Carnegie Mellon

15-826: Multimedia Databases and Data Mining

(Project lecture #1)

Lecture #26: Graph mining - patterns

Christos Faloutsos

1

Carnegie Mellon

Must-read Material – 1-of-2

- [Graph minining textbook] Deepayan Chakrabarti and Christos Faloutsos *Graph Mining: Laws, Tools and Case Studies*, Morgan Claypool, 2012
  - Part I (patterns)

15-826 (c) C. Faloutsos, 2019 2

2

Carnegie Mellon

Must-read Material 2-of-2

- 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) C. Faloutsos, 2019 3

3

Carnegie Mellon

Problem

- Are real graphs random?

15-826 Copyright: C. Faloutsos (2019) 4

4

Carnegie Mellon

Conclusions

- Are real graphs random?
- NO!
  - Static patterns
    - Small diameters
    - Skewed degree distribution
    - Shrinking diameters
  - Weighted
  - Time-evolving

15-826 5

5

Carnegie Mellon

Conclusions

- Are real graphs random?
- NO!
  - Static patterns
    - Small diamete
    - Ske
  - e-evolving

- Many power laws – log-logistic
- Take logarithms

15-826 6

6

Carnegie Mellon

Main outline

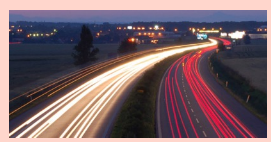
- Introduction
- Indexing
- Mining
  - Graphs – patterns
  - Graphs – generators and tools
  - Association rules
  - …

15-826 (c) C. Faloutsos, 2019 7

7

Carnegie Mellon

Outline



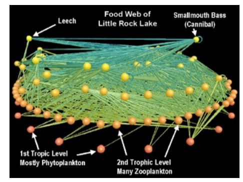
- Introduction – Motivation
- Problem: Patterns in graphs
- Problem#2: Scalability
- Conclusions

15-826 (c) C. Faloutsos, 2019 8

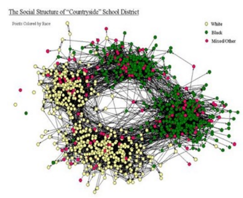
8

Carnegie Mellon

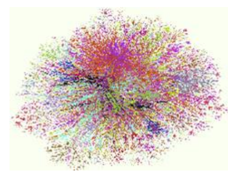
Graphs - why should we care?



Food Web
[Martinez ’91]



Friendship Network
[Moody ’01]

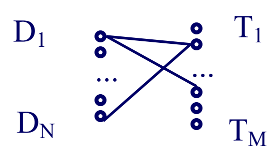


Internet Map
[lumeta.com]

Carnegie Mellon

Graphs - why should we care?

- IR: bi-partite graphs (doc-terms)



- web: hyper-text graph
- ... and more:

Carnegie Mellon

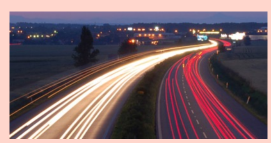
Graphs - why should we care?

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

15-826 (c) C. Faloutsos, 2019 11

11

Carnegie Mellon



Outline

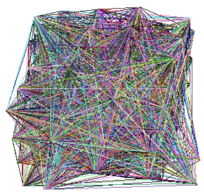
- Introduction – Motivation
- ➡ Problem: Patterns in graphs
  - Static graphs
  - Weighted graphs
  - Time evolving graphs
- Problem#2: Scalability
- Conclusions

15-826 (c) C. Faloutsos, 2019 12

12

Carnegie Mellon

Problem #1 - network and graph mining



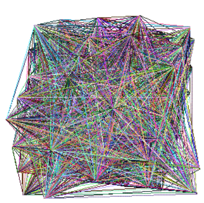
- What does the Internet look like?
- What does FaceBook look like?
- What is 'normal' / 'abnormal' ?
- which patterns/laws hold?

15-826 (c) C. Faloutsos, 2019 13

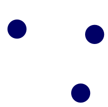
13

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?
  - **anomalies** (rarities) <-> **patterns**

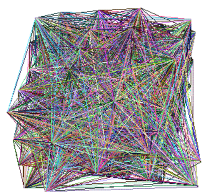


15-826 (c) C. Faloutsos, 2019 14

14

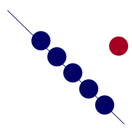
Carnegie Mellon

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?



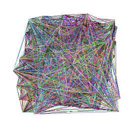
  - **anomalies** (rarities) <-> **patterns**
  - **Large** datasets reveal patterns/anomalies that may be invisible otherwise…

15-826 (c) C. Faloutsos, 2019 15

15

Carnegie Mellon

Graph mining



- Are real graphs random?

15-826 (c) C. Faloutsos, 2019 16

16

Carnegie Mellon

Laws and patterns

- Are real graphs random?
- A: NO!!
  - Diameter ('6 degrees', 'Kevin Bacon')
  - in- and out- degree distributions
  - other (surprising) patterns

- So, let's look at the data

15-826 (c) C. Faloutsos, 2019 17

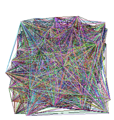
17

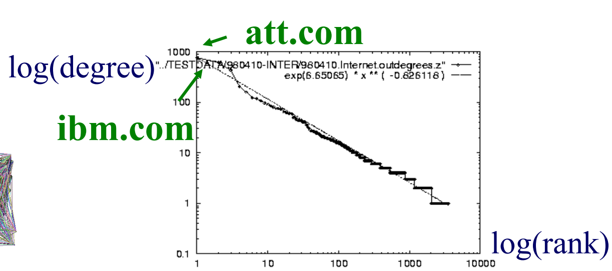
Carnegie Mellon

Solution# S.1

- Power law in the degree distribution [SIGCOMM99]

internet domains





15-826 (c) C. Faloutsos, 2019 18

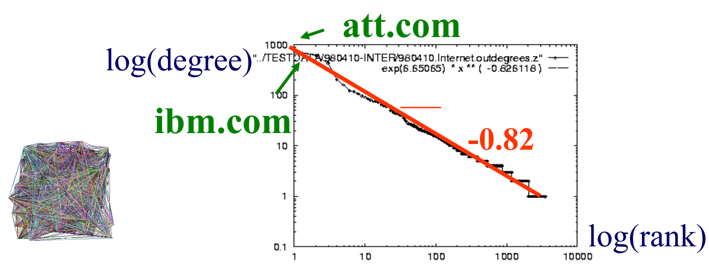
18

Carnegie Mellon

Solution# S.1

- Power law in the degree distribution [SIGCOMM99]

internet domains



15-826 (c) C. Faloutsos, 2019 19

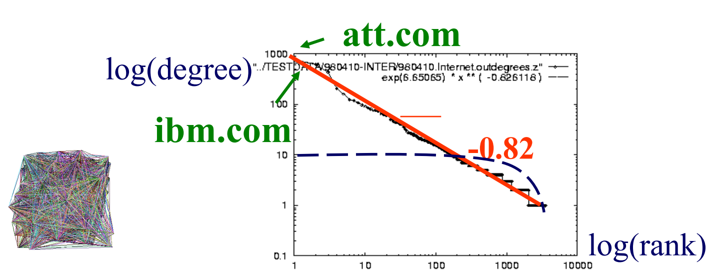
19

Carnegie Mellon

Solution# S.1

- Q: So what?

internet domains



15-826 (c) C. Faloutsos, 2019 20

20

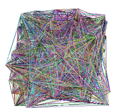
Carnegie Mellon

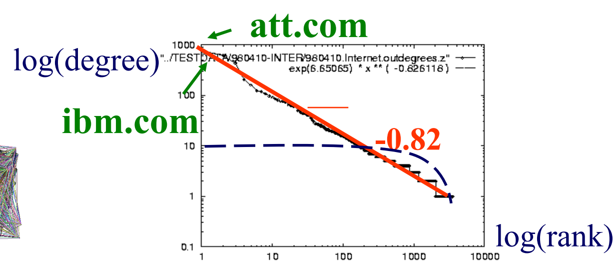
Solution# S.1

- Q: So what?
- A1: # of two-step-away pairs:

= friends of friends (F.O.F.)

internet domains





15-826 (c) C. Faloutsos, 2019 21

21

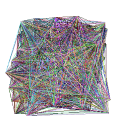
Carnegie Mellon

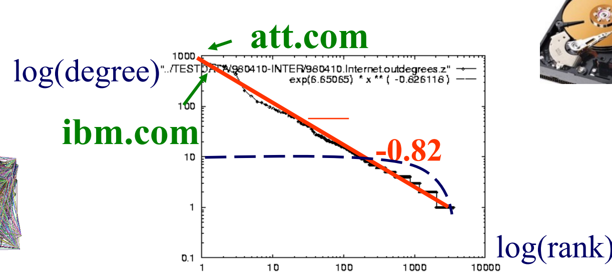
Solution# S.1

- Q: So what?
- A1: # of two-step-away pairs: 100^2 * N= 10 Trillion

= friends of friends (F.O.F.)

internet domains





15-826 (c) C. Faloutsos, 2019 22

22

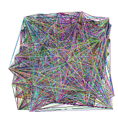
Carnegie Mellon

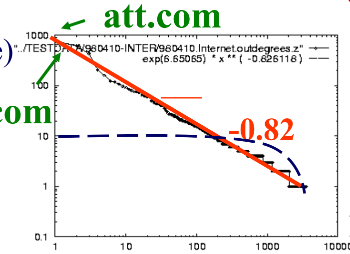
Solution# S.1

- Q: So what?
- A1: # of two-step-away pairs: 100^ … Trillion

= friends of friends (F.O.F.)

internet domains





15-826 (c) C. Faloutsos, 2019 23

23

Carnegie Mellon

Gaussian trap

Solution# S.1

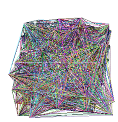
- Q: So what?
- A1: # of two-step-away pairs: O(d_max ^2) ~ 10M^2

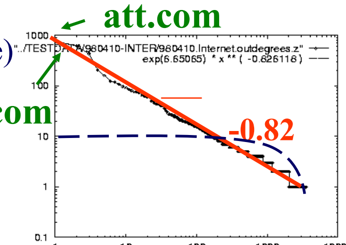
= friends of friends (F.O.F.)

internet domains

⇩

~0.8PB ->
a data center(!)





DCO @ CMU

15-826 (c) C. Faloutsos, 2019 24

24

Carnegie Mellon

Gaussian trap

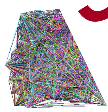
Solution# S.1

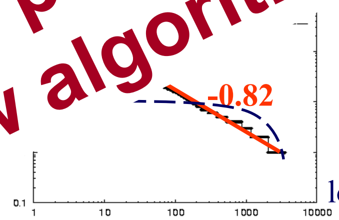
- Q: So what?
- A1: # of two-step-aw ?) ~ 10M^2

inte

Such patterns ->
New algorithms

~0.8PB ->
a data center(!)





