CMU SCS

Graph analysis: laws & tools

Christos Faloutsos
Carnegie Mellon University

CMU SCS

Overall Outline

- Laws (mainly, power laws)
- Generators and
- Tools

15-781/10-701 (c) 2008, C. Faloutsos 3

CMU SCS

Outline

- **Problem definition / Motivation**
- Static & dynamic laws; generators
- Tools: CenterPiece graphs; Tensors
- Other projects (Virus propagation, e-bay fraud detection)
- Conclusions

15-781/10-701 (c) 2008, C. Faloutsos 4

CMU SCS

Motivation

Data mining: ~ find patterns (rules, outliers)
- Problem#1: How do real graphs look like?
- Problem#2: How do they evolve?
- Problem#3: How to generate realistic graphs

TOOLS
- Problem#4: Who is the 'master-mind'?
- Problem#5: Track communities over time

CMU SCS

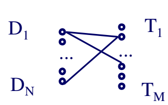
Problem#1: Joint work with

Dr. Deepayan Chakrabarti
(CMU/Yahoo R.L.)

CMU SCS

Graphs - why should we care?

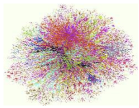
- web: hyper-text graph
- IR: bi-partite graphs (doc-terms)



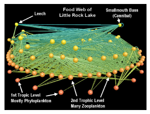
- ... and more:

CMU SCS

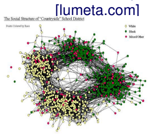
Graphs - why should we care?



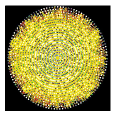
Internet Map
[lumeta.com]



Food Web
[Martinez '91]



Friendship Network
[Moody '01]



Protein Interactions
[genomebiology.com]

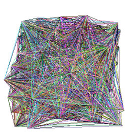
CMU SCS

Graphs - why should we care?

- network of companies & board-of-directors members
- 'viral' marketing
- web-log ('blog') news propagation
- computer network security: email/IP traffic and anomaly detection
- ....

CMU SCS

Problem #1 - network and graph mining



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

Graph mining

- Are real graphs random?

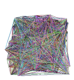
Laws and patterns

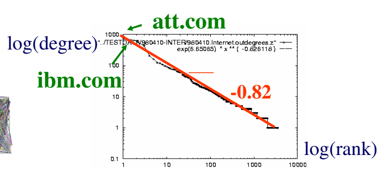
- Are real graphs random?
- A: NO!!
  - Diameter
  - in- and out- degree distributions
  - other (surprising) patterns

Solution#1

- Power law in the degree distribution [SIGCOMM99]

internet domains

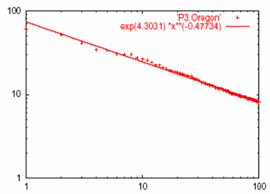




CMU SCS

Solution#1': Eigen Exponent *E*

Eigenvalue



Rank of decreasing eigenvalue

Exponent = slope

E = -0.48

May 2001

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

15-781/10-701 (c) 2008, C. Faloutsos 14

CMU SCS

But:

How about graphs from other domains?

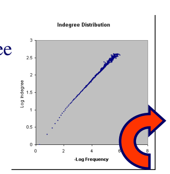
15-781/10-701 (c) 2008, C. Faloutsos 15

CMU SCS

Web

- In- and out-degree distribution of web sites [Barabasi], [IBM-CLEVER]

log indegree



from [Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins]

- log(freq)

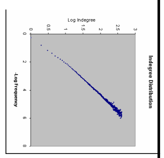
15-781/10-701 (c) 2008, C. Faloutsos 16

CMU SCS

Web

- In- and out-degree distribution of web sites [Barabasi], [IBM-CLEVER]

log(freq)

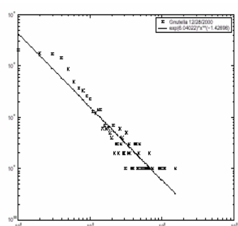


from [Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins]

log indegree

CMU SCS

The Peer-to-Peer Topology



(a) Gnutella snapshot from Dec. 28, 2000 (|r|=0.94)

[Jovanovic+]

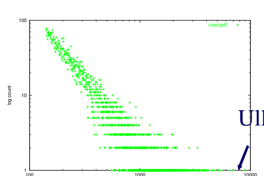
- Frequency versus degree
- Number of adjacent peers follows a power-law

CMU SCS

More power laws:

citation counts: (*citeseer.nj.nec.com* 6/2001)

log(count)



Ullman

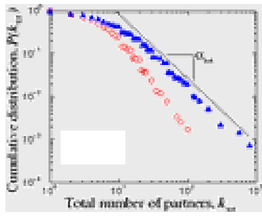
log(#citations)

CMU SCS

Swedish sex-web

Nodes: people (Females; Males)
Links: sexual relationships

Albert Laszlo Barabasi
http://www.nd.edu/~networks/Publication%20Categories/04%20Talks/2005-norway-3hours.ppt



4781 Swedes; 18-74;
59% response rate.

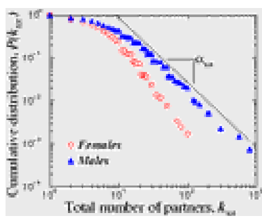
Liljeros et al. *Nature* 2001

CMU SCS

Swedish sex-web

Nodes: people (Females; Males)
Links: sexual relationships

Albert Laszlo Barabasi
http://www.nd.edu/~networks/Publication%20Categories/04%20Talks/2005-norway-3hours.ppt



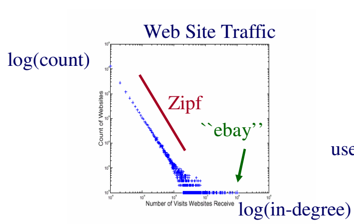
4781 Swedes; 18-74;
59% response rate.

