

Carnegie Mellon

15-826: Multimedia Databases and Data Mining

Lecture #26: Graph mining - patterns
Christos Faloutsos

Carnegie Mellon

Must-read Material

- Michalis Faloutsos, Petros Faloutsos and Christos Faloutsos, On Power-Law Relationships of the Internet Topology, SIGCOMM 1999.
- R. Albert, H. Jeong, and A.-L. Barabasi, Diameter of the World Wide Web Nature, 401, 130-131 (1999).
- Reka Albert and Albert-Laszlo Barabasi Statistical mechanics of complex networks, Reviews of Modern Physics, 74, 47 (2002).
- Jure Leskovec, Jon Kleinberg, Christos Faloutsos Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations, KDD 2005, Chicago, IL, USA

15-826 (c) 2012 C. Faloutsos 2

Carnegie Mellon

Must-read Material (cont'd)

- D. Chakrabarti and C. Faloutsos, Graph Mining: Laws, Generators and Algorithms, in ACM Computing Surveys, 38 (1), 2006
- J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos, Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication, in PKDD 2005, Porto, Portugal

15-826 (c) 2012 C. Faloutsos 3

Carnegie Mellon


Outline

- ➔ Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Tools
- Problem#3: Scalability
- Conclusions

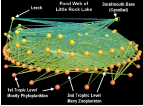
15-826 (c) 2012 C. Faloutsos 4

Carnegie Mellon


Graphs - why should we care?



Friendship Network
[Moody '01]



Food Web
[Martinez '91]



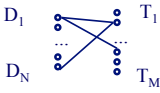
Internet Map
[lumeta.com]

15-826 (c) 2012 C. Faloutsos 5

Carnegie Mellon

Graphs - why should we care?

- IR: bi-partite graphs (doc-terms)
- web: hyper-text graph
- ... and more:



15-826 (c) 2012 C. Faloutsos 6

Carnegie Mellon

Graphs - why should we care?

- ‘viral’ marketing
- web-log (‘blog’) news propagation
- computer network security: email/IP traffic and anomaly detection
-

15-826 (c) 2012 C. Faloutsos 7

Carnegie Mellon

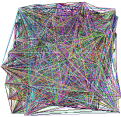
Outline

- Introduction – Motivation
- ➡ • Problem#1: Patterns in graphs
 - Static graphs
 - Weighted graphs
 - Time evolving graphs
- Problem#2: Tools
- Problem#3: Scalability
- Conclusions

15-826 (c) 2012 C. Faloutsos 8

Carnegie Mellon

Problem #1 - network and graph mining

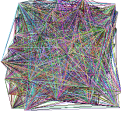


- What does the Internet look like?
- What does FaceBook look like?
- What is ‘normal’/‘abnormal’?
- which patterns/laws hold?


15-826 (c) 2012 C. Faloutsos 9

Carnegie Mellon

Problem #1 - network and graph mining



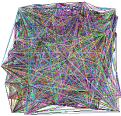
- What does the Internet look like?
- What does FaceBook look like?
- What is 'normal'/'abnormal'?
- which patterns/laws hold?
 - To spot **anomalies** (rarities), we have to discover **patterns**




15-826 (c) 2012 C. Faloutsos 10

Carnegie Mellon

Problem #1 - network and graph mining



- What does the Internet look like?
- What does FaceBook look like?
- What is 'normal'/'abnormal'?
- which patterns/laws hold?
 - To spot **anomalies** (rarities), we have to discover **patterns**
 - **Large** datasets reveal patterns/anomalies that may be invisible otherwise...



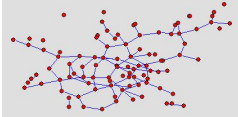
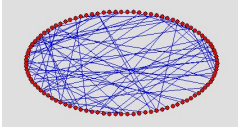
15-826 (c) 2012 C. Faloutsos 11

Carnegie Mellon

Are real graphs random?

- random (Erdos-Renyi) graph – 100 nodes, avg degree = 2
- before layout
- after layout
- No obvious patterns

(generated with: pajek
<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>)



15-826 (c) 2012 C. Faloutsos 12

Carnegie Mellon

Graph mining

- Are real graphs random?

15-826 (c) 2012 C. Faloutsos 13

Carnegie Mellon

Laws and patterns

- Are real graphs random?
- A: NO!!
 - Diameter
 - in- and out- degree distributions
 - other (surprising) patterns
- So, let's look at the data

15-826 (c) 2012 C. Faloutsos 14

Carnegie Mellon

Solution# S.1

- Power law in the degree distribution [SIGCOMM99]

internet domains

15-826 (c) 2012 C. Faloutsos 15

Carnegie Mellon

Solution# S.1

- Power law in the degree distribution [SIGCOMM99]

internet domains

att.com

ibm.com

-0.82

log(rank)

15-826 (c) 2012 C. Faloutsos 16

Carnegie Mellon

Solution# S.2: Eigen Exponent E

Exponent = slope

$E = -0.48$

May 2001

- A2: power law in the eigenvalues of the adjacency matrix

15-826 (c) 2012 C. Faloutsos 17

Carnegie Mellon

Solution# S.2: Eigen Exponent E

Exponent = slope

$E = -0.48$

May 2001

- [Mihail, Papadimitriou '02]: slope is $\frac{1}{2}$ of rank exponent

15-826 (c) 2012 C. Faloutsos 18

Carnegie Mellon

But:

How about graphs from other domains?

15-826 (c) 2012 C. Faloutsos 19

Carnegie Mellon

More power laws:

- web hit counts [w/ A. Montgomery]

Count (log scale)

Web Site Traffic

Zipf

ebay

users sites

in-degree (log scale)

15-826 (c) 2012 C. Faloutsos 20

Carnegie Mellon

epinions.com

- who-trusts-whom [Richardson + Domingos, KDD 2001]

count

Original graph

R-MAT graph

trusts-2000-people user

(out) degree

15-826 (c) 2012 C. Faloutsos 21

Carnegie Mellon

And numerous more

- # of sexual contacts
- Income [Pareto] – '80-20 distribution'
- Duration of downloads [Bestavros+]
- Duration of UNIX jobs ('mice and elephants')
- Size of files of a user
- ...
- 'Black swans'

15-826 (c) 2012 C. Faloutsos 22

Carnegie Mellon


Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
 - Static graphs
 - degree, diameter, eigen,
 - ➡ • triangles
 - cliques
 - Weighted graphs
 - Time evolving graphs
- Problem#2: Tools

15-826 (c) 2012 C. Faloutsos 23

Carnegie Mellon

Solution# S.3: Triangle 'Laws'



- Real social networks have a lot of triangles

15-826 (c) 2012 C. Faloutsos 24

Carnegie Mellon

Solution# S.3: Triangle 'Laws'



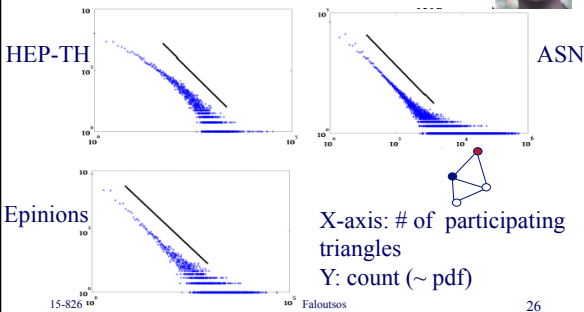

- Real social networks have a lot of triangles
 - Friends of friends are friends
- Any patterns?

15-826 (c) 2012 C. Faloutsos 25

Carnegie Mellon

Triangle Law: #S.3


[Tsourakakis ICDM 2008]



HEP-TH ASN

Epinions

X-axis: # of participating triangles
Y: count (~ pdf)

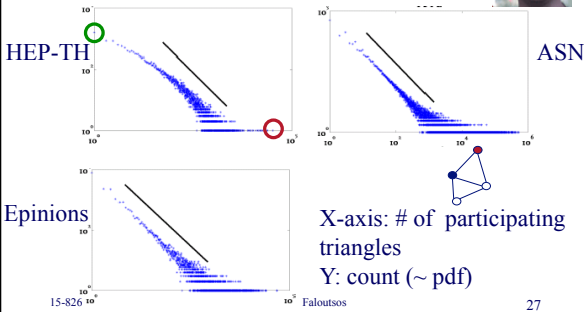



15-826 Faloutsos 26

Carnegie Mellon

Triangle Law: #S.3


[Tsourakakis ICDM 2008]



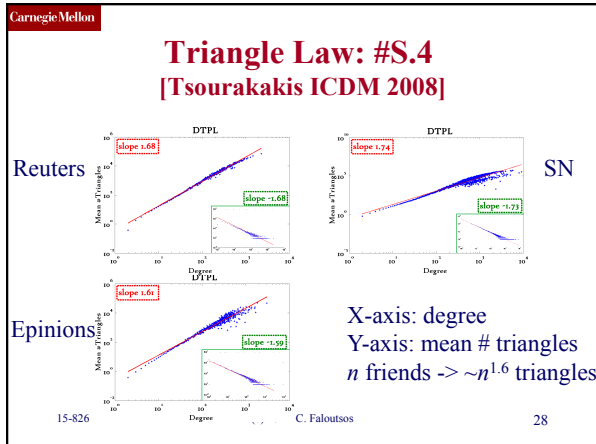
HEP-TH ASN

Epinions

X-axis: # of participating triangles
Y: count (~ pdf)



15-826 Faloutsos 27



Carnegie Mellon

Triangle Law: Computations [Tsourakakis ICDM 2008]

But: triangles are expensive to compute
(3-way join; several approx. algos)
Q: Can we do that quickly?