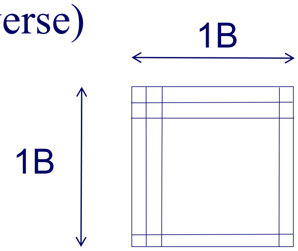
15-826 (c) C. Faloutsos, 2019 25

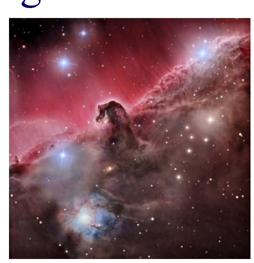
25

Carnegie Mellon

Observation – big-data:

- O(N^2) algorithms are ~intractable - N=1B
- N^2 seconds = 31B years (>2x age of universe)





15-826 (c) C. Faloutsos, 2019 26

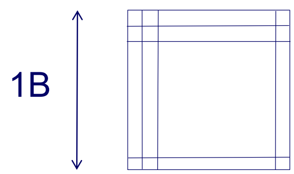
26

Carnegie Mellon

Observation – big-data:

- O(N^2) algorithms are ~intractable - N=1B

- N^2 seconds = ~~31B~~ 31M years
- 1,000 machines

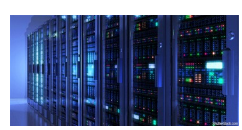


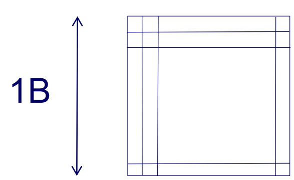
Carnegie Mellon

Observation – big-data:

- O(N^2) algorithms are ~intractable - N=1B

- N^2 seconds = ~~31B~~ 31K years
- 1M machines





Google Y!

Carnegie Mellon

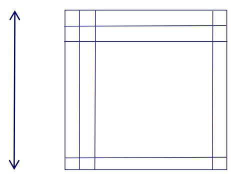
Observation – big-data:

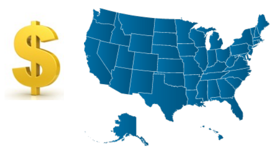
- O(N^2) algorithms are ~intractable - N=1B

3

- N^2 seconds = 31B years
- 10B machines ~ $10Trillion

1B





15-826 (c) C. Faloutsos, 2019 29

29

Carnegie Mellon

Observation – big-data:

- O(N^2) algorithms are ~intractable - N=1B

And parallelism might not help

3

- N^2 seconds = 31B years
- 10B machines ~ $10Trillion

1B

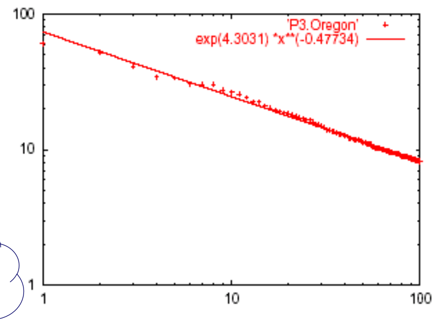
15-826 (c) C. Faloutsos, 2019 30

30

Carnegie Mellon

Solution# S.2: Eigen Exponent *E*

Eigenvalue



Rank of decreasing eigenvalue

$A\,x = \lambda\,x$

Exponent = slope

E = -0.48

May 2001

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

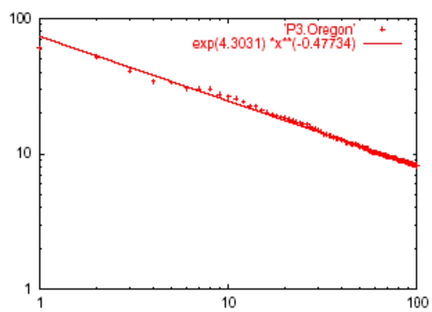
15-826 (c) C. Faloutsos, 2019 31

31

Carnegie Mellon

Solution# S.2: Eigen Exponent *E*

Eigenvalue

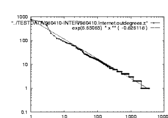


Rank of decreasing eigenvalue

Exponent = slope

E = -0.48

May 2001



- [Mihail, Papadimitriou ' 02]: slope is ½ of rank exponent

15-826 (c) C. Faloutsos, 2019 32

32

Carnegie Mellon

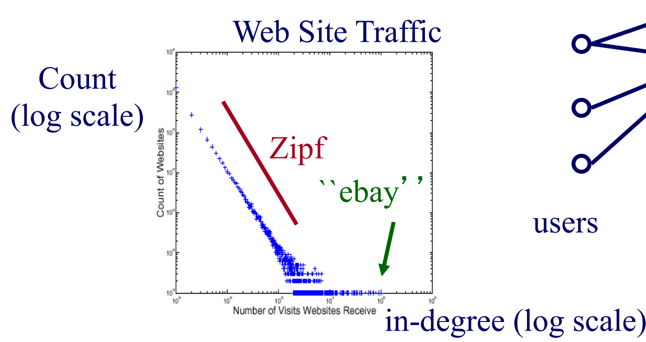
But:

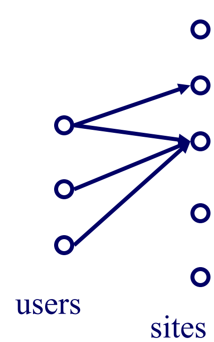
How about graphs from other domains?

Carnegie Mellon

More power laws:

- web hit counts [w/ A. Montgomery]



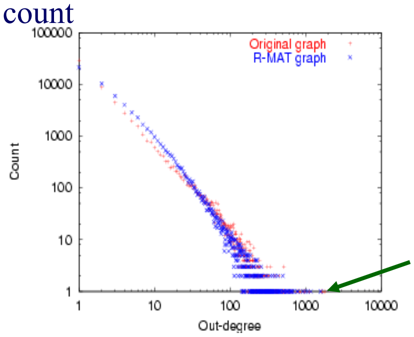


Carnegie Mellon

epinions.com

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

count



trusts-2000-people user

(out) degree

15-826 (c) C. Faloutsos, 2019 35

35

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) C. Faloutsos, 2019 36

36

Carnegie Mellon

List of Static Patterns

- ✔ S.1 degree
- ✔ S.2 eigenvalues
- S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

In textbook

15-826 (c) C. Faloutsos, 2019 37

37

Carnegie Mellon

S.3 small diameters

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

15-826 (c) C. Faloutsos, 2019 38

38

Carnegie Mellon

List of Static Patterns

- ✔ S.1 degree
- ✔ S.2 eigenvalues
- ✔ S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

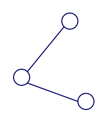
In textbook

15-826 (c) C. Faloutsos, 2019 39

39

Carnegie Mellon

Solution# S.4: Triangle 'Laws'



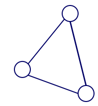
- Real social networks have a lot of triangles

15-826 (c) C. Faloutsos, 2019 40

40

Carnegie Mellon

Solution# S.4: Triangle ‘Laws’



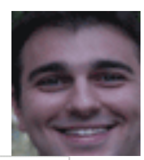
- Real social networks have a lot of triangles
  - Friends of friends are friends
- Any patterns?

15-826 (c) C. Faloutsos, 2019 41

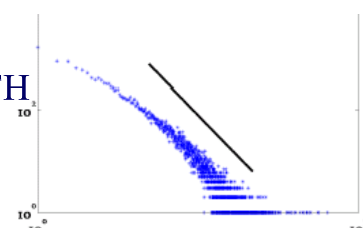
41

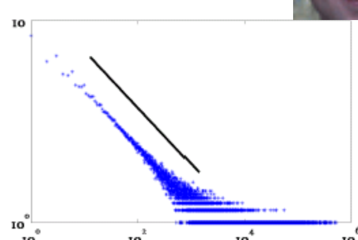
Carnegie Mellon

Triangle Law: #S.4
[Tsourakakis ICDM 2008]



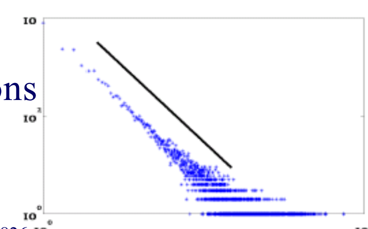
HEP-TH

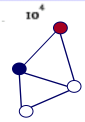




ASN

Epinions





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

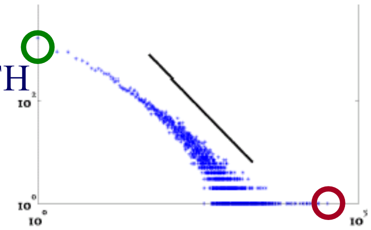
15-826 (c) C. Faloutsos, 2019 42

42

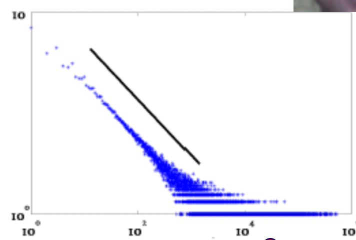
Carnegie Mellon

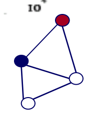
Triangle Law: #S.4
[Tsourakakis ICDM 2008]

HEP-TH

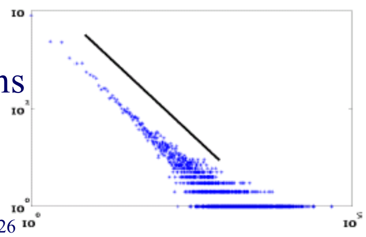


ASN





Epinions

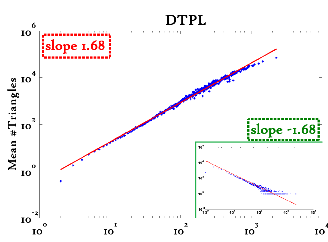


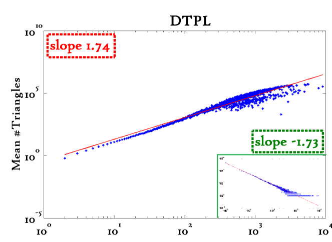
X-axis: # of participating triangles
Y: count (~ pdf)

Carnegie Mellon

Triangle Law: #S.5
[Tsourakakis ICDM 2008]

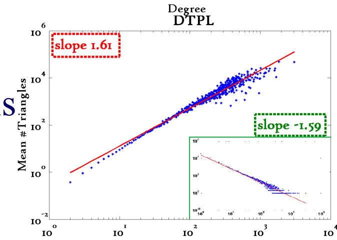
Reuters





SN

Epinions



X-axis: degree
Y-axis: mean # triangles
n friends -> ~$n^{1.6}$ triangles

Carnegie Mellon

details

Triangle Law: Computations
[Tsourakakis ICDM 2008]

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

Carnegie Mellon

details

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!