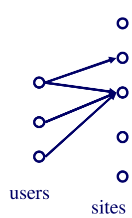
Liljeros et al. *Nature* 2001

CMU SCS

More power laws:

- web hit counts [w/ A. Montgomery]



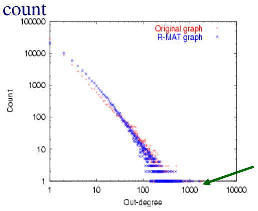


CMU SCS

epinions.com

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

count



trusts-2000-people user

(out) degree

15-781/10-701 (c) 2008, C. Faloutsos 23

CMU SCS

Outline

- Problem definition / Motivation
- Static & **dynamic** laws; generators
- Tools: CenterPiece graphs; Tensors
- Other projects (Virus propagation, e-bay fraud detection)
- Conclusions

15-781/10-701 (c) 2008, C. Faloutsos 24

CMU SCS

Motivation

Data mining: ~ find patterns (rules, outliers)

- Problem#1: How do real graphs look like?
- Problem#2: How do they evolve?
- Problem#3: How to generate realistic graphs

TOOLS

- Problem#4: Who is the 'master-mind'?
- Problem#5: Track communities over time

15-781/10-701 (c) 2008, C. Faloutsos 25

CMU SCS

Problem#2: Time evolution

- with Jure Leskovec (CMU/MLD)

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

15-781/10-701 (c) 2008, C. Faloutsos 26

CMU SCS

Evolution of the Diameter

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

15-781/10-701 (c) 2008, C. Faloutsos 27

CMU SCS

Evolution of the Diameter

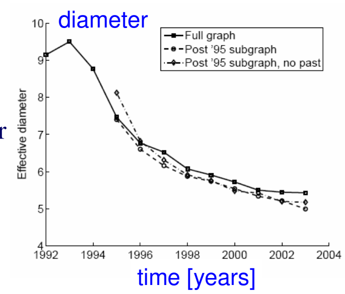
- Prior work on Power Law graphs hints at **slowly growing diameter**:
  - diameter ~ O(log N)
  - diameter ~ O(log log N)
- What is happening in real data?
- Diameter **shrinks** over time

15-781/10-701 (c) 2008, C. Faloutsos 28

CMU SCS

Diameter – ArXiv citation graph

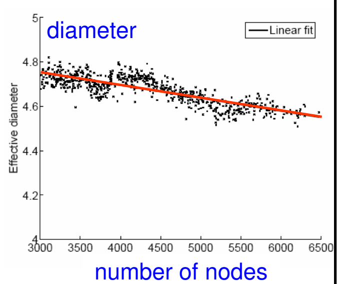
- Citations among physics papers
- 1992 –2003
- One graph per year
- 2003:
  - 29,555 papers,
  - 352,807 citations



CMU SCS

Diameter – "Autonomous Systems"

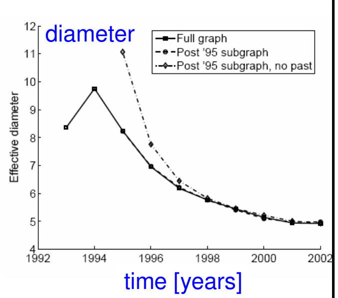
- Graph of Internet
- One graph per day
- 1997 – 2000
- 2000
  - 6,000 nodes
  - 26,000 edges



CMU SCS

Diameter – "Affiliation Network"

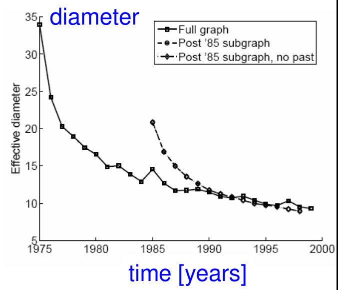
- Graph of collaborations in physics – authors linked to papers
- 10 years of data
- 2002
  - 60,000 nodes
    - 20,000 authors
    - 38,000 papers
  - 133,000 edges



CMU SCS

Diameter – "Patents"

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



15-781/10-701 (c) 2008, C. Faloutsos 32

CMU SCS

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)

15-781/10-701 (c) 2008, C. Faloutsos 33

CMU SCS

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

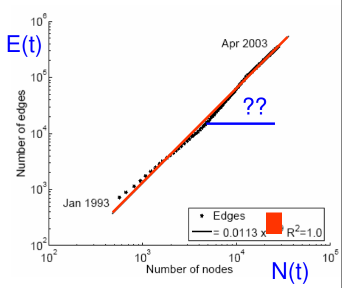
  E(t+1) =? 2 * E(t)
- A: over-doubled!
  - But obeying the ``Densification Power Law''