15-826 (c) 2012 C. Faloutsos 29



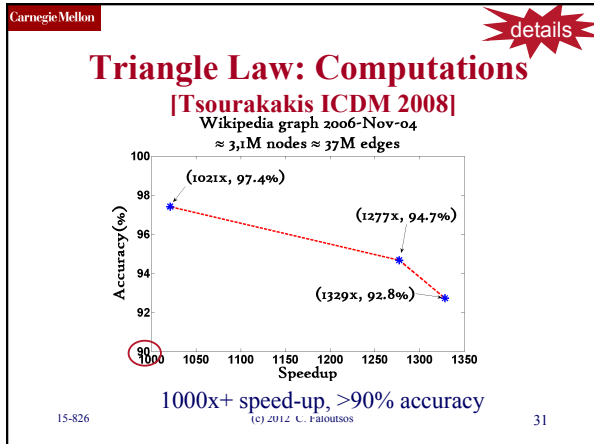
Carnegie Mellon

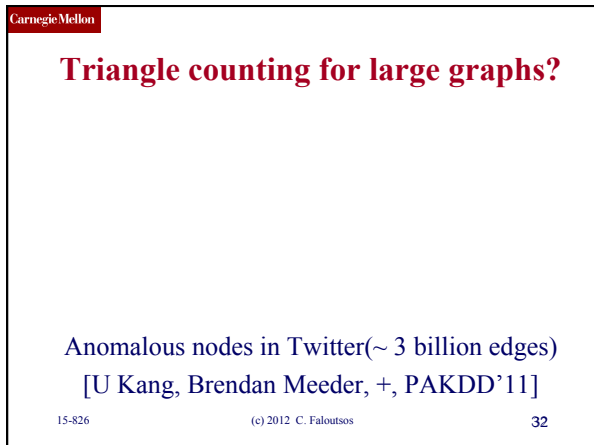
Triangle Law: Computations [Tsourakakis ICDM 2008]

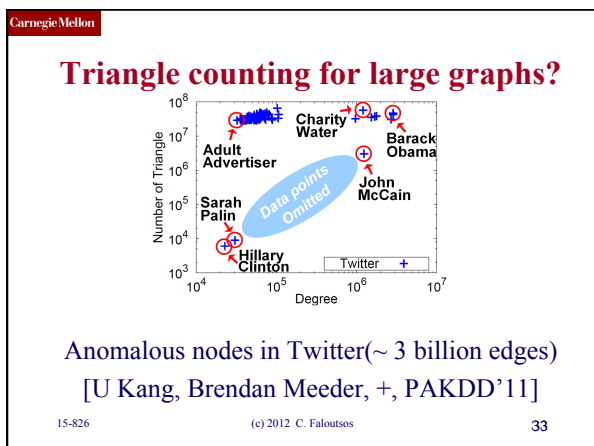
But: triangles are expensive to compute
(3-way join; several approx. algos)
Q: Can we do that quickly?
A: Yes!
#triangles = 1/6 Sum (λ_i^3)
(and, because of skewness (S2) ,
we only need the top few eigenvalues!

15-826 (c) 2012 C. Faloutsos 30









Carnegie Mellon

Any other 'laws'?

Yes!

15-826 (c) 2012 C. Faloutsos 34

Carnegie Mellon

Any other 'laws'?

Yes!

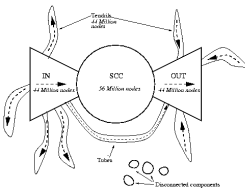
- Small diameter (~ constant!) –
 - six degrees of separation / 'Kevin Bacon'
 - small worlds [Watts and Strogatz]

15-826 (c) 2012 C. Faloutsos 35

Carnegie Mellon

Any other 'laws'?

- Bow-tie, for the web [Kumar+ '99]
- IN, SCC, OUT, 'tendrils'
- disconnected components



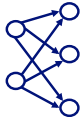
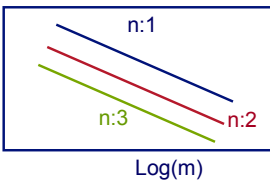
15-826 (c) 2012 C. Faloutsos 36

Carnegie Mellon

Any other 'laws'?

- power-laws in communities (bi-partite cores) [Kumar+, '99]

Log(count)



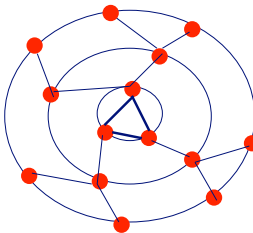
2:3 core
(m:n core)

15-826 (c) 2012 C. Faloutsos 37

Carnegie Mellon

Any other 'laws'?


- “Jellyfish” for Internet [Tauro+ '01]
- core: ~clique
- ~5 concentric layers
- many 1-degree nodes



15-826 (c) 2012 C. Faloutsos 38

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.

15-826 (c) 2012 C. Faloutsos 39

Carnegie Mellon

EigenSpokes

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

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


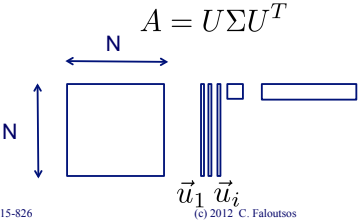
15-826 (c) 2012 C. Faloutsos 40

Carnegie Mellon

EigenSpokes

details

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

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


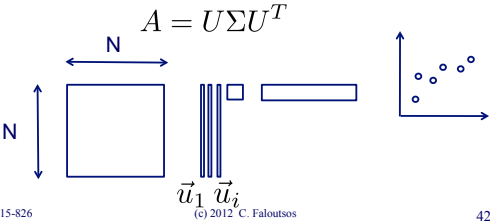
15-826 (c) 2012 C. Faloutsos 41

Carnegie Mellon

EigenSpokes

details

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

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


15-826 (c) 2012 C. Faloutsos 42

Carnegie Mellon details

EigenSpokes

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

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

15-826 (c) 2012 C. Faloutsos 43

Carnegie Mellon details

EigenSpokes

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

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

15-826 (c) 2012 C. Faloutsos 44

Carnegie Mellon

EigenSpokes

- EE plot:
 - Scatter plot of scores of u1 vs u2
 - One would expect
 - Many points @ origin
 - A few scattered ~randomly

$$u_2$$

u1

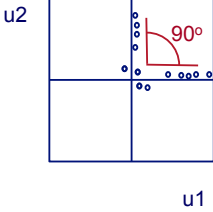
1st Principal component

15-826 (c) 2012 C. Faloutsos 45

Carnegie Mellon

EigenSpokes

- EE plot:
- Scatter plot of scores of u_1 vs u_2
- One would expect
 - Many points @ origin
 - A few scattered ~~~randomly~~



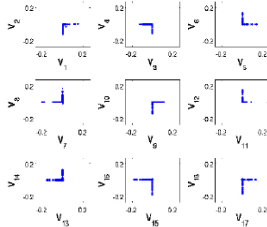
u_2 u_1

15-826 (c) 2012 C. Faloutsos 46

Carnegie Mellon

EigenSpokes - pervasiveness

- Present in mobile social graph
 - across time and space
- Patent citation graph

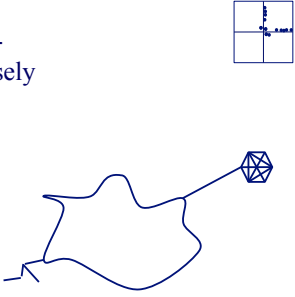


15-826 (c) 2012 C. Faloutsos 47

Carnegie Mellon

EigenSpokes - explanation

Near-cliques, or near-bipartite-cores, loosely connected


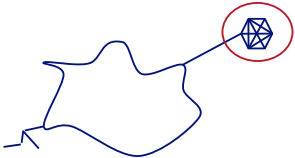


15-826 (c) 2012 C. Faloutsos 48

Carnegie Mellon

EigenSpokes - explanation

Near-cliques, or near-bipartite-cores, loosely connected


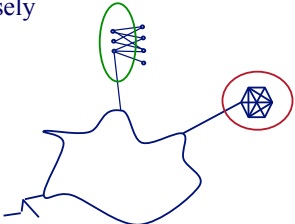


15-826 (c) 2012 C. Faloutsos 49

Carnegie Mellon

EigenSpokes - explanation

Near-cliques, or near-bipartite-cores, loosely connected



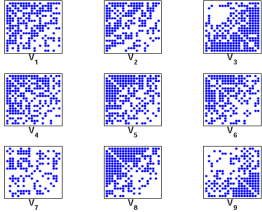
15-826 (c) 2012 C. Faloutsos 50

Carnegie Mellon

EigenSpokes - explanation

Near-cliques, or near-bipartite-cores, loosely connected

spy plot of top 20 nodes



So what?

