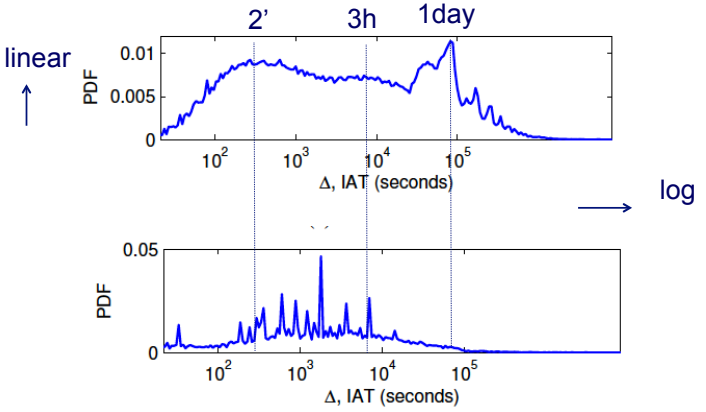
Carnegie Mellon

Human? Robots?



Carnegie Mellon

Human? Robots?



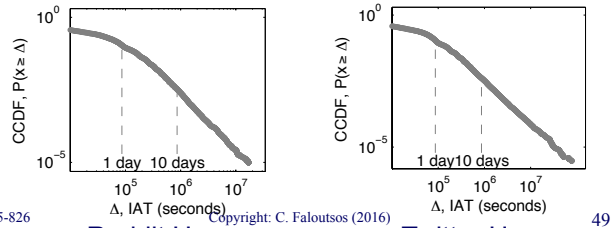
Carnegie Mellon

Pattern Mining

Pattern 1: Distribution of IAT is heavy-tailed

Users can be inactive for long periods of time before making new postings

IAT Complementary Cumulative Distribution Function (CCDF) (log-log axis)



Reddit Users Twitter Users

 Copyright: C. Faloutsos (2016)

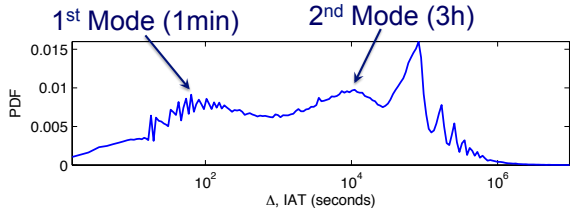
Carnegie Mellon

Pattern Mining

Pattern 2: Bimodal IAT distribution

Users have highly active sections and resting periods

Log-binned histogram of postings IAT



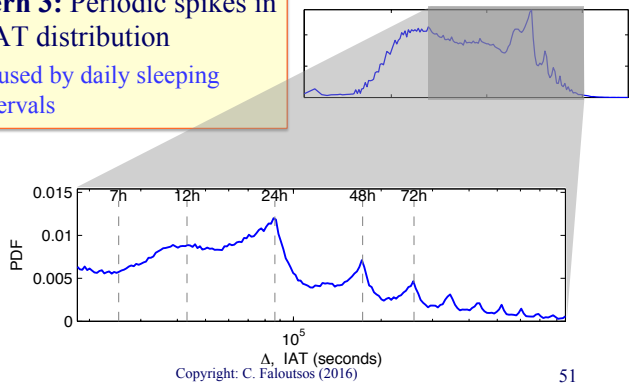
 Copyright: C. Faloutsos (2016)

Carnegie Mellon

Pattern Mining

Pattern 3: Periodic spikes in the IAT distribution

Caused by daily sleeping intervals



 Copyright: C. Faloutsos (2016)

Carnegie Mellon

Experiments: Can RSC-Spotter Detect Bots?