15-781/10-701 (c) 2008, C. Faloutsos 34

CMU SCS

Densification – Physics Citations

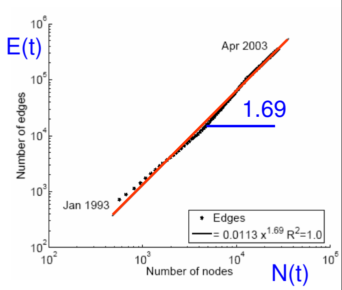
- Citations among physics papers
- 2003:
  - 29,555 papers, 352,807 citations



CMU SCS

Densification – Physics Citations

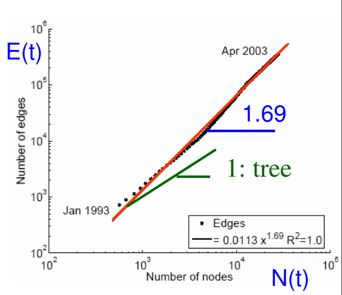
- Citations among physics papers
- 2003:
  - 29,555 papers, 352,807 citations



CMU SCS

Densification – Physics Citations

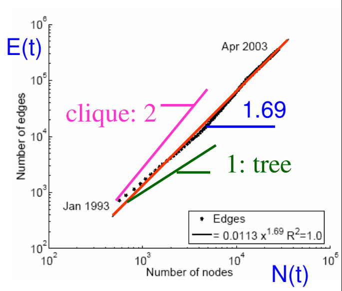
- Citations among physics papers
- 2003:
  - 29,555 papers, 352,807 citations



CMU SCS

Densification – Physics Citations

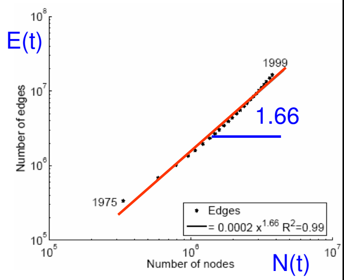
- Citations among physics papers
- 2003:
  - 29,555 papers, 352,807 citations



CMU SCS

Densification – Patent Citations

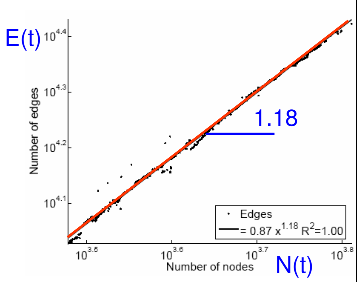
- Citations among patents granted
- 1999
  - 2.9 million nodes
  - 16.5 million edges
- Each year is a datapoint



CMU SCS

Densification – Autonomous Systems

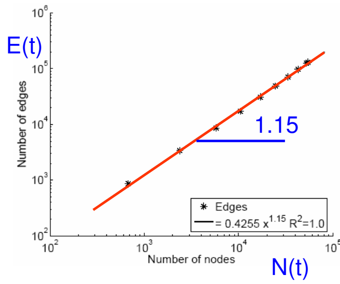
- Graph of Internet
- 2000
  - 6,000 nodes
  - 26,000 edges
- One graph per day



CMU SCS

Densification – Affiliation Network

- Authors linked to their publications
- 2002
  - 60,000 nodes
    - 20,000 authors
    - 38,000 papers
  - 133,000 edges



15-781/10-701 (c) 2008, C. Faloutsos 41

CMU SCS

Outline

- Problem definition / Motivation
- ➡ Static & dynamic laws; **generators**
- Tools: CenterPiece graphs; Tensors
- Other projects (Virus propagation, e-bay fraud detection)
- Conclusions

15-781/10-701 (c) 2008, C. Faloutsos 42

CMU SCS

Motivation

Data mining: ~ find patterns (rules, outliers)

- ✓ Problem#1: How do real graphs look like?
- ✓ Problem#2: How do they evolve?
- Problem#3: How to generate realistic graphs

TOOLS

- Problem#4: Who is the ‘master-mind’?
- Problem#5: Track communities over time

15-781/10-701 (c) 2008, C. Faloutsos 43

CMU SCS

Problem Definition

- Given a growing graph with count of nodes N_1, N_2, …
- Generate a realistic sequence of graphs that will obey all the patterns

CMU SCS

Problem Definition

- Given a growing graph with count of nodes N_1, N_2, …
- Generate a realistic sequence of graphs that will obey all the patterns
  - Static Patterns
    - Power Law Degree Distribution
    - Power Law eigenvalue and eigenvector distribution
    - Small Diameter
  - Dynamic Patterns
    - Growth Power Law
    - Shrinking/Stabilizing Diameters

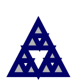
CMU SCS

Problem Definition

- Given a growing graph with count of nodes N_1, N_2, …
- Generate a realistic sequence of graphs that will obey all the patterns