- Extract nodes with high scores
- high connectivity
- Good “communities”

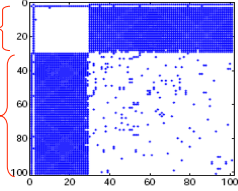
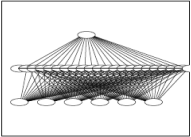
15-826 (c) 2012 C. Faloutsos 51

Carnegie Mellon

Bipartite Communities!

patents from same inventor(s)
'cut-and-paste' bibliography!

magnified bipartite community



15-826 (c) 2012 C. Faloutsos 52

Carnegie Mellon

Outline


- Introduction – Motivation
- Problem#1: Patterns in graphs
 - Static graphs
 - degree, diameter, eigen,
 - triangles
 - cliques
 - ➔ – Weighted graphs
 - Time evolving graphs
- Problem#2: Tools

15-826 (c) 2012 C. Faloutsos 53

Carnegie Mellon

Observations on weighted graphs?

- A: yes - even more 'laws'!



M. McGlohon, L. Akoglu, and C. Faloutsos
Weighted Graphs and Disconnected Components: Patterns and a Generator.
SIG-KDD 2008

15-826 (c) 2012 C. Faloutsos 54

Carnegie Mellon

Observation W.1: Fortification

Q: How do the weights of nodes relate to degree?

15-826 (c) 2012 C. Faloutsos 55

Carnegie Mellon

Observation W.1: Fortification

More donors, more \$?

15-826 (c) 2012 C. Faloutsos 56

Carnegie Mellon

Observation W.1: fortification: Snapshot Power Law

- Weight: super-linear on in-degree
- exponent 'iw': $1.01 < iw < 1.26$

More donors, even more \$

In-weights (\$)

e.g. John Kerry, \$10M received, from 1K donors

Edges (# donors)

15-826 (c) 2012 C. Faloutsos 57

Carnegie Mellon



Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
 - Static graphs
 - Weighted graphs
 - ➡ – Time evolving graphs
- Problem#2: Tools
- ...

15-826 (c) 2012 C. Faloutsos 58

Carnegie Mellon

Problem: Time evolution

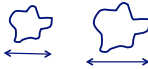
- with Jure Leskovec (CMU -> Stanford) 
- and Jon Kleinberg (Cornell – sabb. @ CMU) 

15-826 (c) 2012 C. Faloutsos 59

Carnegie Mellon

T.1 Evolution of the Diameter

- Prior work on Power Law graphs hints at **slowly growing diameter**:
 - diameter $\sim O(\log N)$
 - diameter $\sim O(\log \log N)$
- What is happening in real data?




15-826 (c) 2012 C. Faloutsos 60

Carnegie Mellon

T.1 Evolution of the Diameter

- Prior work on Power Law graphs hints at **slowly growing diameter**:
 - diameter $\sim O(\log N)$
 - diameter $\sim O(\log \log N)$
- What is happening in real data?
- Diameter **shrinks** over time

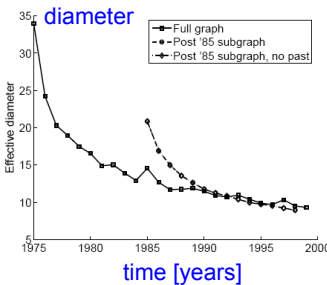


15-826 (c) 2012 C. Faloutsos 61

Carnegie Mellon

T.1 Diameter – “Patents”

- Patent citation network
- 25 years of data
- @1999
 - 2.9 M nodes
 - 16.5 M edges



15-826 (c) 2012 C. Faloutsos 62

Carnegie Mellon

T.2 Temporal Evolution of the Graphs

- $N(t)$... nodes at time t
- $E(t)$... edges at time t
- Suppose that
 - $N(t+1) = 2 * N(t)$
- Q: what is your guess for
 - $E(t+1) = ? 2 * E(t)$

15-826 (c) 2012 C. Faloutsos 63

Carnegie Mellon

T.2 Temporal Evolution of the Graphs

- $N(t)$... nodes at time t
- $E(t)$... edges at time t
- Suppose that
 - $N(t+1) = 2 * N(t)$
- Q: what is your guess for
 - $E(t+1) = ? * E(t)$
- A: over-doubled!
 - But obeying the “Densification Power Law”

15-826 (c) 2012 C. Faloutsos 64

Carnegie Mellon

T.2 Densification – Patent Citations

- Citations among patents granted
- @1999
 - 2.9 M nodes
 - 16.5 M edges
- Each year is a datapoint

15-826 (c) 2012 C. Faloutsos 65

Carnegie Mellon

Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
 - Static graphs
 - Weighted graphs
 - ➔ – Time evolving graphs
- Problem#2: Tools
- ...

15-826 (c) 2012 C. Faloutsos 66

Carnegie Mellon

More on Time-evolving graphs

M. McGlohon, L. Akoglu, and C. Faloutsos
Weighted Graphs and Disconnected Components: Patterns and a Generator.
SIG-KDD 2008

15-826 (c) 2012 C. Faloutsos 67

Carnegie Mellon

[Gelling Point]

- Most real graphs display a gelling point
- After gelling point, they exhibit typical behavior. This is marked by a spike in diameter.

15-826 (c) 2012 C. Faloutsos 68

Carnegie Mellon

Observation T.3: NLCC behavior

Q: How do NLCC's emerge and join with the GCC?

(`NLCC` = non-largest conn. components)

- Do they continue to grow in size?
- or do they shrink?
- or stabilize?

15-826 (c) 2012 C. Faloutsos 69


Carnegie Mellon

Observation T.3: NLCC behavior

Q: How do NLCC's emerge and join with the GCC?

(`NLCC` = non-largest conn. components)

- Do they continue to grow in size?
- or do they shrink?
- or stabilize?



15-826 (c) 2012 C. Faloutsos 70

Carnegie Mellon

Observation T.3: NLCC behavior

Q: How do NLCC's emerge and join with the GCC?

(`NLCC` = non-largest conn. components)

YES - Do they continue to grow in size?

YES - or do they shrink?

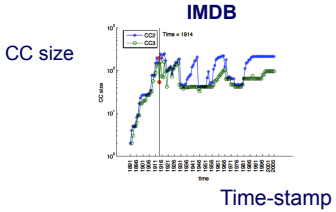
YES - or stabilize?

15-826 (c) 2012 C. Faloutsos 71

Carnegie Mellon

Observation T.3: NLCC behavior

- After the gelling point, the GCC takes off, but NLCC's remain ~constant (actually, **oscillate**).



15-826 (c) 2012 C. Faloutsos 72

Carnegie Mellon

Timing for Blogs

- with Mary McGlohon (CMU->Google)
- Jure Leskovec (CMU->Stanford)
- Natalie Glance (now at Google)
- Mat Hurst (now at MSR)

[SDM'07]

15-826 (c) 2012 C. Faloutsos 73

Carnegie Mellon

T.4 : popularity over time

in links

lag: days after post

Post popularity drops-off – exponentially?

@t + lag @t

15-826 (c) 2012 C. Faloutsos 74

Carnegie Mellon

T.4 : popularity over time

in links (log)

days after post (log)

Post popularity drops-off – exponentially? **POWER LAW!**
Exponent?