Methodology

Datasets

Users were manually labeled as bot or humans

Reddit		Twitter	
1,963 Humans	37 Bots	1353 Humans	64 Bots

Training

Same size for train and test subsets (preserved class distribution)

Baseline features:

1. IAT Histogram — Log-binned IAT histogram
2. Entropy — Entropy of the IAT histogram
3. Week Hist. — # of postings for day of week
4. All features — Combination of 1, 2 and 3

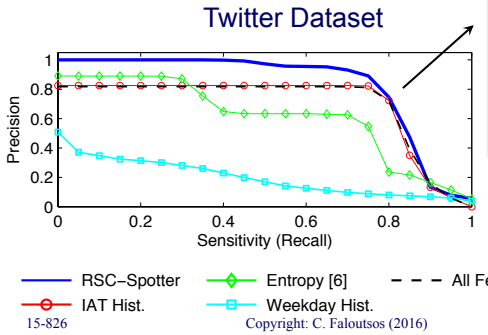
 Copyright: C. Faloutsos (2016)

Carnegie Mellon

Experiments: Can RSC-Spotter Detect Bots?

Precision vs. Sensitivity Curves

Good performance: curve close to the top



Precision > 94%
Sensitivity > 70%

With strongly imbalanced datasets

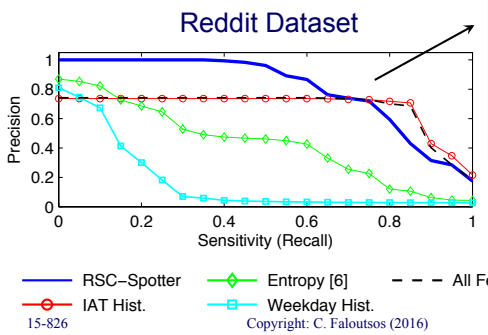
humans >> # bots

Carnegie Mellon

Experiments: Can RSC-Spotter Detect Bots?

Precision vs. Sensitivity Curves

Good performance: curve close to the top



Precision > 96%
Sensitivity > 47%

With strongly imbalanced datasets

humans >> # bots

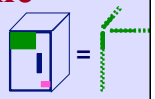
Carnegie Mellon

Roadmap

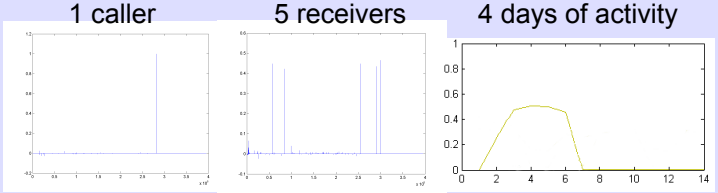
- Introduction – Motivation
- Unsupervised
  - static graphs
  - Time-evolving graphs
    - 'copyCatch' in FaceBook
    - Patterns of IAT
    - [tensors]
- Supervised: BP
- Conclusions

Carnegie Mellon

Anomaly detection in time-evolving graphs



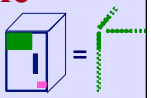
- Anomalous communities in phone call data:
  - European country, 4M clients, data over 2 weeks



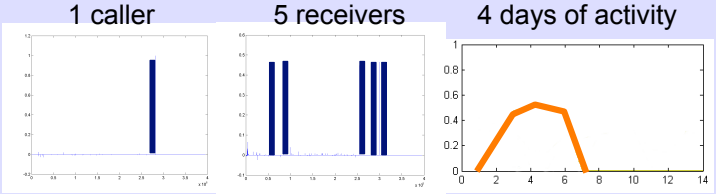
~200 calls to EACH receiver on EACH day!

Carnegie Mellon

Anomaly detection in time-evolving graphs



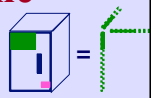
- Anomalous communities in phone call data:
  - European country, 4M clients, data over 2 weeks



~200 calls to EACH receiver on EACH day!

Carnegie Mellon

Anomaly detection in time-evolving graphs



- Anomalous communities in phone call data:
  - European country, 4M clients, data over 2 weeks

Miguel Araujo, Spiros Papadimitriou, Stephan Günnemann, Christos Faloutsos, Prithwish Basu, Ananthram Swami, Evangelos Papalexakis, Danai Koutra. *Com2: Fast Automatic Discovery of Temporal (Comet) Communities.* PAKDD 2014, Tainan, Taiwan.