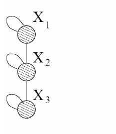
Idea: Self-similarity

- Leads to power laws
- Communities within communities
- …



CMU SCS

Kronecker Product – a Graph



1	1	0
1	1	1
0	1	1

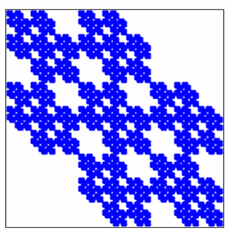
G_1

Adjacency matrix

CMU SCS

Kronecker Product – a Graph

- Continuing multiplying with G_1 we obtain G_4 and so on …

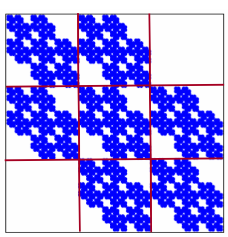


G_4 adjacency matrix

CMU SCS

Kronecker Product – a Graph

- Continuing multiplying with G_1 we obtain G_4 and so on …

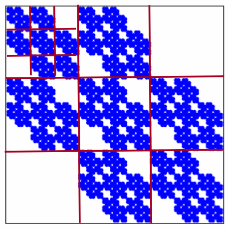


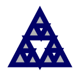
G_4 adjacency matrix

CMU SCS

Kronecker Product – a Graph

- Continuing multiplying with G_1 we obtain G_4 and so on …





G_4 adjacency matrix

15-781/10-701 (c) 2008, C. Faloutsos 51

CMU SCS

Properties:

- We can prove that
  - Degree distribution is multinomial ~ power law
  - Diameter: constant
  - Eigenvalue distribution: multinomial
  - First eigenvector: multinomial
- See [Leskovec+, PKDD'05] for proofs

15-781/10-701 (c) 2008, C. Faloutsos 56

CMU SCS

Problem Definition

- Given a growing graph with nodes N_1, N_2, …
- Generate a realistic sequence of graphs that will obey all the patterns
  - Static Patterns
    - ✔ Power Law Degree Distribution
    - ✔ Power Law eigenvalue and eigenvector distribution
    - ✔ Small Diameter
  - Dynamic Patterns
    - ✔ Growth Power Law
    - ✔ Shrinking/Stabilizing Diameters
- First and only generator for which we can **prove** all these properties

15-781/10-701 (c) 2008, C. Faloutsos 57

CMU SCS

skip

Stochastic Kronecker Graphs

- Create $N_1 \times N_1$ probability matrix P_1
- Compute the k^{th} Kronecker power P_k
- For each entry p_{uv} of P_k include an edge (u,v) with probability p_{uv}

P_1:

0.4	0.2
0.1	0.3

Kronecker multiplication →

P_k:

0.16	0.08	0.08	0.04
0.04	0.12	0.02	0.06
0.04	0.02	0.12	0.06
0.01	0.03	0.03	0.09

→ Instance Matrix G_2

flip biased coins

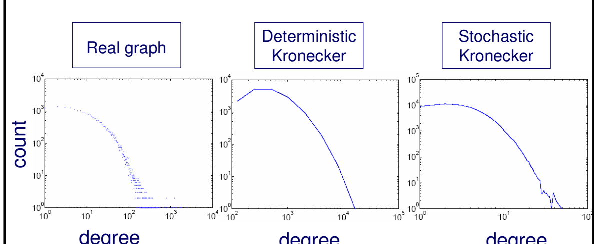
CMU SCS

Experiments

- How well can we match real graphs?
  - Arxiv: physics citations:
    - 30,000 papers, 350,000 citations
    - 10 years of data
  - U.S. Patent citation network
    - 4 million patents, 16 million citations
    - 37 years of data
  - Autonomous systems – graph of internet
    - Single snapshot from January 2002
    - 6,400 nodes, 26,000 edges
- We show both static and temporal patterns