15-826 (c) 2012 C. Faloutsos 75

Carnegie Mellon

T.4 : popularity over time

in links (log)

days after post (log)

Post popularity drops-off – exponentially? **NO**
POWER LAW!
Exponent? -1.6

- close to -1.5: Barabasi's stack model
- and like the zero-crossings of a random walk

15-826 (c) 2012 C. Faloutsos 76

Carnegie Mellon

-1.5 slope

J. G. Oliveira & A.-L. Barabási Human Dynamics: The Correspondence Patterns of Darwin and Einstein.
Nature **437**, 1251 (2005). [[PDF](#)]

15-826 (c) 2012 C. Faloutsos 77

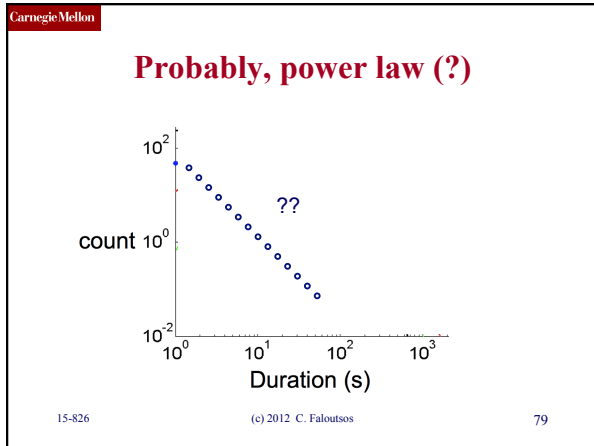
Carnegie Mellon

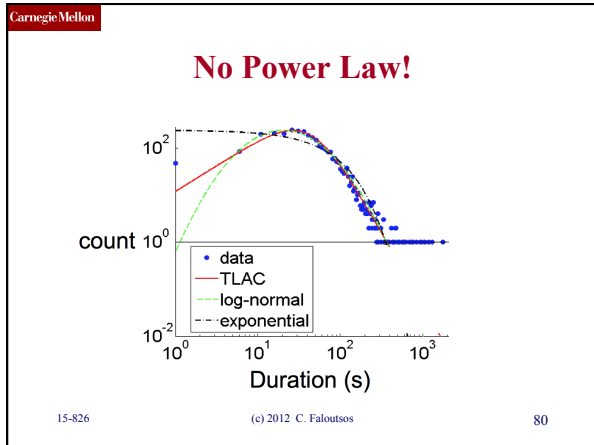
T.5: duration of phonecalls

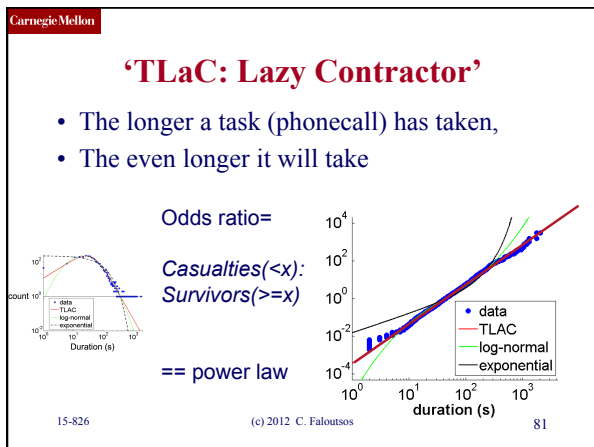
Surprising Patterns for the Call Duration Distribution of Mobile Phone Users

Pedro O. S. Vaz de Melo, Leman Akoglu, Christos Faloutsos, Antonio A. F. Loureiro
PKDD 2010

15-826 (c) 2012 C. Faloutsos 78





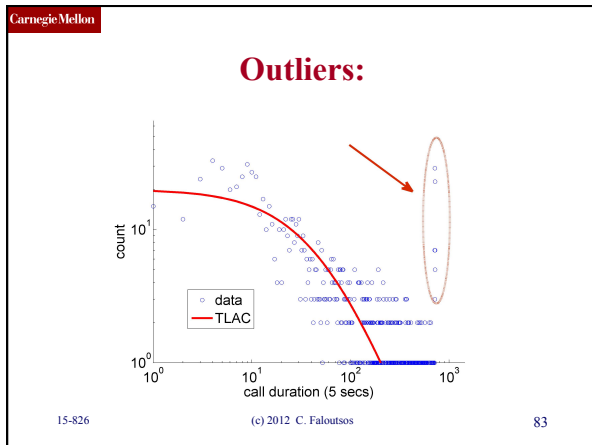


Carnegie Mellon

Data Description

- Data from a private mobile operator of a large city
 - 4 months of data
 - 3.1 million users
 - more than 1 billion phone records
- Over 96% of ‘talkative’ users obeyed a TLAC distribution (‘talkative’: >30 calls)

15-826 (c) 2012 C. Faloutsos 82



Carnegie Mellon


Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Tools
 - ➡ – OddBall (anomaly detection)
 - Belief Propagation
 - Immunization
- Problem#3: Scalability
- Conclusions

15-826 (c) 2012 C. Faloutsos 84

Carnegie Mellon

OddBall: Spotting Anomalies in Weighted Graphs



Lemnan Akoglu, Mary McGlohon, Christos Faloutsos
Carnegie Mellon University
School of Computer Science

PAKDD 2010, Hyderabad, India

Carnegie Mellon

Main idea

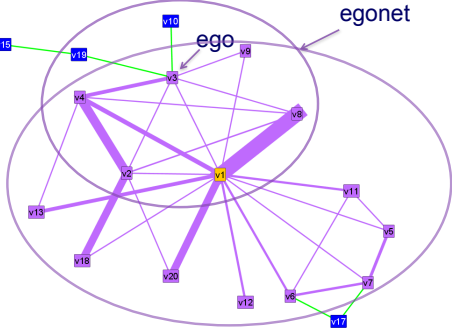
For each node,

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

15-826 (c) 2012 C. Faloutsos 86

Carnegie Mellon

What is an egonet?

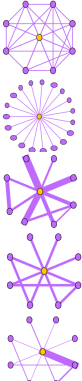


15-826 (c) 2012 C. Faloutsos 87

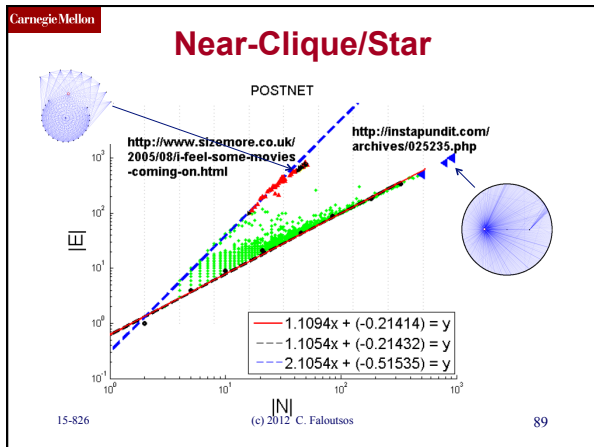
Carnegie Mellon

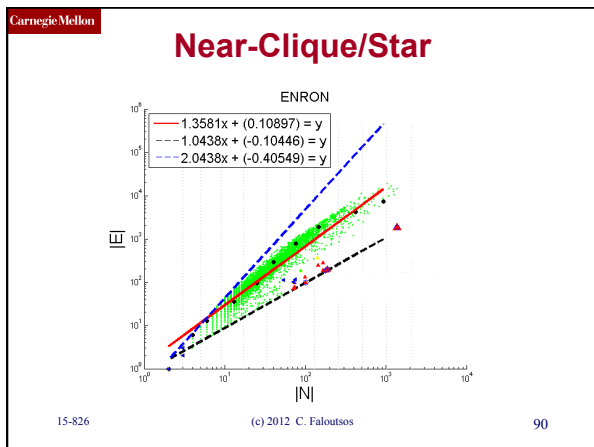
Selected Features

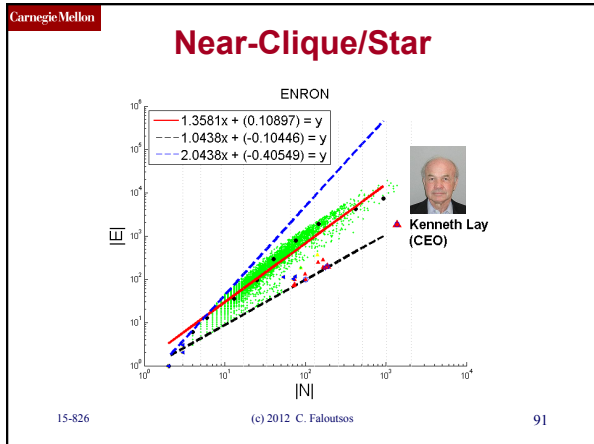
- N_i : number of neighbors (degree) of ego i
- E_i : number of edges in egonet i
- W_i : total weight of egonet i
- $\lambda_{w,i}$: principal eigenvalue of the **weighted** adjacency matrix of egonet I