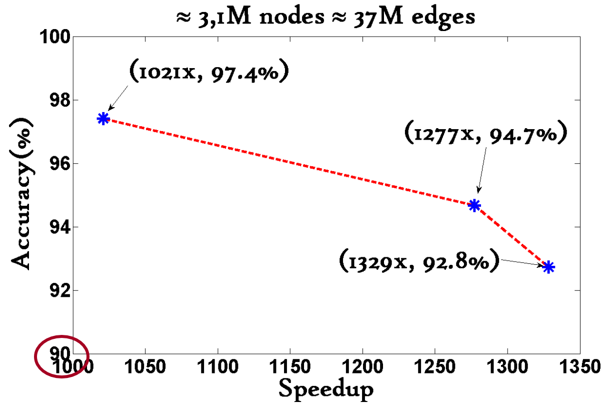
Carnegie Mellon

details

Triangle Law: Computations

[Tsourakakis ICDM 2008]

Wikipedia graph 2006-Nov-04
≈ 3,1M nodes ≈ 37M edges



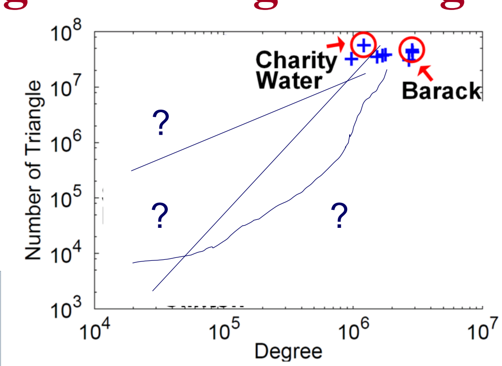
1000x+ speed-up, >90% accuracy

15-826 (c) C. Faloutsos, 2019 47

47

Carnegie Mellon

Triangle counting for large graphs?



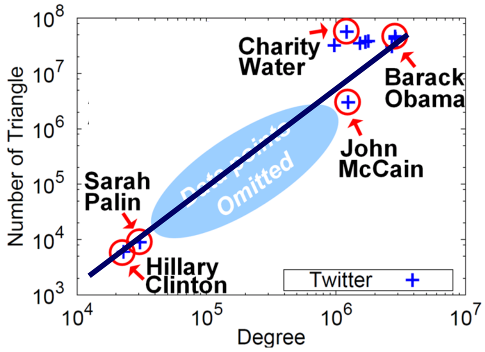
Anomalous nodes in Twitter(~ 3 billion edges)
[U Kang, Brendan Meeder, +, PAKDD’11]


15-826 (c) C. Faloutsos, 2019 48

48

Carnegie Mellon

Triangle counting for large graphs?



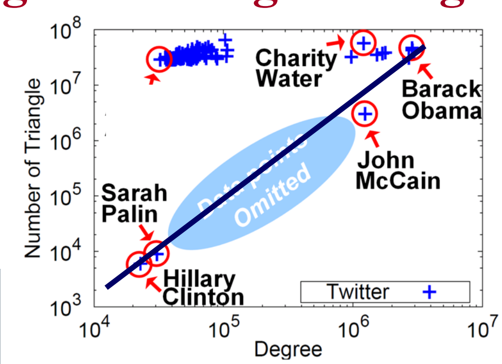
Anomalous nodes in Twitter(~ 3 billion edges)
[U Kang, Brendan Meeder, +, PAKDD’11]

15-826 (c) C. Faloutsos, 2019 49

49

Carnegie Mellon

Triangle counting for large graphs?



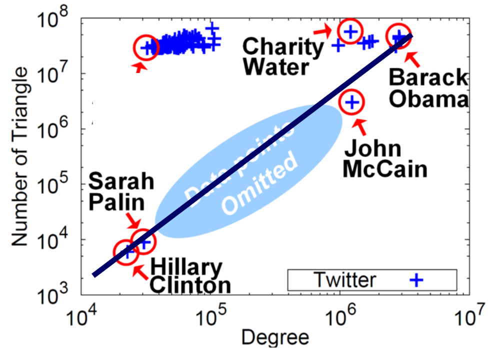
Anomalous nodes in Twitter(~ 3 billion edges)
[U Kang, Brendan Meeder, +, PAKDD’11]

15-826 (c) C. Faloutsos, 2019 50

50

Carnegie Mellon

Triangle counting for large graphs?



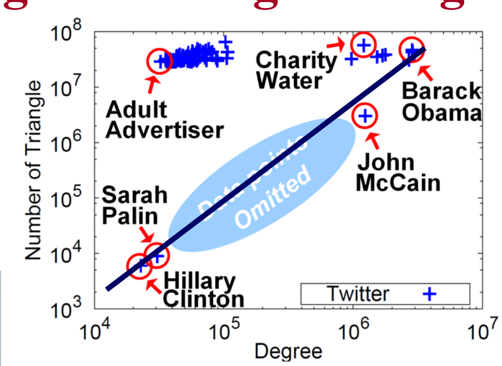
Anomalous nodes in Twitter(~ 3 billion edges)
[U Kang, Brendan Meeder, +, PAKDD'11]

15-826 (c) C. Faloutsos, 2019 51

51

Carnegie Mellon

Triangle counting for large graphs?



Anomalous nodes in Twitter(~ 3 billion edges)
[U Kang, Brendan Meeder, +, PAKDD'11]

15-826 (c) C. Faloutsos, 2019 52

52

Carnegie Mellon

List of Static Patterns

- ✔ S.1 degree
- ✔ S.2 eigenvalues
- ✔ S.3 small diameter
- ✔ S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

In textbook

15-826 (c) C. Faloutsos, 2019 53

53

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).

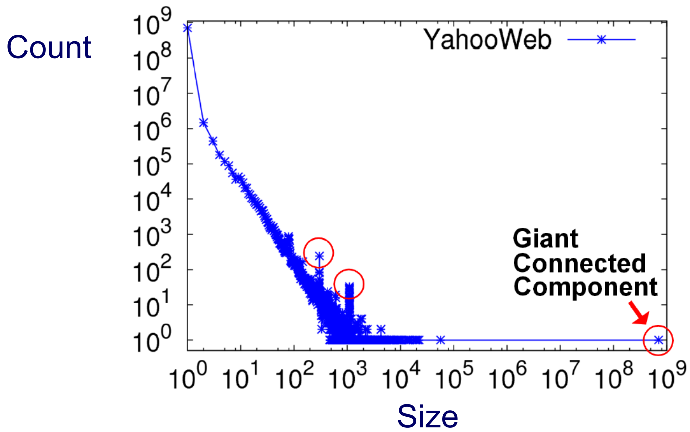
15-826 (c) C. Faloutsos, 2019 54

54

Carnegie Mellon

S.6: NLCC

- Connected Components – 4 observations:



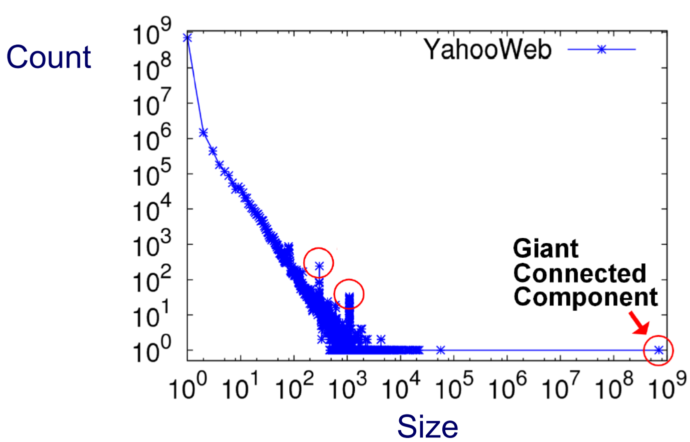
15-826 (c) C. Faloutsos, 2019 55

55

Carnegie Mellon

S.6: NLCC

- Connected Components



1) 10K x larger than next

15-826 (c) C. Faloutsos, 2019 56

56

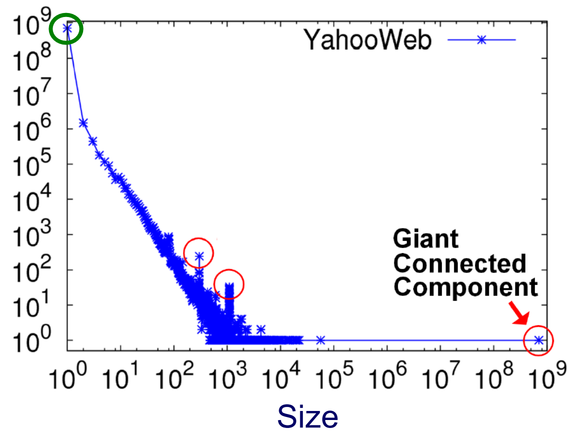
Carnegie Mellon

S.6: NLCC

- Connected Components

Count

2) ~0.7B singleton nodes



15-826 (c) C. Faloutsos, 2019 57

57

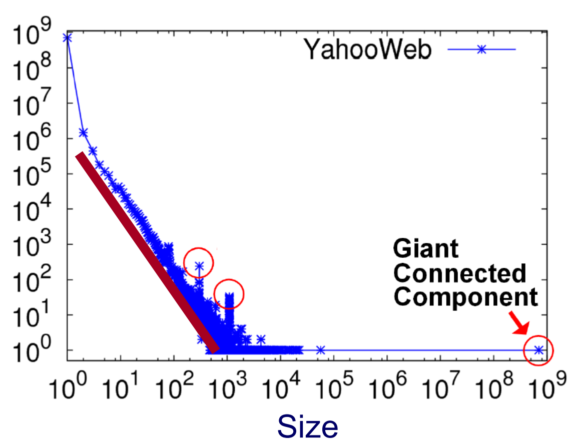
Carnegie Mellon

S.6: NLCC

- Connected Components

Count

3) SLOPE!