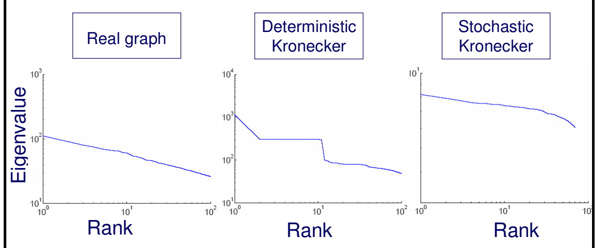
CMU SCS

Arxiv – Degree Distribution



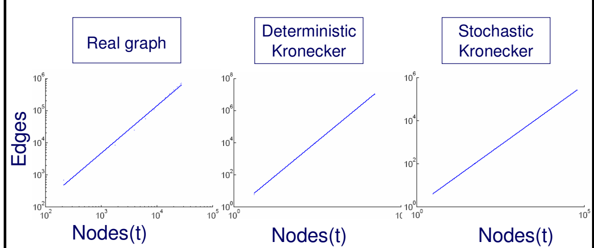
CMU SCS

Arxiv – Scree Plot



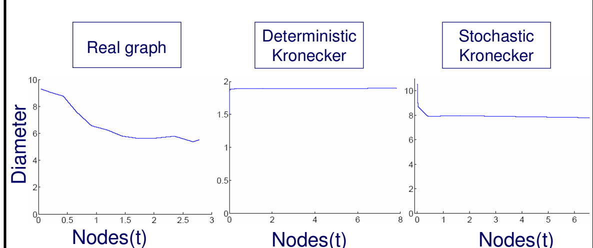
CMU SCS

Arxiv – Densification



CMU SCS

Arxiv – Effective Diameter



CMU SCS

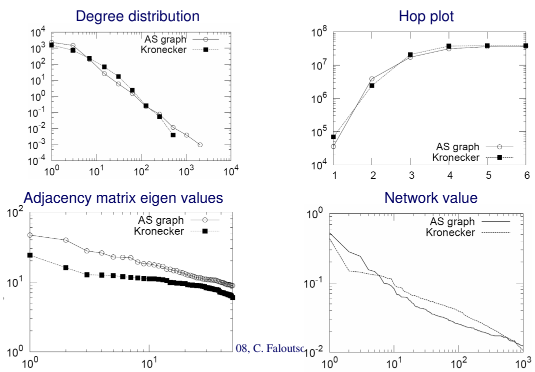
(Q: how to fit the parm's?)

A:

- Stochastic version of Kronecker graphs +
- Max likelihood +
- Metropolis sampling
- [Leskovec+, ICML'07]

CMU SCS

Experiments on real AS graph



CMU SCS

Conclusions

- Kronecker graphs have:
  - All the static properties
    - ✓Heavy tailed degree distributions
    - ✓Small diameter
    - ✓Multinomial eigenvalues and eigenvectors
  - All the temporal properties
    - ✓Densification Power Law
    - ✓Shrinking/Stabilizing Diameters
  - We can formally prove these results

CMU SCS

Outline

- Problem definition / Motivation
- Static & dynamic laws; generators
- Tools: **CenterPiece graphs**; Tensors
- Other projects (Virus propagation, e-bay fraud detection)
- Conclusions

15-781/10-701 (c) 2008, C. Faloutsos 70

CMU SCS

Motivation

Data mining: ~ find patterns (rules, outliers)

- ✓ Problem#1: How do real graphs look like?
- ✓ Problem#2: How do they evolve?
- ✓ Problem#3: How to generate realistic graphs

TOOLS

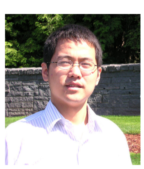
- Problem#4: Who is the 'master-mind'?
- Problem#5: Track communities over time

15-781/10-701 (c) 2008, C. Faloutsos 71

CMU SCS

Problem#4: MasterMind – 'CePS'

- w/ Hanghang Tong, KDD 2006
- htong <at> cs.cmu.edu

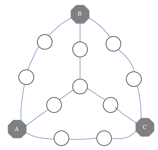


15-781/10-701 (c) 2008, C. Faloutsos 72

CMU SCS

Center-Piece Subgraph(CePS)

- **Given** Q query nodes
- **Find** Center-piece ($\leq b$)

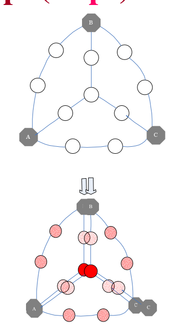


- App.
  - Social Networks
  - Law Inforcement, …
- Idea:
  - Proximity -> random walk with restarts

CMU SCS

Center-Piece Subgraph(Ceps)

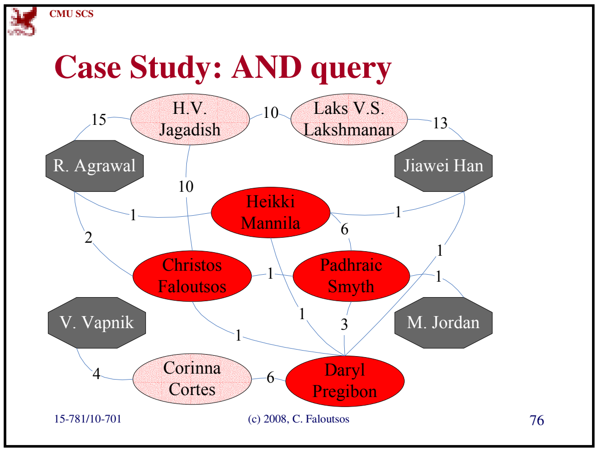
- **Given** Q query nodes
- **Find** Center-piece ($\leq b$)