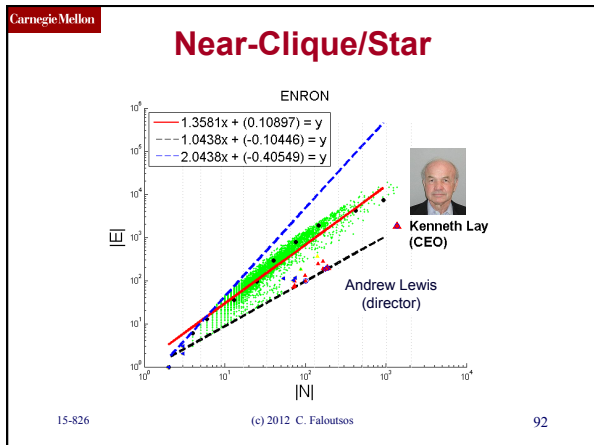


15-826 (c) 2012 C. Faloutsos










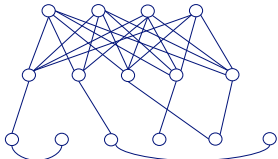
- Carnegie Mellon
- ### Outline
- Introduction – Motivation
 - Problem#1: Patterns in graphs
 - Problem#2: Tools
 - OddBall (anomaly detection)
 - ➡ – Belief Propagation
 - Immunization
 - Problem#3: Scalability
 - Conclusions
- 15-826 (c) 2012 C. Faloutsos 93

Carnegie Mellon

E-bay Fraud detection



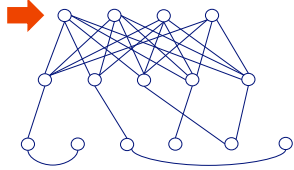
w/ Polo Chau & Shashank Pandit, CMU
[www'07]



15-826 (c) 2012 C. Faloutsos 94

Carnegie Mellon

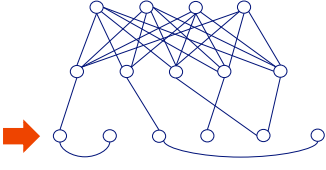
E-bay Fraud detection



15-826 (c) 2012 C. Faloutsos 95

Carnegie Mellon

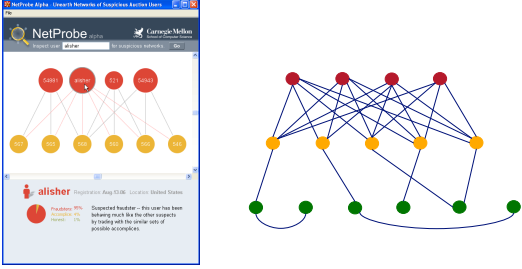
E-bay Fraud detection



15-826 (c) 2012 C. Faloutsos 96

Carnegie Mellon

E-bay Fraud detection - NetProbe



The screenshot shows the NetProbe application interface with a search bar and a network graph. The graph has four red nodes at the top, four yellow nodes in the middle, and four green nodes at the bottom. The network graph diagram to the right shows a similar structure with red, yellow, and green nodes connected by blue lines.

15-826 (c) 2012 C. Faloutsos 97

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?')

15-826 (c) 2012 C. Faloutsos 98

Carnegie Mellon

Outline

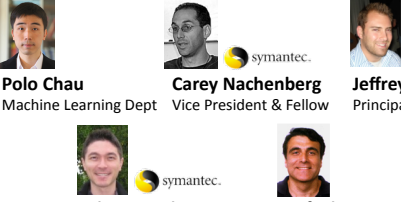
- Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Tools
 - OddBall (anomaly detection)
 - ➡ – Belief Propagation – antivirus app
 - Immunization
- Problem#3: Scalability
- Conclusions

15-826 (c) 2012 C. Faloutsos 99

Carnegie Mellon

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

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

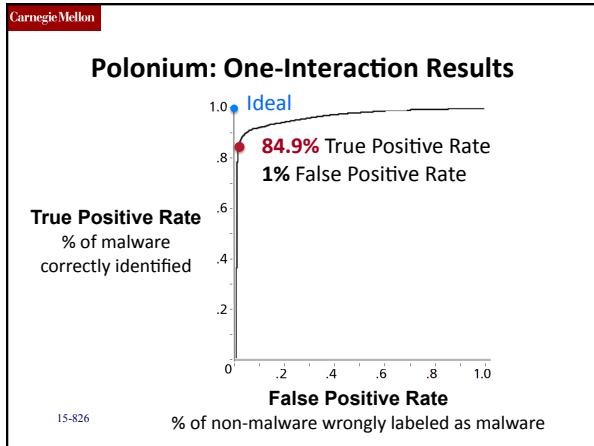
15-826 (c) 2012 C. Faloutsos 101

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

15-826 (c) 2012 C. Faloutsos 102



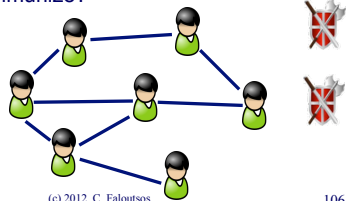
- Carnegie Mellon
- ### Outline
- Introduction – Motivation
 - Problem#1: Patterns in graphs
 - Problem#2: Tools
 - OddBall (anomaly detection)
 - Belief propagation
 - ➡ – Immunization
 - Problem#3: Scalability -PEGASUS
 - Conclusions
- 15-826 (c) 2012 C. Faloutsos 104

- Carnegie Mellon
- ### Immunization and epidemic thresholds
- Q1: which nodes to immunize?
 - Q2: will a virus vanish, or will it create an epidemic?
- 15-826 (c) 2012 C. Faloutsos 105

Carnegie Mellon

Q1: Immunization:

- Given
 - a network,
 - k vaccines, and
 - the virus details
- Which nodes to immunize?

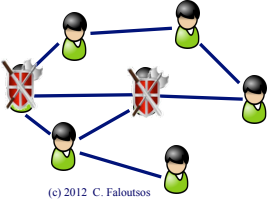


15-826 (c) 2012 C. Faloutsos 106

Carnegie Mellon

Q1: Immunization:

- Given
 - a network,
 - k vaccines, and
 - the virus details
- Which nodes to immunize?

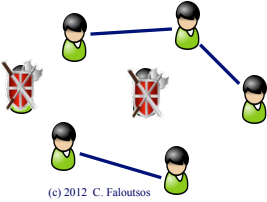


15-826 (c) 2012 C. Faloutsos 107

Carnegie Mellon

Q1: Immunization:

- Given
 - a network,
 - k vaccines, and
 - the virus details
- Which nodes to immunize?



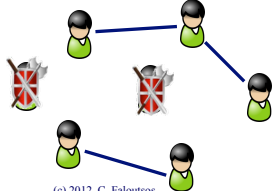
15-826 (c) 2012 C. Faloutsos 108

Carnegie Mellon

Q1: Immunization:

- Given
 - a network,
 - k vaccines, and
 - the virus details
- Which nodes to immunize?

A: immunize the ones that maximally raise the 'epidemic threshold' [Tong+, ICDM'10]



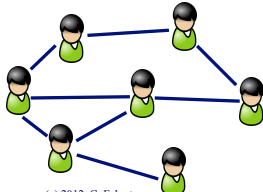
15-826 (c) 2012 C. Faloutsos 109

Carnegie Mellon

Q2: will a virus take over?

- Flu-like virus (no immunity, 'SIS')
- Mumps (life-time immunity, 'SIR')
- Pertussis (finite-length immunity, 'SIRS')

β : attack prob
 δ : heal prob



15-826 (c) 2012 C. Faloutsos 110

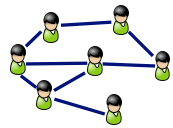
Carnegie Mellon

Q2: will a virus take over?

- Flu-like virus (no immunity, 'SIS')
- Mumps (life-time immunity, 'SIR')
- Pertussis (finite-length immunity, 'SIRS')

β : attack prob
 δ : heal prob

A: depends on connectivity (avg degree? Max degree? variance? Something else?)



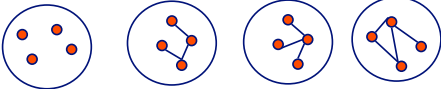
15-826 (c) 2012 C. Faloutsos 111

Carnegie Mellon details

Epidemic threshold τ

What should τ depend on?