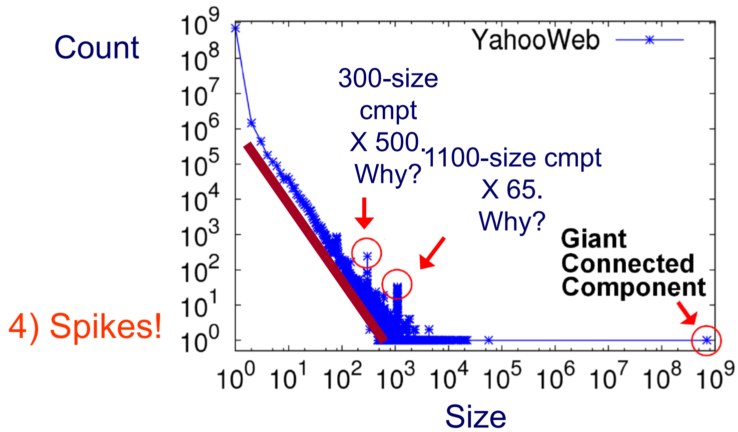


15-826 (c) C. Faloutsos, 2019 58

58

S.6: NLCC

- Connected Components

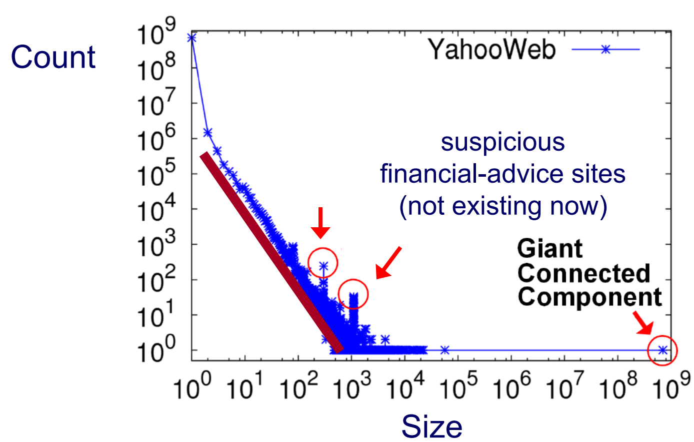


15-826 (c) C. Faloutsos, 2019 59

59

S.6: NLCC

- Connected Components

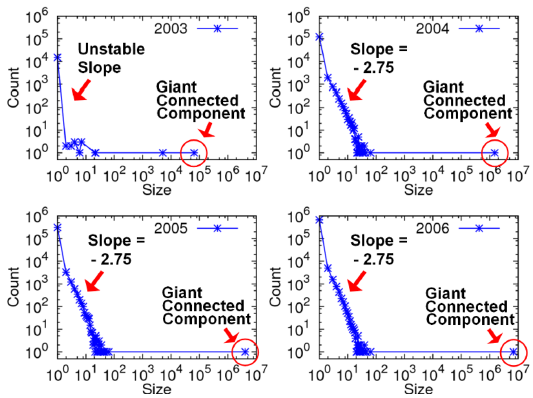


15-826 (c) C. Faloutsos, 2019 60

60

S.6: persists over time

- Connected Components over Time
- **LinkedIn: 7.5M nodes and 58M edges**



Stable tail slope after the gelling point

15-826 (c) C. Faloutsos, 2019 61

61

List of Static Patterns

- ✔ S.1 degree
- ✔ S.2 eigenvalues
- ✔ S.3 small diameter
- ✔ S.4/5 Triangle laws

} In textbook

- ✔ (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

15-826 (c) C. Faloutsos, 2019 62

62

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.

Useful for fraud detection!

15-826 (c) C. Faloutsos, 2019 63

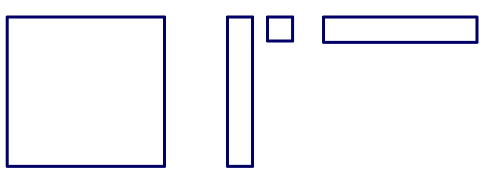
63

Carnegie Mellon

EigenSpokes

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

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



15-826 (c) C. Faloutsos, 2019 64

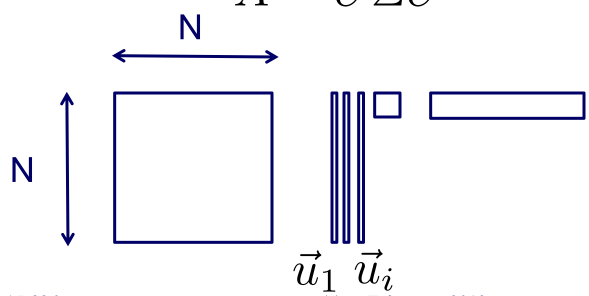
64

Carnegie Mellon

EigenSpokes

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

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

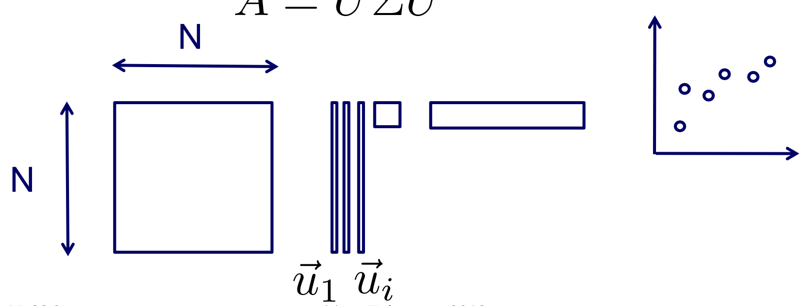


Carnegie Mellon

EigenSpokes

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

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

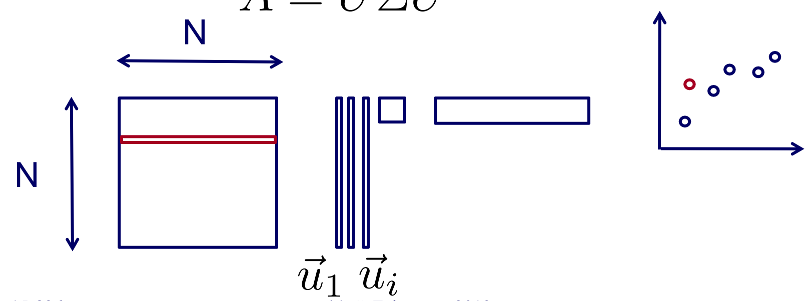


Carnegie Mellon

EigenSpokes

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

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



15-826 (c) C. Faloutsos, 2019 67

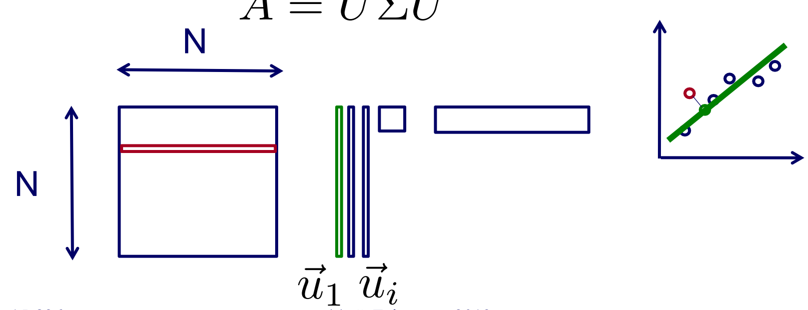
67

Carnegie Mellon

EigenSpokes

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

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



15-826 (c) C. Faloutsos, 2019 68

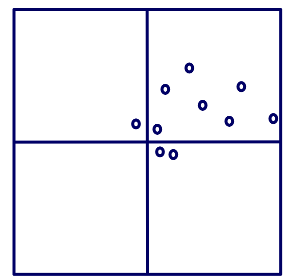
68

Carnegie Mellon

EigenSpokes

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

2nd Principal component
u2



u1
1st Principal component

15-826 (c) C. Faloutsos, 2019 69

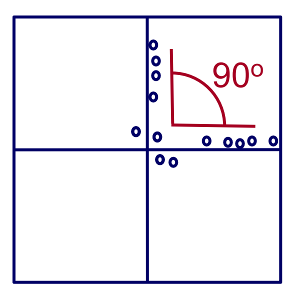
69

Carnegie Mellon

EigenSpokes

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

u2



u1

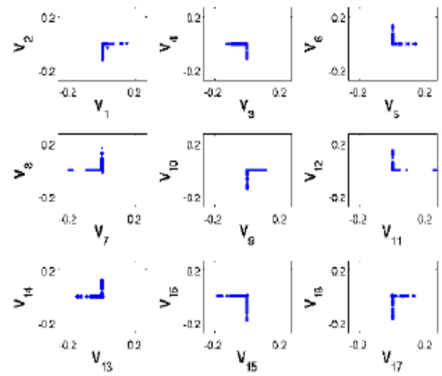
15-826 (c) C. Faloutsos, 2019 70

70

Carnegie Mellon

EigenSpokes - pervasiveness

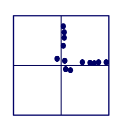
- Present in mobile social graph
  - across time and space
- Patent citation graph



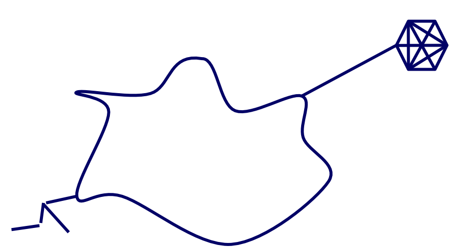
71

Carnegie Mellon

EigenSpokes - explanation



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

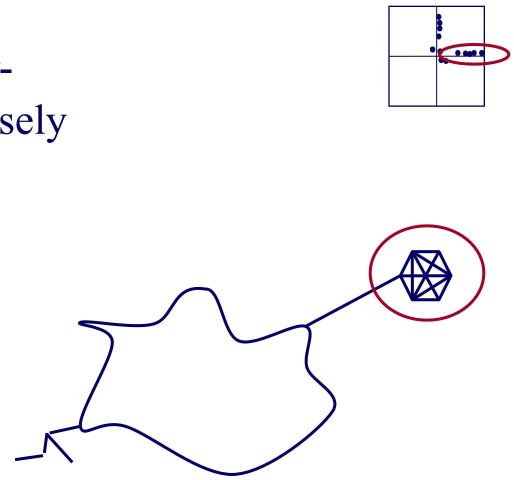


72

Carnegie Mellon

EigenSpokes - explanation

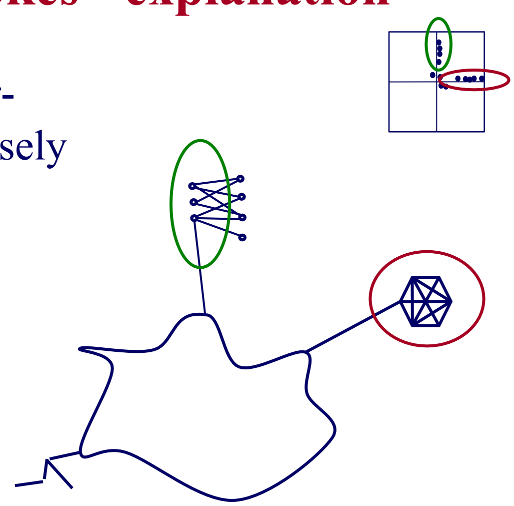
Near-cliques, or near-bipartite-cores, loosely connected



Carnegie Mellon

EigenSpokes - explanation

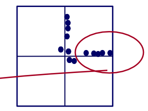
Near-cliques, or near-bipartite-cores, loosely connected



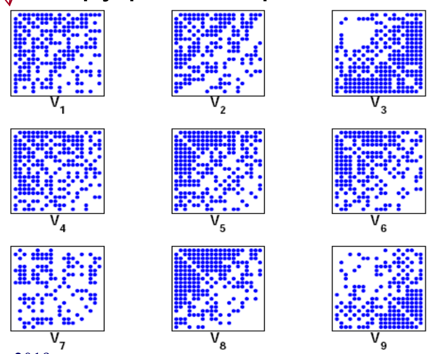
Carnegie Mellon

EigenSpokes - explanation

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



spy plot of top 20 nodes



So what?