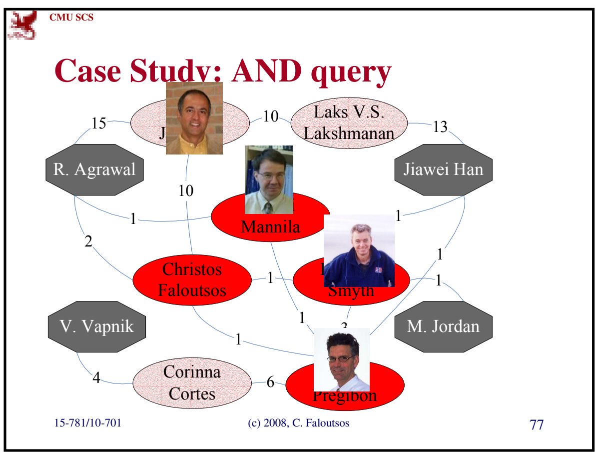


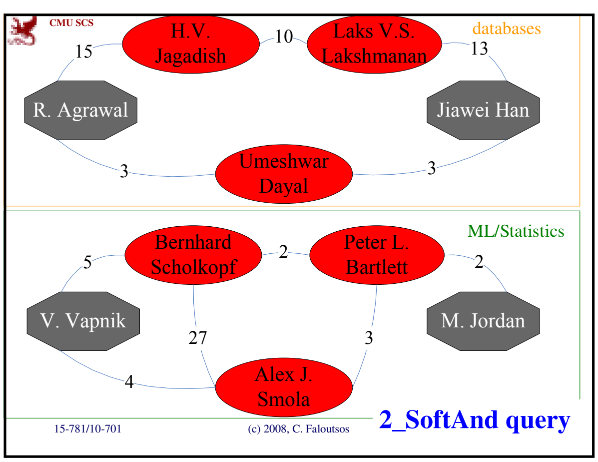
- App.
  - Social Networks
  - Law Inforcement, …
- Idea:
  - Proximity -> random walk with restarts

CMU SCS

Case Study: AND query

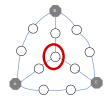

CMU SCS
Case Study: AND query
H.V. Jagadish
Laks V.S. Lakshmanan
R. Agrawal
Jiawei Han
Heikki Mannila
Christos Faloutsos
Padhraic Smyth
V. Vapnik
M. Jordan
Corinna Cortes
Daryl Pregibon
15
10
13
10
1
1
6
2
1
1
1
1
3
1
4
6
15-781/10-701
(c) 2008, C. Faloutsos
76


CMU SCS
Case Study: AND query
Laks V.S. Lakshmanan
R. Agrawal
Jiawei Han
Mannila
Christos Faloutsos
Smyth
V. Vapnik
M. Jordan
Corinna Cortes
Pregibon
15
10
13
10
1
1
2
1
1
1
1
1
4
6
15-781/10-701
(c) 2008, C. Faloutsos
77


CMU SCS
databases
H.V. Jagadish
Laks V.S. Lakshmanan
R. Agrawal
Jiawei Han
Umeshwar Dayal
15
10
13
3
3
ML/Statistics
Bernhard Scholkopf
Peter L. Bartlett
V. Vapnik
M. Jordan
Alex J. Smola
5
2
2
27
3
4
2_SoftAnd query
15-781/10-701
(c) 2008, C. Faloutsos

CMU SCS

Conclusions



- Q1:How to measure the importance?
- A1: RWR+K_SoftAnd
- Q2: How to find connection subgraph?
- A2:"Extract" Alg.
- Q3:How to do it efficiently?
- A3:Graph Partition (Fast CePS)
  - ~90% quality
  - 6:1 speedup; 150x speedup (ICDM'06, b.p. award)

15-781/10-701 (c) 2008, C. Faloutsos 79

CMU SCS

Outline

- Problem definition / Motivation
- Static & dynamic laws; generators
- ➡ Tools: CenterPiece graphs; **Tensors**
- Other projects (Virus propagation, e-bay fraud detection)
- Conclusions

15-781/10-701 (c) 2008, C. Faloutsos 80

CMU SCS

Motivation

Data mining: ~ find patterns (rules, outliers)

- ✓ Problem#1: How do real graphs look like?
- ✓ Problem#2: How do they evolve?
- ✓ Problem#3: How to generate realistic graphs

TOOLS

- ✓ Problem#4: Who is the 'master-mind'?
- Problem#5: Track communities over time

15-781/10-701 (c) 2008, C. Faloutsos 81

CMU SCS

Tensors for time evolving graphs

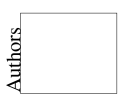
- [Jimeng Sun+ KDD'06]
- [" , SDM'07]
- [CF, Kolda, Sun, SDM'07 and SIGMOD'07 tutorial]

15-781/10-701 (c) 2008, C. Faloutsos 82

CMU SCS

Social network analysis

- Static: find community structures

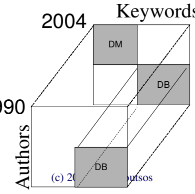


15-781/10-701 83

CMU SCS

Social network analysis

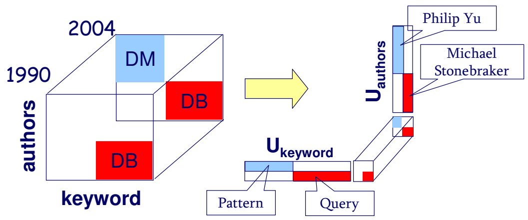
- Static: find community structures
- Dynamic: monitor community structure evolution; spot abnormal individuals; abnormal time-stamps



15-781/10-701 (c) 20...utsos 84

CMU SCS

Application 1: Multiway latent semantic indexing (LSI)



- Projection matrices specify the clusters
- Core tensors give cluster activation level

CMU SCS

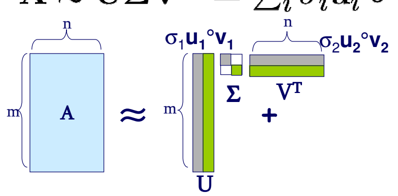
Crash course

- On SVD / spectral methods
- And tensors

CMU SCS

SVD as spectral decomposition

$$\mathbf{A} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



- Best rank-k approximation in L2 and Frobenius
- SVD only works for static matrices (a single 2nd order tensor)