- avg. degree? and/or highest degree?
- and/or variance of degree?
- and/or third moment of degree?
- and/or diameter?



15-826 (c) 2012 C. Faloutsos 112

Carnegie Mellon details

Epidemic threshold

- [Theorem] We have no epidemic, if

$$\beta/\delta < \tau = 1/\lambda_{1,A}$$

15-826 (c) 2012 C. Faloutsos 113

Carnegie Mellon details

Epidemic threshold

- [Theorem] We have no epidemic, if

recovery prob. δ

epidemic threshold τ

$$\beta/\delta < \tau = 1/\lambda_{1,A}$$

attack prob. β

largest eigenvalue of adj. matrix A

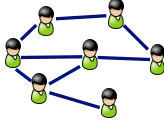
Proof: [Wang+03] (for SIS=flu only)

15-826 (c) 2012 C. Faloutsos 114


Carnegie Mellon details

A2: will a virus take over?

- For **all** typical virus propagation models (flu, mumps, pertussis, HIV, etc)
- The **only** connectivity measure that matters, is $1/\lambda_1$ the first eigenvalue of the adj. matrix [Prakash+, '10, arxiv]

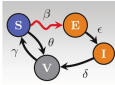


15-826 (c) 2012 C. Faloutsos 115

Carnegie Mellon 

Thresholds for some models

- $s = \text{effective strength}$
- $s < 1$: below threshold



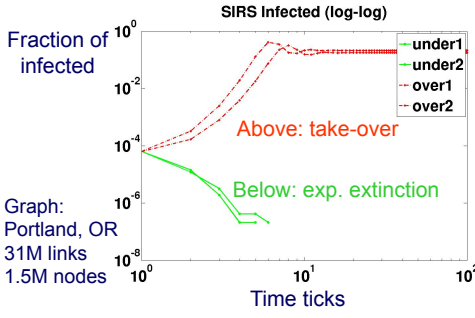
Models	Effective Strength (s)	Threshold (tipping point)
SIS, SIR, SIRS, SEIR	$s = \lambda \cdot \left(\frac{\beta}{\delta} \right)$	$s = 1$
SIV, SEIV	$s = \lambda \cdot \left(\frac{\beta\gamma}{\delta(\gamma + \theta)} \right)$	
SI ₁ I ₂ V ₁ V ₂ (H.I.I.V.)	$s = \lambda \cdot \left(\frac{\beta_1 v_2 + \beta_2 \epsilon}{v_2(\epsilon + v_1)} \right)$	

15-826 (c) 2012 C. Faloutsos 117

Carnegie Mellon

A2: will a virus take over?

SIRS Infected (log-log)



Graph: Portland, OR
31M links
1.5M nodes

15-826 (c) 2012 C. Faloutsos 117

Carnegie Mellon

Q1: Immunization:

- Given
 - a network,
 - k vaccines, and
 - the virus details
- Which nodes to immunize?

A: immunize the ones that maximally raise the 'epidemic threshold' [Tong+, ICDM'10]

15-826 (c) 2012 C. Faloutsos 118

Carnegie Mellon

Q1: Immunization:

- Given
 - a network,
 - k vaccines, and
 - the virus details
- Which nodes to immunize?

A: immunize the ones that

Max eigen-drop $\Delta\lambda$ for any virus!

15-826 (c) 2012 C. Faloutsos 119

Carnegie Mellon


Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Tools
 - OddBall (anomaly detection)
 - Belief propagation
 - Immunization
- ➡ Problem#3: Scalability -PEGASUS
- Conclusions


15-826 (c) 2012 C. Faloutsos 120

Carnegie Mellon

Scalability



- Google: > 450,000 processors in clusters of ~2000 processors each [Barroso, Dean, Hölzle, "Web Search for a Planet: The Google Cluster Architecture" IEEE Micro 2003]
- Yahoo: 5Pb of data [Fayyad, KDD'07]
- Problem: machine failures, on a daily basis
- How to parallelize data mining tasks, then?
- A: map/reduce – hadoop (open-source clone) <http://hadoop.apache.org/>



15-826 (c) 2012 C. Faloutsos 121

Carnegie Mellon


Outline – Algorithms & results

	Centralized	Hadoop/PEGASUS
Degree Distr.	old	old
Pagerank	old	old
→ Diameter/ANF	old	HERE
Conn. Comp	old	HERE
Triangles	done	HERE
Visualization	started	

15-826 (c) 2012 C. Faloutsos 122

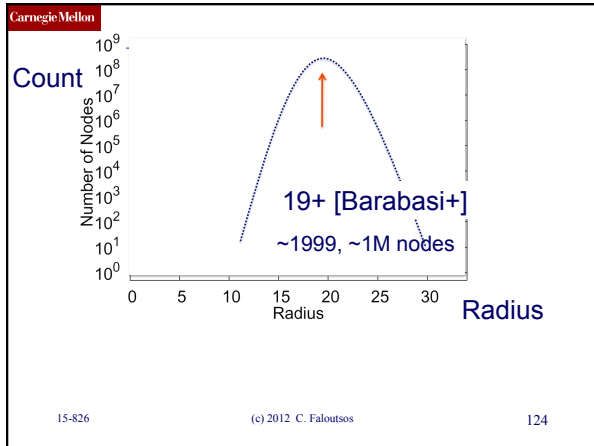
Carnegie Mellon

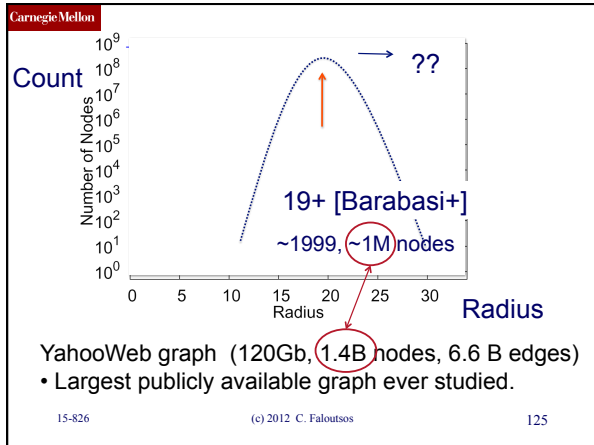
HADI for diameter estimation

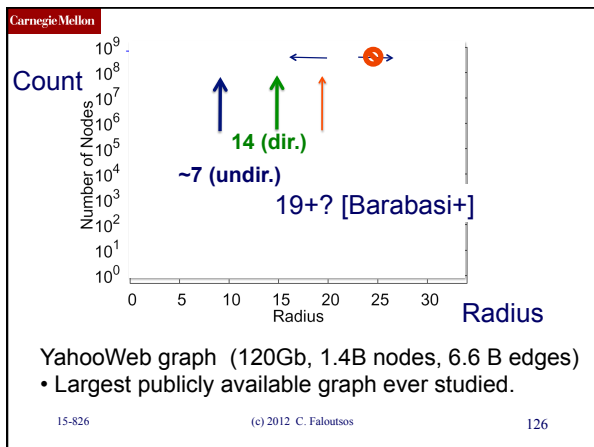


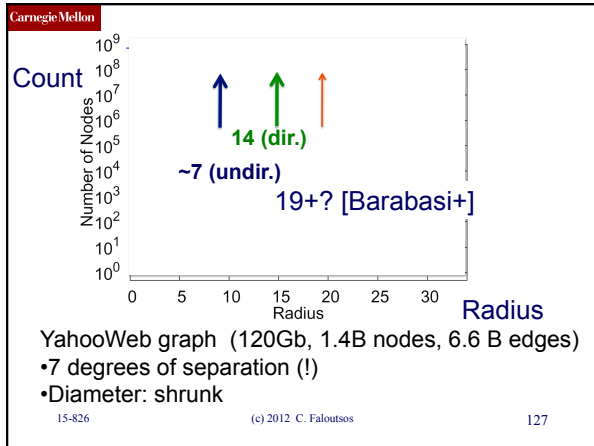
- *Radius Plots for Mining Tera-byte Scale Graphs* U Kang, Charalampos Tsourakakis, Ana Paula Appel, Christos Faloutsos, Jure Leskovec, SDM'10
- Naively: diameter needs $O(N^2)$ space and up to $O(N^3)$ time – **prohibitive** ($N \sim 1B$)
- Our HADI: linear on E ($\sim 10B$)
 - Near-linear scalability wrt # machines
 - Several optimizations \rightarrow 5x faster