Carnegie Mellon

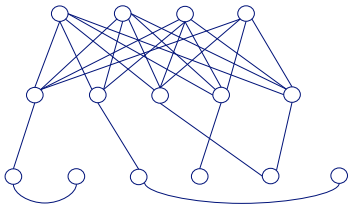
Roadmap

- Introduction – Motivation
- Unsupervised
  - static graphs
  - Time-evolving graphs
- Supervised: BP
  - Ebay fraud ('NetProbe')
  - Malware (Polonium)
- Conclusions

Carnegie Mellon

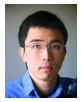
E-bay Fraud detection

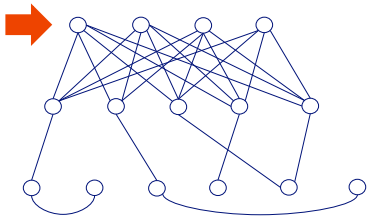
w/ Polo Chau & Shashank Pandit, CMU [www'07]



Carnegie Mellon

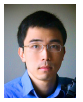
E-bay Fraud detection

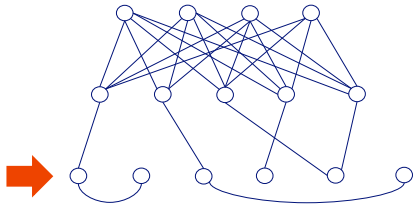




Carnegie Mellon

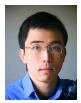
E-bay Fraud detection

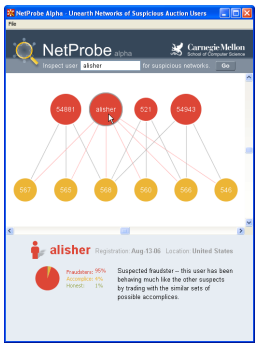


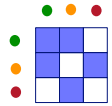


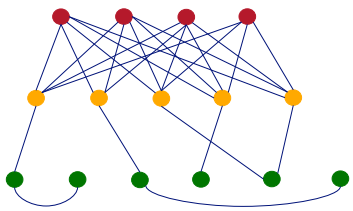
Carnegie Mellon

E-bay Fraud detection - NetProbe



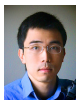






Carnegie Mellon

Popular press



USA TODAY

The Washington Post

Los Angeles Times

And less desirable attention:

- E-mail from 'Belgium police' ('copy of your code?')

Carnegie Mellon

Roadmap

- Introduction – Motivation
- Part#1: Patterns in graphs
  - Patterns
  - Anomaly / fraud detection
    - CopyCatch
    - Spectral methods ('fBox')
    - Belief Propagation; antivirus app
- Part#2: time-evolving graphs; tensors
- Conclusions

Carnegie Mellon

Roadmap

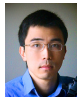
- Introduction – Motivation
- Unsupervised
  - static graphs
  - Time-evolving graphs
- Supervised: BP
  - Ebay fraud ('NetProbe')
  - Malware (Polonium)
- Conclusions

Carnegie Mellon

PATENT PENDING

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

SDM 2011, Mesa, Arizona



Polo Chau
Machine Learning Dept

symantec.

Carey Nachenberg
Vice President & Fellow

symantec.

Jeffrey Wilhelm
Principal Software Engineer

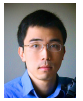
symantec.

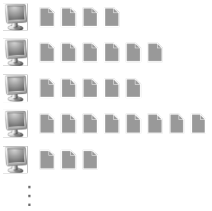
Adam Wright
Software Engineer

Prof. Christos Faloutsos
Computer Science Dept

Carnegie Mellon

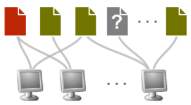
Polonium: The Data





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

- 50+ million machines
- 900+ million executable files

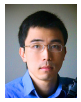


Constructed a machine-file bipartite graph (0.2 TB+)