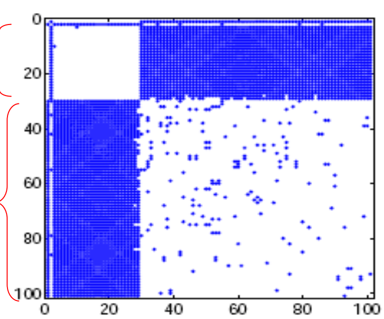
- Extract nodes with high *scores*
- high connectivity
- Good "communities"

Carnegie Mellon

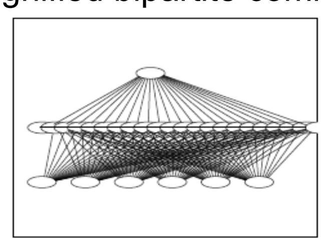
Bipartite Communities!

patents from same inventor(s)

\`cut-and-paste' bibliography!



magnified bipartite community



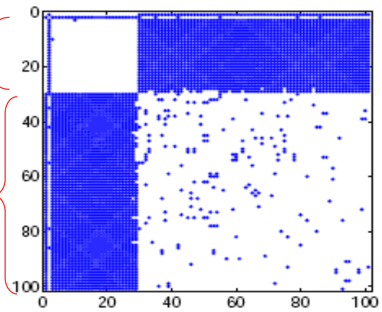
Useful for fraud detection!

Carnegie Mellon

Bipartite Communities!

IP – port scanners

victims



Useful for fraud detection!

15-826 (c) C. Faloutsos, 2019 77

77

Carnegie Mellon

List of Static Patterns

- ✔ • S.1 degree
- ✔ • S.2 eigenvalues
- ✔ • S.3 small diameter
- ✔ • S.4/5 Triangle laws
- ✔ • (S.6) NLCC non-largest conn. components
- ✔ • (S.7) eigen plots
- • (S.8) radius plot

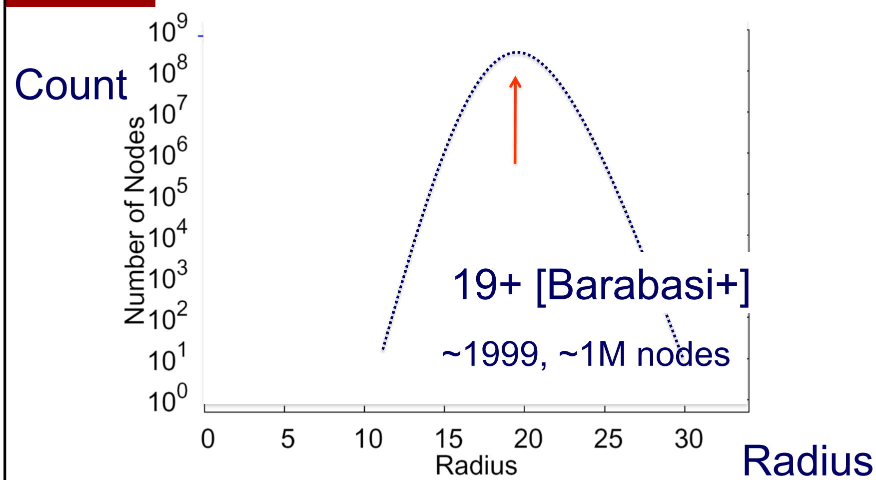
In textbook

15-826 (c) C. Faloutsos, 2019 78

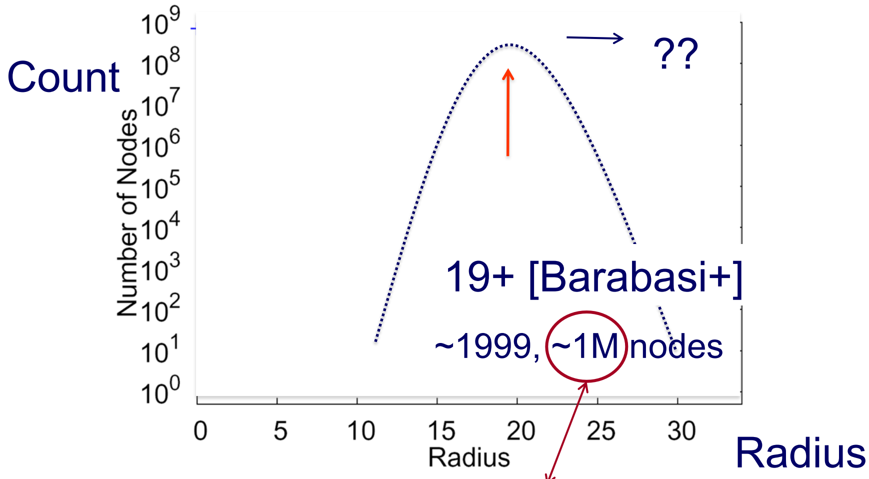
78

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~1B)
- Our HADI: linear on E (~10B)
  – Near-linear scalability wrt # machines
  – Several optimizations -> 5x faster



Carnegie Mellon



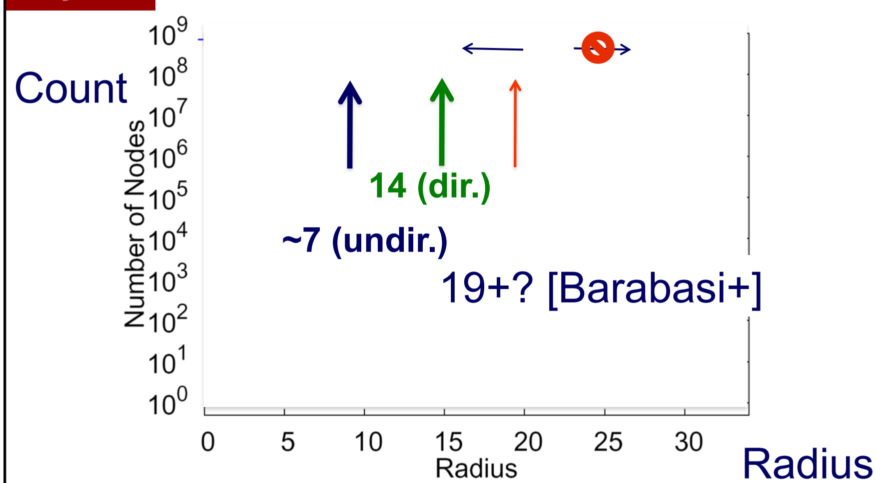
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- Largest publicly available graph ever studied.

15-826 (c) C. Faloutsos, 2019 81

81

Carnegie Mellon

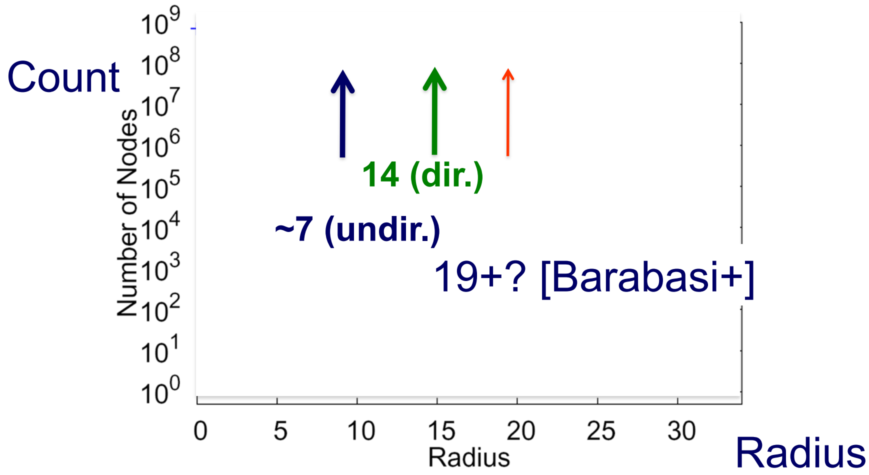


YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- Largest publicly available graph ever studied.

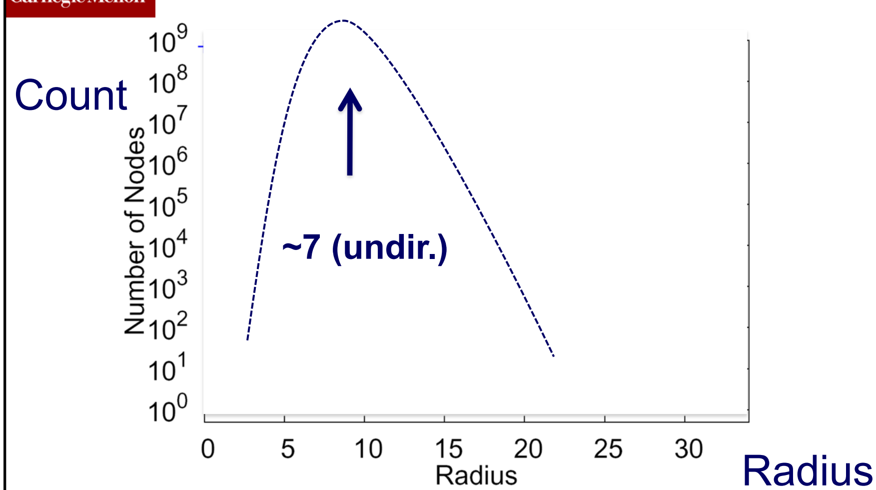
15-826 (c) C. Faloutsos, 2019 82

82



YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
•7 degrees of separation (!)
•Diameter: shrunk

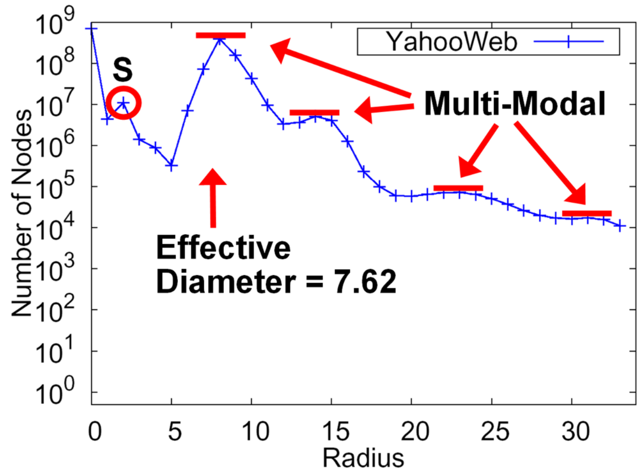
83



YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
Q: Shape?