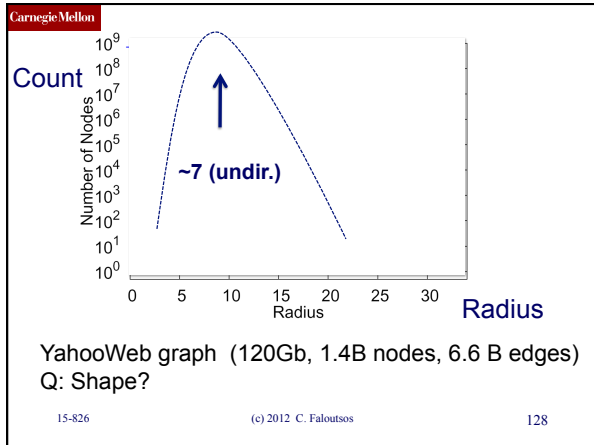
15-826 (c) 2012 C. Faloutsos 123

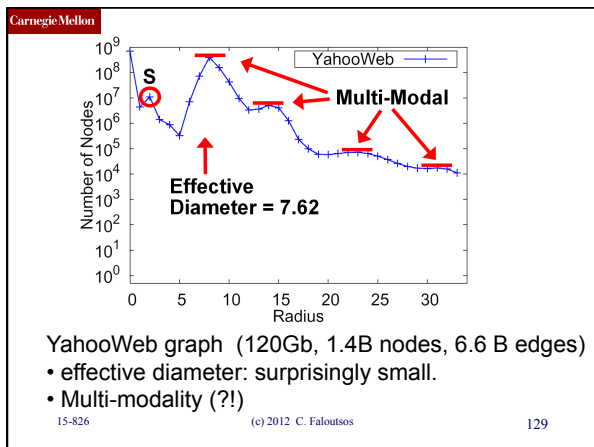


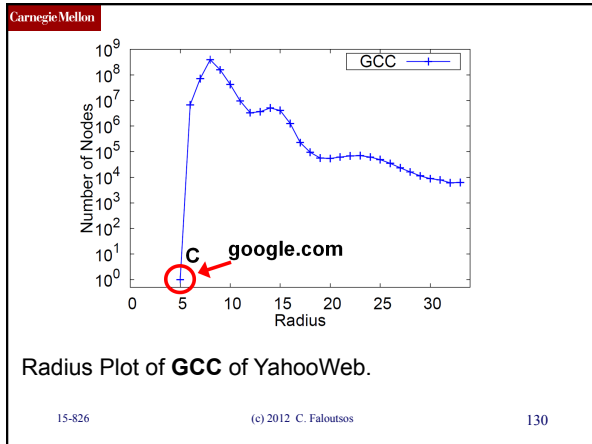


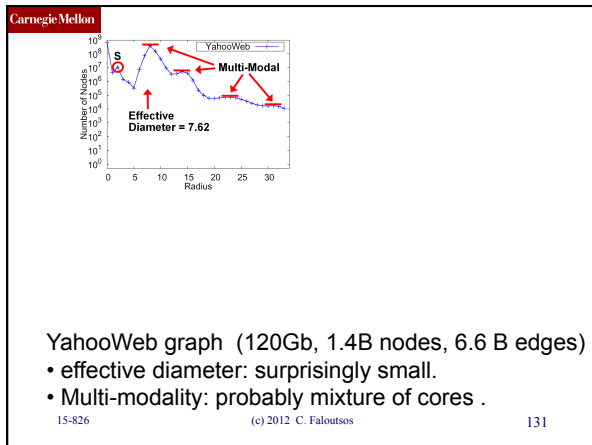


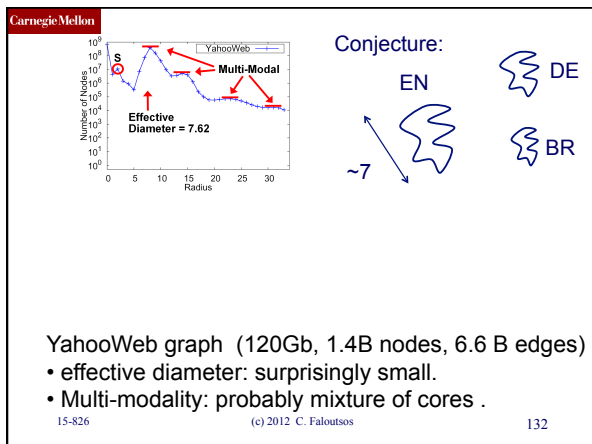


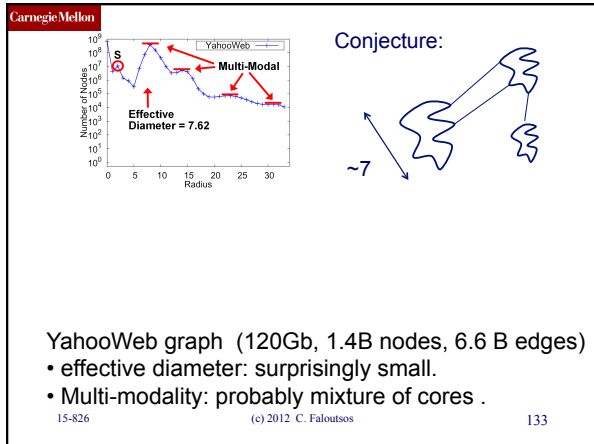


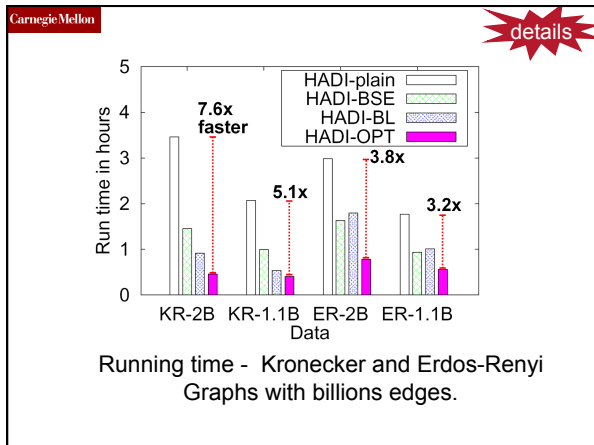












Outline – Algorithms & results

	Centralized	Hadoop/PEGASUS
Degree Distr.	old	old
Pagerank	old	old
Diameter/ANF	old	HERE
→ Conn. Comp	old	HERE
Triangles		HERE
Visualization	started	

15-826 (c) 2012 C. Faloutsos 135

Carnegie Mellon

Generalized Iterated Matrix Vector Multiplication (GIMV)

PEGASUS: A Peta-Scale Graph Mining System - Implementation and Observations.
U Kang, Charalampos E. Tsourakakis, and Christos Faloutsos.
(ICDM) 2009, Miami, Florida, USA.
Best Application Paper (runner-up).

15-826 (c) 2012 C. Faloutsos 136

Carnegie Mellon

Generalized Iterated Matrix Vector Multiplication (GIMV) details

- PageRank
- proximity (RWR)
- Diameter
- Connected components
- (eigenvectors,
- Belief Prop.
- ...)

Matrix – vector Multiplication (iterated)

15-826 (c) 2012 C. Faloutsos 137

Carnegie Mellon

Example: GIM-V At Work

- Connected Components – 4 observations:

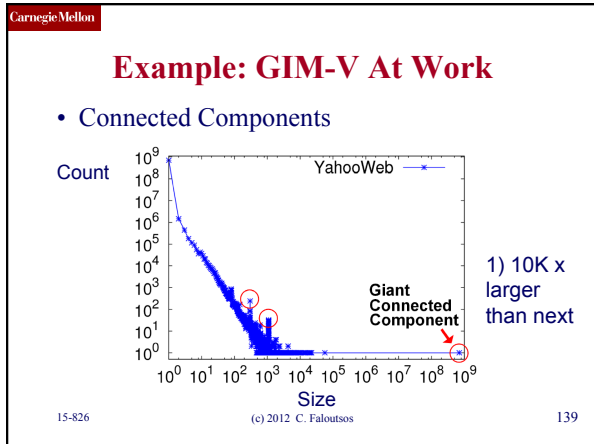
Count

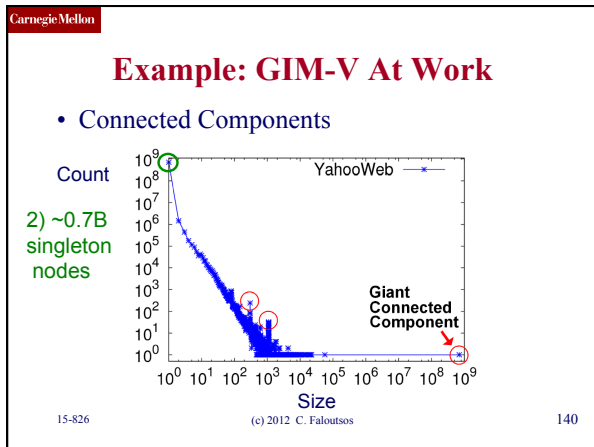
YahooWeb

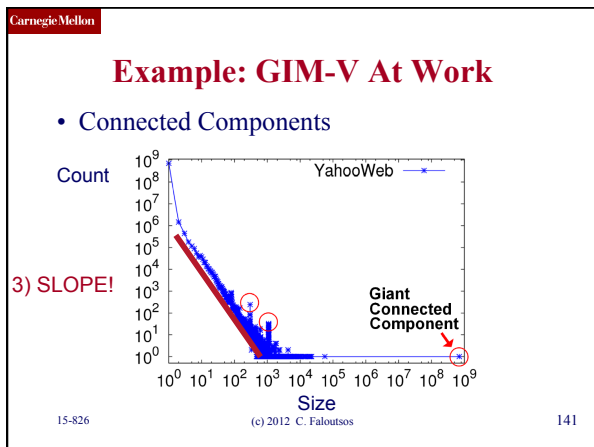
Size

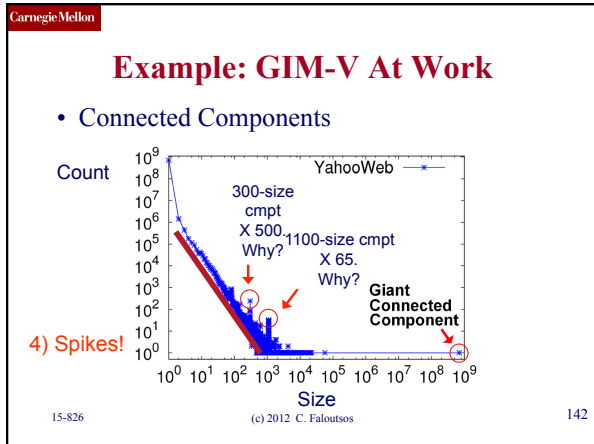
Giant Connected Component

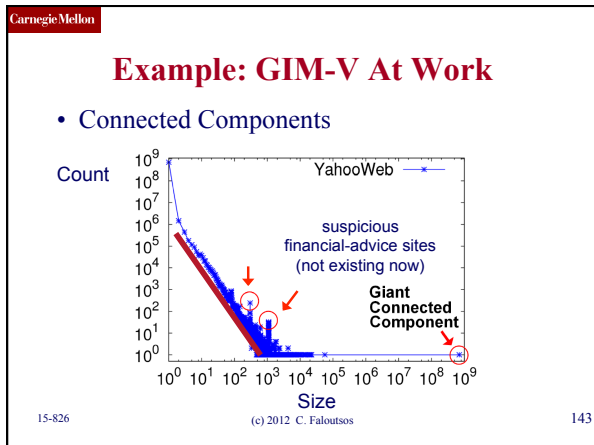
15-826 (c) 2012 C. Faloutsos 138

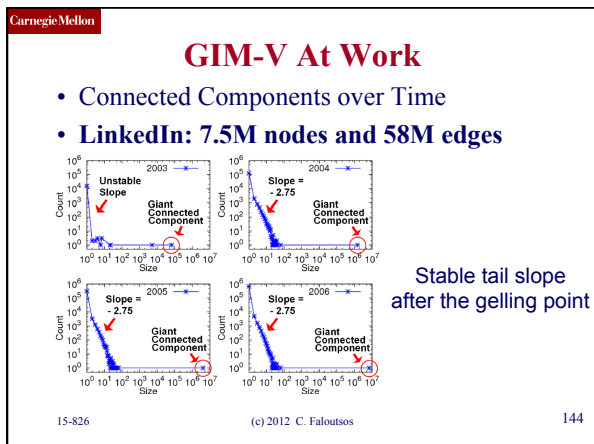












Carnegie Mellon

Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Tools
- Problem#3: Scalability
- ➔ • Conclusions

15-826 (c) 2012 C. Faloutsos 145

Carnegie Mellon

OVERALL CONCLUSIONS – low level:

- Several new **patterns** (fortification, shrinking diameter, triangle-laws, conn. components, etc)
- New **tools**:
 - anomaly detection (OddBall), belief propagation, immunization
- **Scalability**: PEGASUS / hadoop

15-826 (c) 2012 C. Faloutsos 146

Carnegie Mellon

OVERALL CONCLUSIONS – high level

- **BIG DATA**: Large datasets reveal patterns/ outliers that are invisible otherwise

Number of Nodes

Radius

Effective Diameter = 7.62

Multi-Modal

YahooWeb

15-826 (c) 2012 C. Faloutsos 147

Carnegie Mellon

References

- Leman Akoglu, Christos Faloutsos: *RTG: A Recursive Realistic Graph Generator Using Random Typing*. ECML/PKDD (1) 2009: 13-28
- Deepayan Chakrabarti, Christos Faloutsos: *Graph mining: Laws, generators, and algorithms*. ACM Comput. Surv. 38(1): (2006)

15-826 (c) 2012 C. Faloutsos 148

Carnegie Mellon

References

- Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jure Leskovec, Christos Faloutsos: *Epidemic thresholds in real networks*. ACM Trans. Inf. Syst. Secur. 10(4): (2008)
- Deepayan Chakrabarti, Jure Leskovec, Christos Faloutsos, Samuel Madden, Carlos Guestrin, Michalis Faloutsos: *Information Survival Threshold in Sensor and P2P Networks*. INFOCOM 2007: 1316-1324

15-826 (c) 2012 C. Faloutsos 149

Carnegie Mellon

References

- Christos Faloutsos, Tamara G. Kolda, Jimeng Sun: *Mining large graphs and streams using matrix and tensor tools*. Tutorial, SIGMOD Conference 2007: 1174

15-826 (c) 2012 C. Faloutsos 150

Carnegie Mellon

References

- T. G. Kolda and J. Sun. *Scalable Tensor Decompositions for Multi-aspect Data Mining*. In: ICDM 2008, pp. 363-372, December 2008.

15-826 (c) 2012 C. Faloutsos 151

Carnegie Mellon

References

- Jure Leskovec, Jon Kleinberg and Christos Faloutsos *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations*, KDD 2005 (Best Research paper award).
- Jure Leskovec, Deepayan Chakrabarti, Jon M. Kleinberg, Christos Faloutsos: *Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication*. PKDD 2005: 133-145

15-826 (c) 2012 C. Faloutsos 152

Carnegie Mellon

References

- Jimeng Sun, Yinglian Xie, Hui Zhang, Christos Faloutsos. *Less is More: Compact Matrix Decomposition for Large Sparse Graphs*, SDM, Minneapolis, Minnesota, Apr 2007.
- Jimeng Sun, Spiros Papadimitriou, Philip S. Yu, and Christos Faloutsos, *GraphScope: Parameter-free Mining of Large Time-evolving Graphs* ACM SIGKDD Conference, San Jose, CA, August 2007

15-826 (c) 2012 C. Faloutsos 153

Carnegie Mellon

References

- Jimeng Sun, Dacheng Tao, Christos Faloutsos: *Beyond streams and graphs: dynamic tensor analysis*. KDD 2006: 374-383

15-826 (c) 2012 C. Faloutsos 154

Carnegie Mellon

References

- Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan, *Fast Random Walk with Restart and Its Applications*, ICDM 2006, Hong Kong.
- Hanghang Tong, Christos Faloutsos, *Center-Piece Subgraphs: Problem Definition and Fast Solutions*, KDD 2006, Philadelphia, PA

15-826 (c) 2012 C. Faloutsos 155

Carnegie Mellon


References







- Hanghang Tong, Christos Faloutsos, Brian Gallagher, Tina Eliassi-Rad: Fast best-effort pattern matching in large attributed graphs. KDD 2007: 737-746

15-826 (c) 2012 C. Faloutsos 156

Carnegie Mellon

(Project info)

www.cs.cmu.edu/~pegasus 

 Akoglu, Leman	 Kang, U	 Koutra, Danae	 McGlohon, Mary	 Prakash, Aditya	 Tong, Hanghang
---	--	---	--	---	--

Thanks to: NSF IIS-0705359, IIS-0534205,
CTA-INARC; Yahoo (M45), LLNL, IBM, SPRINT,
Google, INTEL, HP, iLab