See also PARAFAC

CMU SCS

SVD - Example

- $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$ - example:

$$\begin{array}{c} \\ \text{CS} \\ \\ \\ \\ \text{MD} \\ \\ \end{array}\begin{array}{c} \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-781/10-701 (c) 2008, C. Faloutsos 88

CMU SCS

SVD - Example

- $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$ - example:

CS-concept

MD-concept

$$\begin{array}{c} \\ \text{CS} \\ \\ \\ \\ \text{MD} \\ \\ \end{array}\begin{array}{c} \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-781/10-701 (c) 2008, C. Faloutsos 89

CMU SCS

SVD - Example

- $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$ - example:

doc-to-concept similarity matrix

CS-concept

MD-concept

$$\begin{array}{c} \\ \text{CS} \\ \\ \\ \\ \text{MD} \\ \\ \end{array}\begin{array}{c} \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-781/10-701 (c) 2008, C. Faloutsos 90

SVD - Example

- $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$ - example:

'strength' of CS-concept

$$\begin{array}{cc} & \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\ \begin{array}{c} \text{CS} \\ \\ \\ \\ \text{MD} \\ \\ \\ \end{array} & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

SVD - Example

- $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$ - example:

term-to-concept similarity matrix

CS-concept

$$\begin{array}{cc} & \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\ \begin{array}{c} \text{CS} \\ \\ \\ \\ \text{MD} \\ \\ \\ \end{array} & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

SVD - Example

- $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$ - example:

term-to-concept similarity matrix

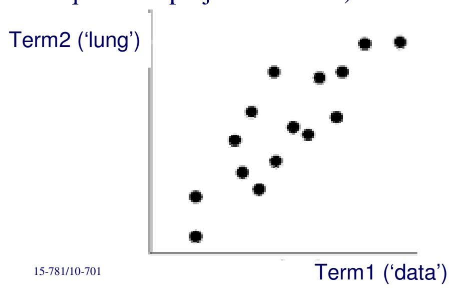
CS-concept

$$\begin{array}{cc} & \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\ \begin{array}{c} \text{CS} \\ \\ \\ \\ \text{MD} \\ \\ \\ \end{array} & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

CMU SCS

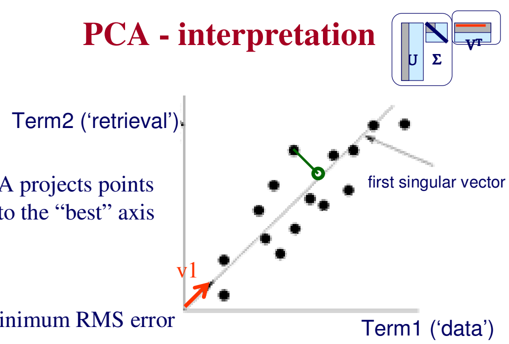
PCA interpretation

- best axis to project on: ('best' = min sum of squares of projection errors)



CMU SCS

PCA - interpretation

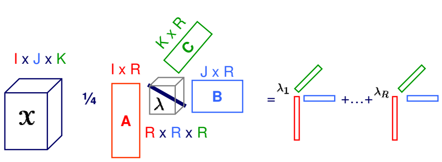


PCA projects points Onto the "best" axis

- minimum RMS error

CMU SCS

Goal: extension to >=3 modes



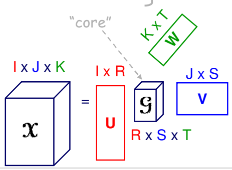
$$\mathfrak{X} \approx [\![\lambda\,;\mathbf{A},\mathbf{B},\mathbf{C}]\!] = \sum_r \lambda_r\, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

Specially Structured Tensors

- Tucker Tensor

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}$$
$$= \sum_r \sum_s \sum_t g_{rst}\, \mathbf{u}_r \circ \mathbf{v}_s \circ \mathbf{w}_t$$
$$\equiv [\![\mathcal{G} \,; \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$$

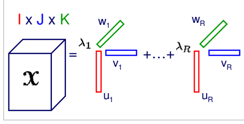
Our Notation



- Kruskal Tensor

$$\mathcal{X} = \sum_r \lambda_r\, \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r$$
$$\equiv [\![\lambda \,; \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$$

Our Notation



End of crash course

Bibliographic data (DBLP)

- Papers from VLDB and KDD conferences
- Construct 2nd order tensors with yearly windows with <author, keywords>
- Each tensor: 4584×3741
- 11 timestamps (years)

CMU SCS

Multiway LSI

Authors	Keywords	Year
michael carey, michael stonebraker, h. jagadish, hector garcia-molina	queri,parallel,optimization,concurr, objectorient	1995
surajit chaudhuri,mitch cherniack,michael stonebraker,ugur etintemel	distribut,systems,view,storage,servic,process,cache	2004
jiawei han,jian pei,philip s. yu, jianyong wang,charu c. aggarw	streams,pattern,support, cluster, gener,queri	2004

DB

DM

- Two groups are correctly identified: Databases and Data mining
- People and concepts are drifting over time