84

Carnegie Mellon



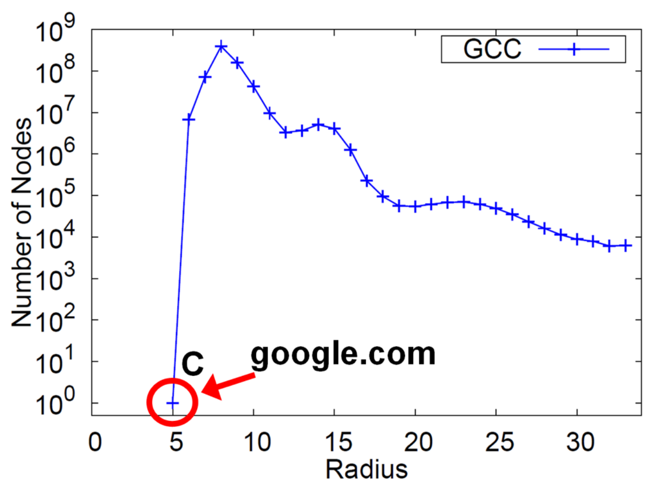
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- effective diameter: surprisingly small.
- Multi-modality (?!)

15-826 (c) C. Faloutsos, 2019 85

85

Carnegie Mellon

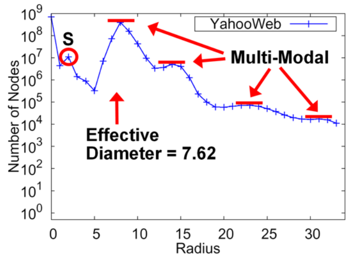


Radius Plot of **GCC** of YahooWeb.

15-826 (c) C. Faloutsos, 2019 86

86

Carnegie Mellon



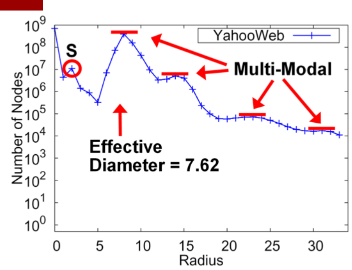
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- effective diameter: surprisingly small.
- Multi-modality: probably mixture of cores .

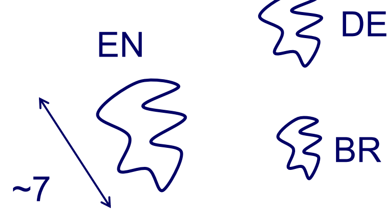
15-826 (c) C. Faloutsos, 2019 87

87

Carnegie Mellon



Conjecture:



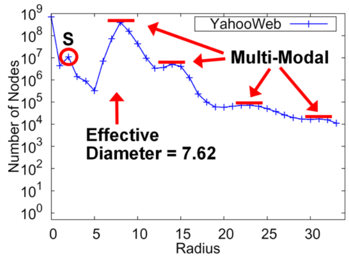
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)

- effective diameter: surprisingly small.
- Multi-modality: probably mixture of cores .

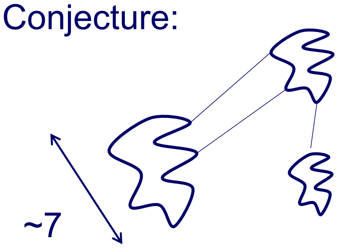
15-826 (c) C. Faloutsos, 2019 88

88

Carnegie Mellon



Conjecture:



YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
- effective diameter: surprisingly small.
- Multi-modality: probably mixture of cores .

15-826 (c) C. Faloutsos, 2019 89

89

Carnegie Mellon

List of Static Patterns

- ✔ S.1 degree
- ✔ S.2 eigenvalues
- ✔ S.3 small diameter
- ✔ S.4/5 Triangle laws

(S.1–S.4/5: In textbook)

- ✔ (S.6) NLCC non-largest conn. components
- ✔ (S.7) eigen plots
- ✔ (S.8) radius plot
- Other observations / patterns?

15-826 (c) C. Faloutsos, 2019 90

90

Carnegie Mellon

Attention: Phase1

List of Static Patterns

- ✔ S.1 degree
- ✔ S.2 eigenvalues
- ✔ S.3 small diameter
- ✔ S.4/5 Triangle laws

In textbook

- ✔ (S.6) NLCC non-largest conn. components
- ✔ (S.7) eigen plots
- ✔ (S.8) radius plot
- Other observations / patterns?

15-826 (c) C. Faloutsos, 2019 91

91

Carnegie Mellon

Any other 'laws'?

Yes!

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

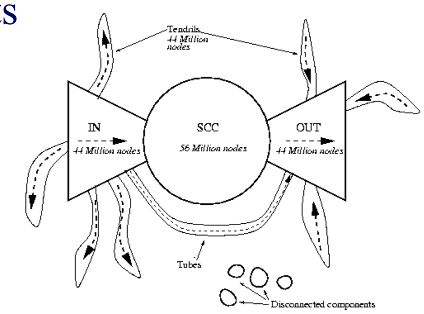
15-826 (c) C. Faloutsos, 2019 92

92

Carnegie Mellon

Any other 'laws'?

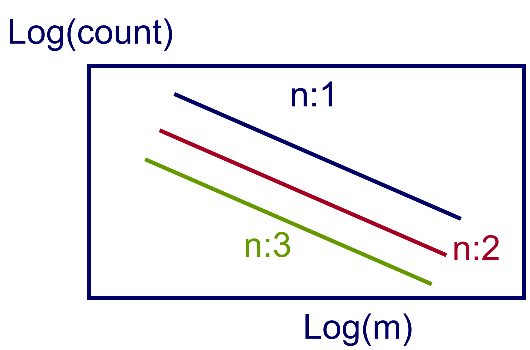
- Bow-tie, for the web [Kumar+ '99]
- IN, SCC, OUT, 'tendrils'
- disconnected components

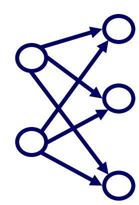


Carnegie Mellon

Any other 'laws'?

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



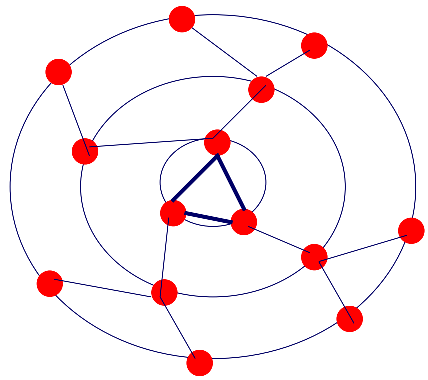


2:3 core
(m:n core)

Carnegie Mellon

Any other ‘laws’?

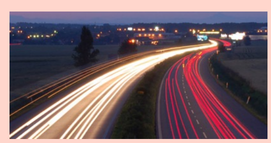
- “Jellyfish” for Internet [Tauro+ ’01]
- core: ~clique
- ~5 concentric layers
- many 1-degree nodes



15-826 (c) C. Faloutsos, 2019 95

95

Carnegie Mellon



Outline

- Introduction – Motivation
- Problem: Patterns in graphs
  - Static graphs
    - degree, diameter, eigen,
    - Triangles
  - ➡ Weighted graphs
  - Time evolving graphs
- Problem#2: Scalability
- Conclusions

15-826 (c) C. Faloutsos, 2019 96

96

CarnegieMellon

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) C. Faloutsos, 2019 97

97

CarnegieMellon

Observation W.1: Fortification

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

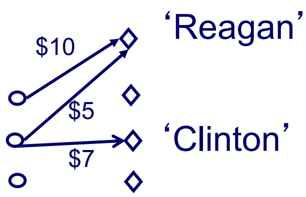
15-826 (c) C. Faloutsos, 2019 98

98

Carnegie Mellon

Observation W.1: Fortification

More donors, more $?



15-826 (c) C. Faloutsos, 2019 99

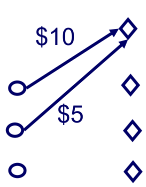
99

Carnegie Mellon

Observation W.1: fortification: Snapshot Power Law

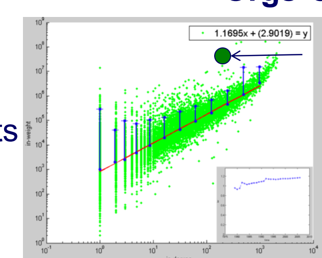
- Weight: super-linear on in-degree
- exponent 'iw' : 1.01 < iw < 1.26

More donors, <u>even</u> more $



Orgs-Candidates

In-weights ($)



Edges (# donors)

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

15-826 (c) C. Faloutsos, 2019 100

100

Carnegie Mellon

Outline

- Introduction – Motivation
- Problem: Patterns in graphs
  - Static graphs
  - Weighted graphs
  - Time evolving graphs
- Problem#2: Scalability
- Conclusions

15-826 (c) C. Faloutsos, 2019 101

101

Carnegie Mellon

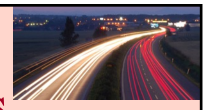
Problem: Time evolution

- with Jure Leskovec (CMU -> Stanford)

- and Jon Kleinberg (Cornell – sabb. @ CMU)

15-826 (c) C. Faloutsos, 2019 102

102



List of Dynamic Patterns

- D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
- D.5 Eigenvalue over time
- D.6 Popularity over time
- D.7 phonecall duration

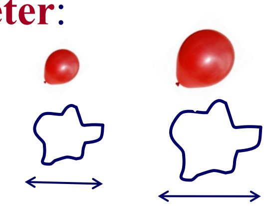
In textbook

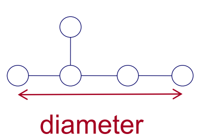
15-826 (c) C. Faloutsos, 2019 103

103

D.1 Evolution of the Diameter

- Prior work on Power Law graphs hints at **slowly growing diameter**:
  - [diameter ~ O($N^{1/3}$)]
  - diameter ~ O(log N)
  - diameter ~ O(log log N)
- What is happening in real data?