- 1 billion nodes (machines and files)
- 37 billion edges

Carnegie Mellon

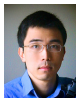
Polonium: Key Ideas

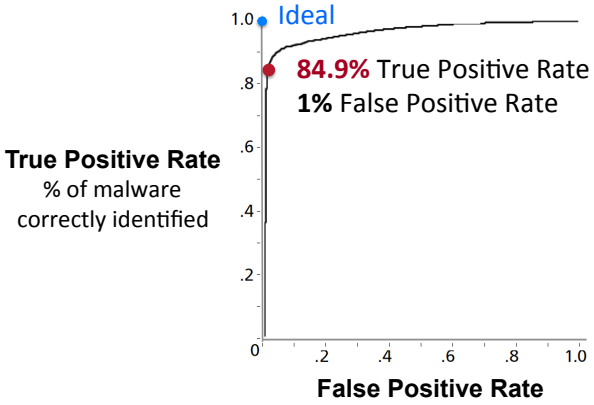


- Use Belief Propagation to propagate domain knowledge in machine-file graph to detect malware
- Use "guilt-by-association" (i.e., homophily)
  - E.g., files that appear on machines with many bad files are more likely to be bad
- Scalability: handles 37 billion-edge graph

Carnegie Mellon

Polonium: One-Interaction Results





Carnegie Mellon

Roadmap

- Introduction – Motivation
- Unsupervised
- Supervised: BP (Belief Propagation)
  - Ebay fraud ('NetProbe')
  - Malware (Polonium)
  - Intuition behind BP
- Conclusions

Carnegie Mellon

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

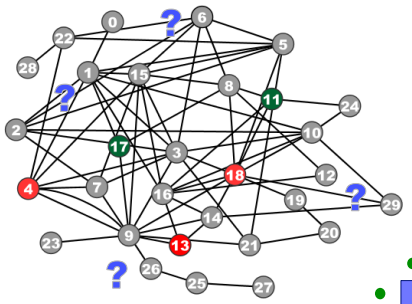
Danai Koutra
U Kang
Hsing-Kuo Kenneth Pao

Tai-You Ke
Duen Horng (Polo) Chau
Christos Faloutsos

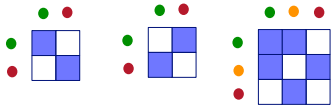
ECML PKDD, 5-9 September 2011, Athens, Greece

Carnegie Mellon

Problem Definition: GBA techniques



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



Carnegie Mellon

Are they related?

- RWR (Random Walk with Restarts)
  - google's pageRank ('*if my friends 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

Carnegie Mellon

Are they related? YES!

- RWR (Random Walk with Restarts)
  - google's pageRank ('*if my friends 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

Carnegie Mellon

Correspondence of Methods

Method	Matrix		Unknown		known
RWR	$[\mathbf{I} - c\ \underline{\mathbf{A}}\mathbf{D}^{-1}]$	×	$\mathbf{x}$	=	$(1\text{-}c)\mathbf{y}$
SSL	$[\mathbf{I} + a(\mathbf{D} - \underline{\mathbf{A}})]$	×	$\mathbf{x}$	=	$\mathbf{y}$
FABP	$[\mathbf{I} + a\ \mathbf{D} - c'\underline{\mathbf{A}}]$	×	$\mathbf{b}_h$	=	$\boldsymbol{\phi}_h$

adjacency matrix — final labels/ beliefs — prior labels/ beliefs

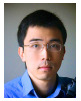
Carnegie Mellon

Cast

Akoglu, Leman

Araujo, Miguel

Beutel, Alex


Chau, Polo

Kang, U

Koutra, Danai

Lee, Jay Yoon

Prakash, Aditya

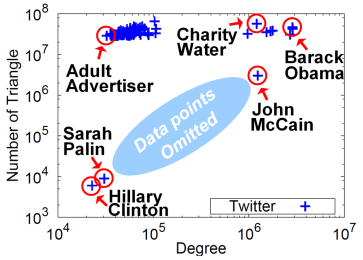
Papalexakis, Vagelis

Shah, Neil

Carnegie Mellon

CONCLUSION#1 – Big data

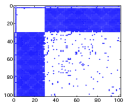
- **Patterns** **Anomalies**
- **Large** datasets reveal patterns/outliers that are invisible otherwise



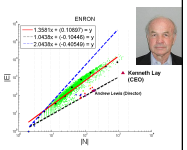
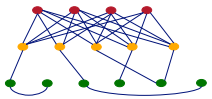
Carnegie Mellon

Conclusions, cont'd

- Spectral/tensor methods: spot lockstep behavior



- OddBall: strange, single nodes



- BP: good for the supervised setting (ie., some labels are given)