15-781/10-701 (c) 2008, C. Faloutsos 100

CMU SCS

Conclusions

Tensor-based methods:

- spot patterns and anomalies on time evolving graphs, and
- on streams (monitoring)

15-781/10-701 (c) 2008, C. Faloutsos 101

CMU SCS

Outline

- Problem definition / Motivation
- Static & dynamic laws; generators
- Tools: CenterPiece graphs; Tensors
- Other tools (**Virus propagation**, e-bay fraud detection)
- Conclusions

15-781/10-701 (c) 2008, C. Faloutsos 102

CMU SCS

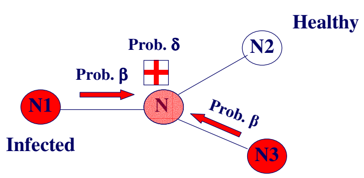
Virus propagation

- How do viruses/rumors/blog-influence propagate?
- Will a flu-like virus linger, or will it become extinct soon?

CMU SCS

The model: SIS

- 'Flu' like: Susceptible-Infected-Susceptible
- Virus 'strength' s= β/δ



CMU SCS

Epidemic threshold τ

of a graph: the value of τ, such that

if strength $s = \beta / \delta < \tau$

an epidemic can not happen

Thus,

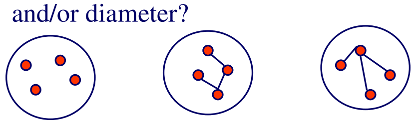
- given a graph
- compute its epidemic threshold

CMU SCS

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?



CMU SCS

Epidemic threshold

- [Theorem] We have no epidemic, if

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

CMU SCS

Epidemic threshold

- [Theorem] We have no epidemic, if

recovery prob.

epidemic threshold

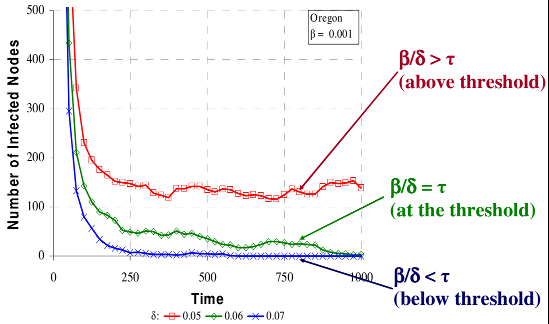
$$\beta/\delta < \tau = 1/\lambda_{1,A}$$

attack prob.

largest eigenvalue of adj. matrix ***A***

Proof: [Wang+03]

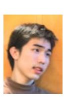
Experiments (Oregon)



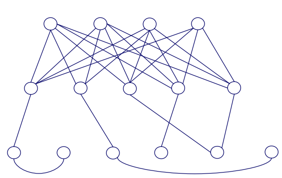
Outline

- Problem definition / Motivation
- Static & dynamic laws; generators
- Tools: CenterPiece graphs; Tensors
- Other projects (Virus propagation, **e-bay fraud detection**)
- Conclusions

E-bay Fraud detection

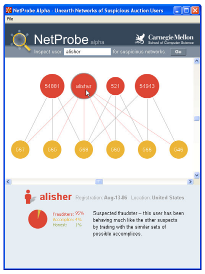


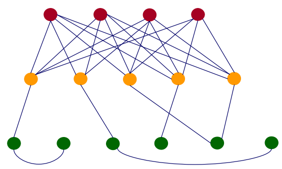

w/ Polo Chau & Shashank Pandit, CMU [WWW'07]



CMU SCS

E-bay Fraud detection - NetProbe





CMU SCS

OVERALL CONCLUSIONS

- Graphs pose a wealth of fascinating problems
- self-similarity and power laws work, when textbook methods fail!
- New patterns (shrinking diameter!)
- New generator: Kronecker

CMU SCS

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

CMU SCS

References

- Jure Leskovec, Jon Kleinberg and Christos Faloutsos *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations* KDD 2005, Chicago, IL. ("Best Research Paper" award).
- Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos *Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication* (ECML/PKDD 2005), Porto, Portugal, 2005.

15-781/10-701 (c) 2008, C. Faloutsos 115

CMU SCS

References

- Jure Leskovec and Christos Faloutsos, *Scalable Modeling of Real Graphs using Kronecker Multiplication*, ICML 2007, Corvallis, OR, USA
- Jimeng Sun, Dacheng Tao, Christos Faloutsos *Beyond Streams and Graphs: Dynamic Tensor Analysis,* KDD 2006, Philadelphia, PA

15-781/10-701 (c) 2008, C. Faloutsos 116

CMU SCS

References

- Jimeng Sun, Yinglian Xie, Hui Zhang, Christos Faloutsos. *Less is More: Compact Matrix Decomposition for Large Sparse Graphs*, SDM, Minneapolis, Minnesota, Apr 2007. [pdf]

15-781/10-701 (c) 2008, C. Faloutsos 117

Thank you!

- Christos Faloutsos
  www.cs.cmu.edu/~christos
  Wean Hall 7107

For more info on tensors:
www.cs.cmu.edu/~christos/TALKS/SIGMOD-07-tutorial/
3h version: www.cs.cmu.edu/~christos/TALKS/SDM-tut-07/