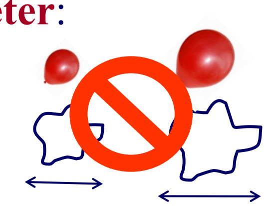
15-826 (c) C. Faloutsos, 2019 104

104

Carnegie Mellon

D.1 Evolution of the Diameter

- Prior work on Power Law graphs hints at **slowly growing diameter**:
  - [diameter ~ O($N^{1/3}$)]
  - diameter ~ O(log N)
  - diameter ~ O(log log N)



- What is happening in real data?
- Diameter **shrinks** over time

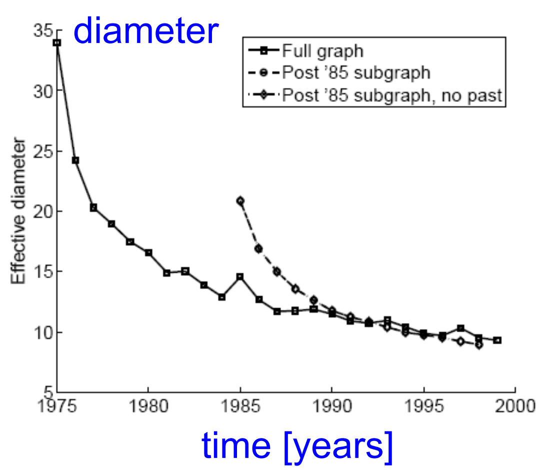
15-826 (c) C. Faloutsos, 2019 105

105

Carnegie Mellon

D.1 Diameter – "Patents"

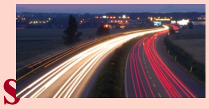
- Patent citation network
- 25 years of data
- @1999
  - 2.9 M nodes
  - 16.5 M edges



15-826 (c) C. Faloutsos, 2019 106

106

Carnegie Mellon



List of Dynamic Patterns

- ✔ D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
- D.5 Eigenvalue over time
- D.6 Popularity over time
- D.7 phonecall duration

In textbook

15-826 (c) C. Faloutsos, 2019 107

107

Carnegie Mellon

D.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) C. Faloutsos, 2019 108

108

Carnegie Mellon

D.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)

- A: over-doubled!
  - But obeying the ``Densification Power Law''

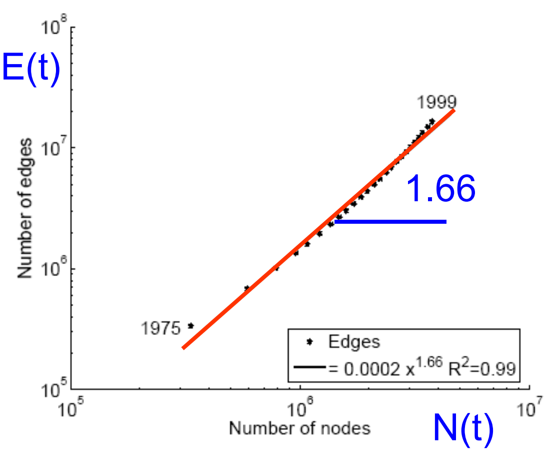
15-826 (c) C. Faloutsos, 2019 109

109

Carnegie Mellon

D.2 Densification – Patent Citations

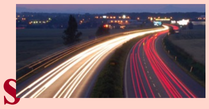
- Citations among patents granted
- @1999
  - 2.9 M nodes
  - 16.5 M edges
- Each year is a datapoint



15-826 (c) C. Faloutsos, 2019 110

110

Carnegie Mellon



List of Dynamic Patterns

- ✔ D.1 diameter
- ✔ D.2 densification
- D.3 gelling point
- D.4 NLCC over time
- D.5 Eigenvalue over time
- D.6 Popularity over time
- D.7 phonecall duration

In textbook

Carnegie Mellon

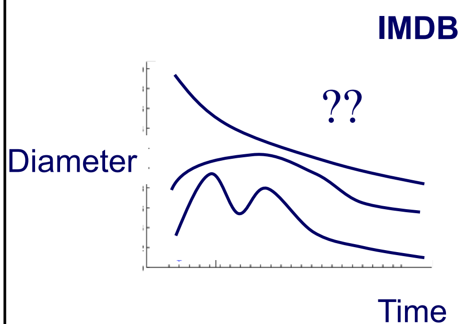
More on Time-evolving graphs

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

Carnegie Mellon

D.3 Gelling Point

- Diameter, over time



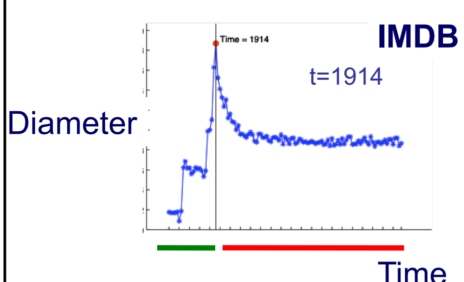
15-826 (c) C. Faloutsos, 2019 113

113

Carnegie Mellon

D.3 Gelling Point

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



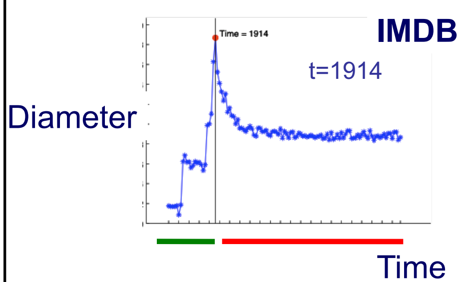
15-826 (c) C. Faloutsos, 2019 114

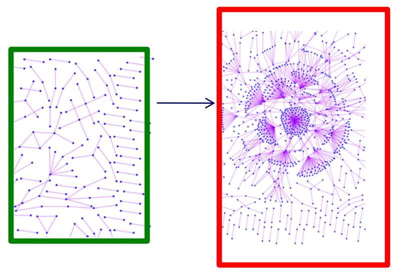
114

Carnegie Mellon

D.3 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) C. Faloutsos, 2019 115

115

Carnegie Mellon

List of Dynamic Patterns

- ✔ D.1 diameter
- ✔ D.2 densification
- ✔ D.3 gelling point
- D.4 NLCC over time
- D.5 Eigenvalue over time
- D.6 Popularity over time
- D.7 phonecall duration

$\Big\}$ In textbook (D.1 – D.5)

15-826 (c) C. Faloutsos, 2019 116

116

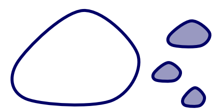
Carnegie Mellon

Observation D.4: 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) C. Faloutsos, 2019 117

117

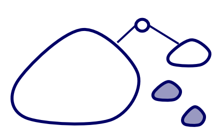
Carnegie Mellon

Observation D.4: 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 <u>shrink</u>?
- or stabilize?



15-826 (c) C. Faloutsos, 2019 118

118

Carnegie Mellon

Observation D.4: 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) C. Faloutsos, 2019 119

119

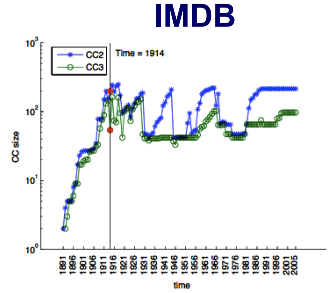
Carnegie Mellon

Observation D.4: NLCC behavior

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

IMDB

CC size

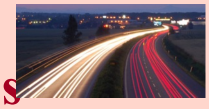


Time-stamp

15-826 (c) C. Faloutsos, 2019 120

120

Carnegie Mellon



List of Dynamic Patterns

- D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
- ~~D.5 Eigenvalue over time~~
- D.6 Popularity over time
- D.7 phonecall duration

In textbook

15-826 (c) C. Faloutsos, 2019 121

121

Carnegie Mellon

Timing for Blogs

Cascading Behavior in Large Blog Graphs: Patterns and a model

Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, Matthew Hurst

SDM'07

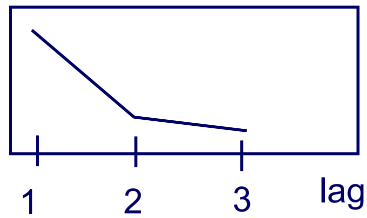
15-826 (c) C. Faloutsos, 2019 122

122

Carnegie Mellon

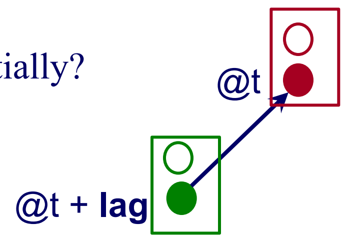
D.6 : popularity over time

in links



lag: days after post

Post popularity drops-off – exponentially?



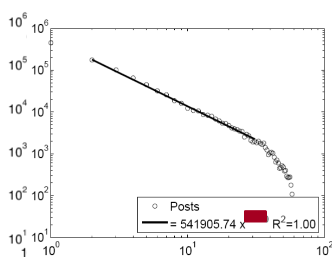
15-826 (c) C. Faloutsos, 2019 123

123

Carnegie Mellon

D.6 : popularity over time

in links
(**log**)



days after post
(**log**)

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

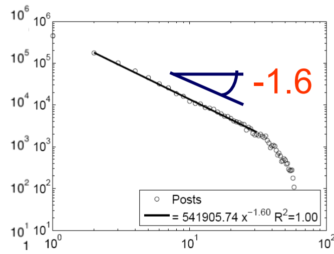
15-826 (c) C. Faloutsos, 2019 124

124

Carnegie Mellon

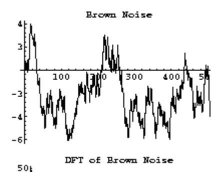
D.6 : popularity over time

in links
(**log**)



days after post
(**log**)

Post popularity drops-off – exponentially?
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) C. Faloutsos, 2019 125

125

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]

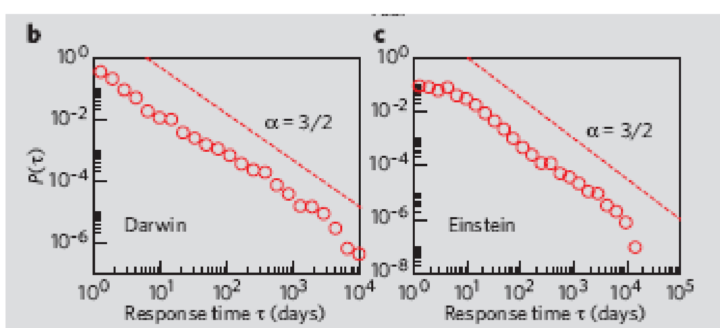
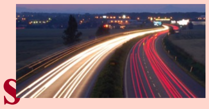


Figure 1 | The correspondence patterns of Darwin and Einstein.

1 126

126

Carnegie Mellon



List of Dynamic Patterns

- D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
- ~~D.5 Eigenvalue over time~~
- D.6 Popularity over time
- D.7 phonecall duration

In textbook

15-826 (c) C. Faloutsos, 2019 127

127

Carnegie Mellon

D.7: 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

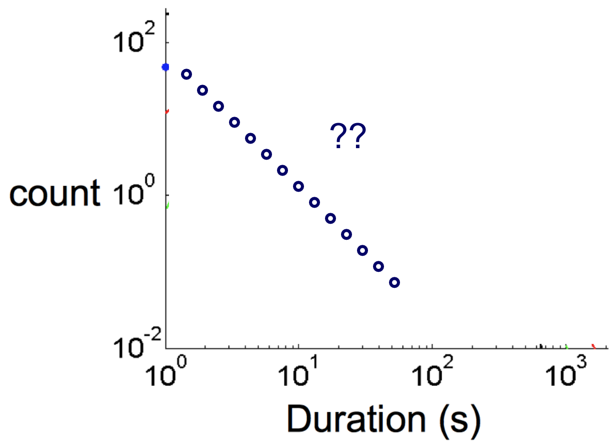
PKDD 2010

15-826 (c) C. Faloutsos, 2019 128

128

Carnegie Mellon

Probably, power law (?)

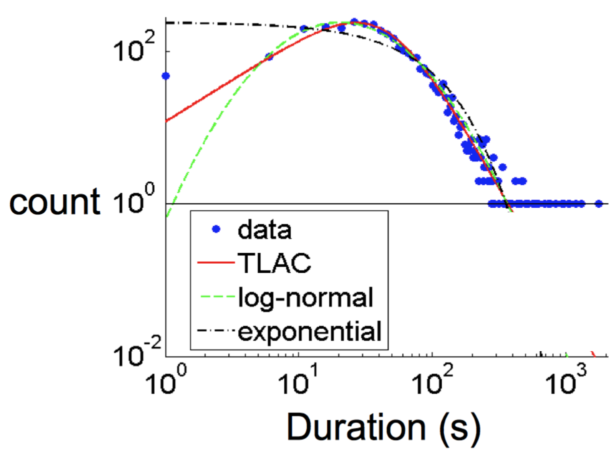


15-826 (c) C. Faloutsos, 2019 129

129

Carnegie Mellon

No Power Law!



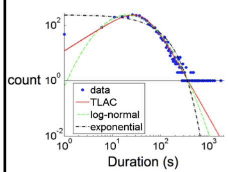
15-826 (c) C. Faloutsos, 2019 130

130

Carnegie Mellon

'TLaC: Lazy Contractor'

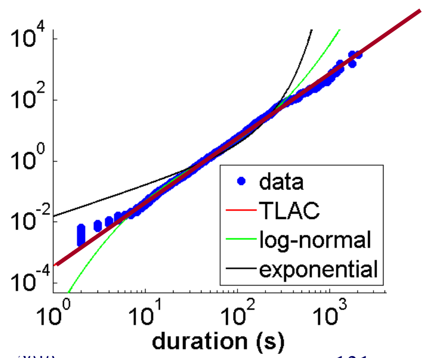
- The longer a task (phonecall) has taken,
- The even longer it will take



Odds ratio=

Casualties(<x): Survivors(>=x)

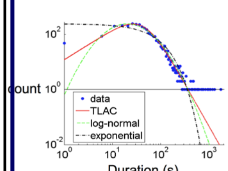
== power law



Carnegie Mellon

Log-logistic distribution

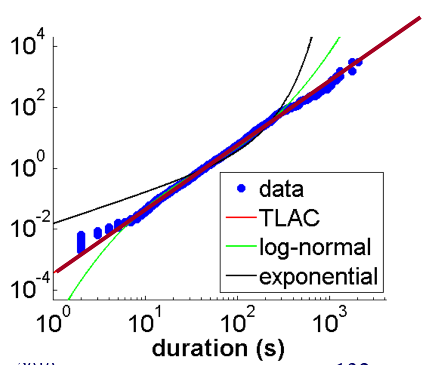
- CDF(t)/(1- CDF(t)) == OR(t)
- For log-logistic: log[OR(t)] = $\beta + \rho$*log(t)



Odds ratio=

Casualties(<x): Survivors(>=x)

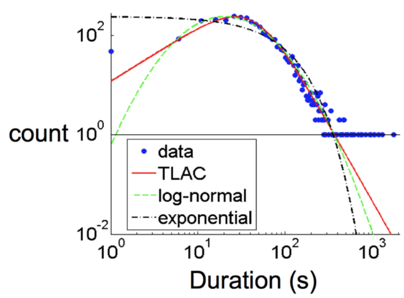
== power law



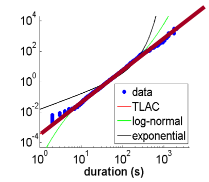
Carnegie Mellon

Log-logistic distribution

- CDF(t)/(1- CDF(t)) == OR(t)
- For log-logistic: log[OR(t)] = β + ρ*log(t)



OR(t)



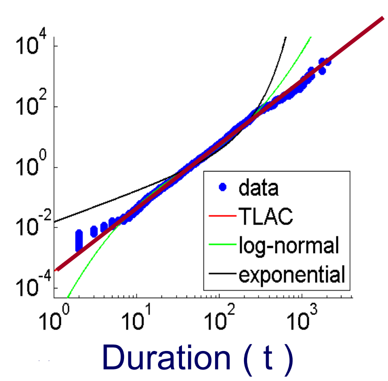
- PDF looks like hyperbola;
- and, if clipped, like power-law

Carnegie Mellon

Log-logistic distribution

- CDF(t)/(1- CDF(t)) == OR(t)
- For log-logistic: $\boxed{\log[OR(t)] = \beta + \rho*\log(t)}$

OR(t)

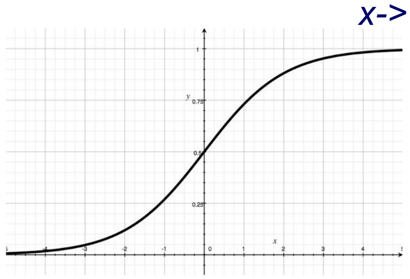


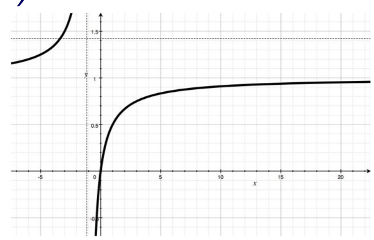
Carnegie Mellon

Log-logistic distribution

- Logistic distribution: CDF -> sigmoid
- LOG-Logistic distribution:

x-> ln(x)





CDF(x) = 1/(1+exp(-x))

CDF(x) = 1/(1+1/x)

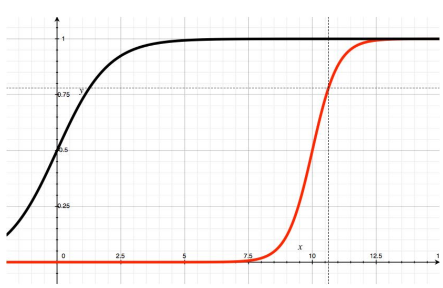
15-826 (c) C. Faloutsos, 2019 135

135

Carnegie Mellon

Log-logistic distribution

- Logistic distribution: CDF -> sigmoid
- LOG-Logistic distribution:



CDF(x) = 1/(1+exp(-(x-m)/s))

CDF(x) = 1/(1+exp(-(ln(x)-m)/s))

15-826 (c) C. Faloutsos, 2019 136

136

Carnegie Mellon

Attention: Phase1

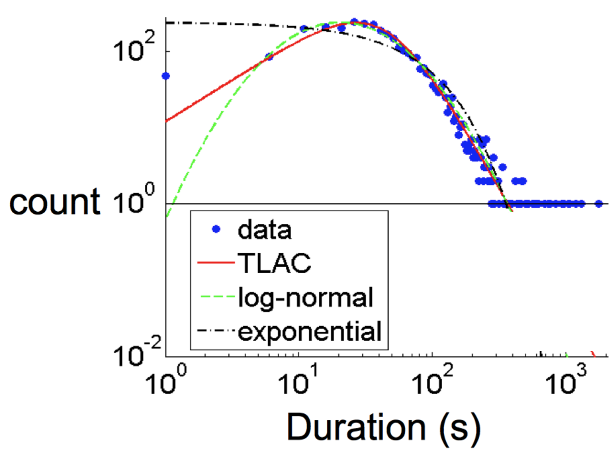
Log-logistic distribution

Nice 1 page description: section II of

Pravallika Devineni, Danai Koutra, Michalis Faloutsos, and Christos Faloutsos.
If walls could talk: Patterns and anomalies in Facebook wallposts.
ASONAM 2015, pp 367-374.

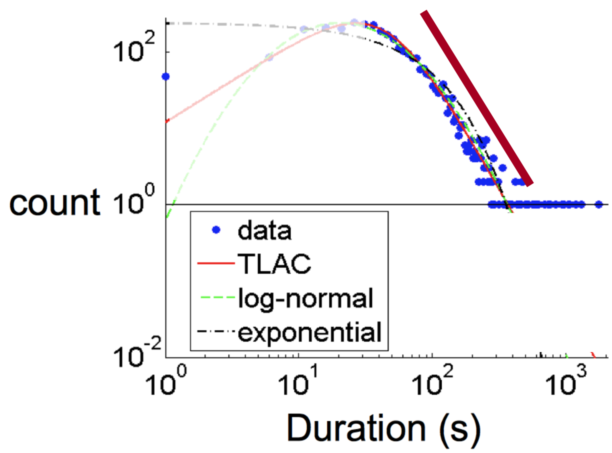
Carnegie Mellon

Log-logistic: ~ power law



Carnegie Mellon

Log-logistic: ~ power law



139

Carnegie Mellon

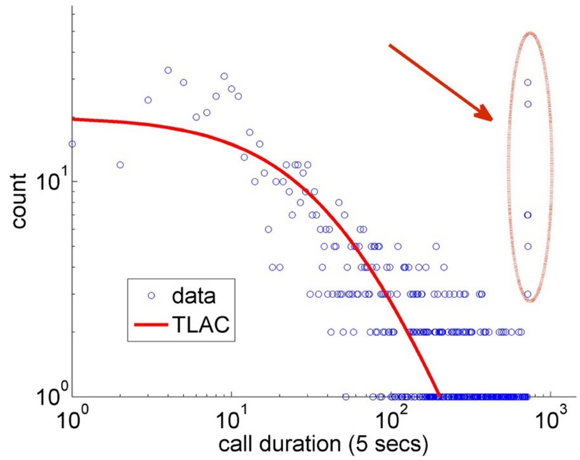
Data Description

- 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)

140

Carnegie Mellon

Outliers:



15-826 (c) C. Faloutsos, 2019 141

141

Carnegie Mellon

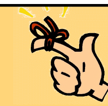
Conclusions

- Are real graphs random?
- NO!
  - Static patterns
    - Small diameters
    - Skewed degree distribution
    - Shrinking diameters
  - Weighted
  - Time-evolving

15-826 Copyright: C. Faloutsos (2019) 142

142

Carnegie Mellon



Conclusions

- Are real graphs random?
- NO!
  - Static patterns
    - Small diamete
    - Ske
  - ie-evolving

- Many power laws – log-logistic
- Take logarithms

15-826 Copyright: C. Faloutsos (2019) 143

143

Carnegie Mellon

Next lecture:

- Anomaly detection tools (OddBall, etc)

15-826 Copyright: C. Faloutsos (2019) 144

144