

CMU SCS

15-826: Multimedia Databases and Data Mining

Lecture #27: Graph mining -
Communities and a paradox

Christos Faloutsos

15-826 Copyright: C. Faloutsos (2012) #1

CMU SCS

Must-read material

Fully Automatic Cross-Associations,
by D. Chakrabarti, S. Papadimitriou, D.
Modha and C. Faloutsos, in KDD 2004
(pages 79-88), Washington, USA

15-826 Copyright: C. Faloutsos (2012) #2

CMU SCS

Detailed outline


- ➡ Motivation
 - Hard clustering – k pieces
 - Hard co-clustering – (k,l) pieces
 - Hard clustering – optimal # pieces
 - Observations

15-826 Copyright: C. Faloutsos (2012) -3

CMU SCS

Problem

- Given a graph, and k
- Break it into k (disjoint) communities

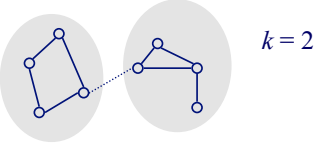


15-826 Copyright: C. Faloutsos (2012) -4

CMU SCS

Problem

- Given a graph, and k
- Break it into k (disjoint) communities



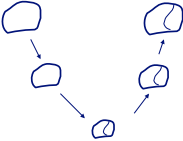
$k = 2$

15-826 Copyright: C. Faloutsos (2012) -5

CMU SCS

Solution #1: METIS

- Arguably, the best algorithm
- Open source, at
 - <http://www.cs.umn.edu/~metis>
- and *many* related papers, at same url
- Main idea:
 - coarsen the graph;
 - partition;
 - un-coarsen





15-826 Copyright: C. Faloutsos (2012) -6

CMU SCS

Solution #1: METIS

- G. Karypis and V. Kumar. *METIS 4.0: Unstructured graph partitioning and sparse matrix ordering system*. TR, Dept. of CS, Univ. of Minnesota, 1998.
- <and many extensions>

15-826 Copyright: C. Faloutsos (2012) -7

CMU SCS

Solution #2

(problem: hard clustering, k pieces)

Spectral partitioning:

- Consider the 2nd smallest eigenvector of the (normalized) Laplacian

15-826 Copyright: C. Faloutsos (2012) -8

CMU SCS

Solutions #3, ...

Many more ideas:

- Clustering on the A^2 (square of adjacency matrix) [Zhou, Woodruff, PODS'04]
- Minimum cut / maximum flow [Flake+, KDD'00]
- ...

15-826 Copyright: C. Faloutsos (2012) -9

CMU SCS

Detailed outline

- Motivation
- Hard clustering – k pieces
- ➔ • Hard co-clustering – (k, l) pieces
- Hard clustering – optimal # pieces
- Soft clustering – matrix decompositions
- Observations

15-826 Copyright: C. Faloutsos (2012) -10

CMU SCS

Problem definition

- Given a bi-partite graph, and k, l
- Divide it into k row groups and l row groups
- (Also applicable to uni-partite graph)

15-826 Copyright: C. Faloutsos (2012) -11

CMU SCS

Co-clustering

- Given data matrix and the number of row and column groups k and l
- Simultaneously
 - Cluster rows into k disjoint groups
 - Cluster columns into l disjoint groups

15-826 Copyright: C. Faloutsos (2012) -12

CMU SCS details

Co-clustering

- Let X and Y be discrete random variables
 - X and Y take values in $\{1, 2, \dots, m\}$ and $\{1, 2, \dots, n\}$
 - $p(X, Y)$ denotes the joint probability distribution—if not known, it is often estimated based on co-occurrence data
 - Application areas: text mining, market-basket analysis, analysis of browsing behavior, etc.
- Key Obstacles in Clustering Contingency Tables
 - High Dimensionality, Sparsity, Noise
 - Need for robust and scalable algorithms

Reference:
1. Dhillon et al. Information-Theoretic Co-clustering, KDD'03

CMU SCS

$$\begin{matrix} k \\ m \end{matrix}
 \begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix}
 \begin{matrix} l \\ k \end{matrix}
 \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ 2 & .2 \end{bmatrix}
 \begin{matrix} n \\ l \\ n \end{matrix}
 \begin{bmatrix} .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & .28 & .36 & .36 \end{bmatrix}
 =
 \begin{matrix} n \\ n \end{matrix}
 \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}
 \begin{matrix} n \\ n \end{matrix}
 \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & .028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}$$

eg, terms x documents

15-826 Copyright: C. Faloutsos (2012) -14

CMU SCS

$$\begin{matrix} k \\ m \end{matrix}
 \begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix}
 \begin{matrix} l \\ k \end{matrix}
 \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ 2 & .2 \end{bmatrix}
 \begin{matrix} n \\ l \\ n \end{matrix}
 \begin{bmatrix} .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & .28 & .36 & .36 \end{bmatrix}
 =
 \begin{matrix} n \\ n \end{matrix}
 \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}
 \begin{matrix} n \\ n \end{matrix}
 \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & .028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}$$

med. doc | med. terms
cs doc | cs terms
term group x doc. group | common terms

doc x doc group

term x term-group

Copyright: C. Faloutsos (2012) -15

CMU SCS

Co-clustering

Observations

- uses KL divergence, instead of L2
- the middle matrix is **not** diagonal
 - we'll see that again in the Tucker tensor decomposition
- s/w at:
www.cs.utexas.edu/users/dml/Software/cocluster.html

15-826 Copyright: C. Faloutsos (2012) -16

CMU SCS

Detailed outline

- Motivation
- Hard clustering – k pieces
- Hard co-clustering – (k,l) pieces
- ➔ • Hard clustering – optimal # pieces
- Soft clustering – matrix decompositions
- Observations

15-826 Copyright: C. Faloutsos (2012) -17

CMU SCS

Problem with Information Theoretic Co-clustering

- Number of row and column groups must be specified

Desiderata:

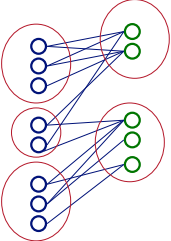
- ✓ Simultaneously discover row and column groups
- ✗ Fully Automatic: No “magic numbers”
- ✓ Scalable to large graphs

15-826 Copyright: C. Faloutsos (2012) -18

CMU SCS

Graph partitioning

- Documents x terms
- Customers x products
- Users x web-sites

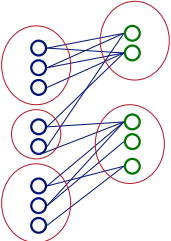


15-826 Copyright: C. Faloutsos (2012) #19

CMU SCS

Graph partitioning

- Documents x terms
- Customers x products
- Users x web-sites



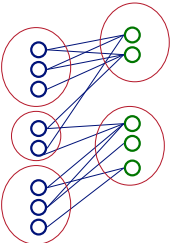
- Q: HOW MANY PIECES?

15-826 Copyright: C. Faloutsos (2012) #20

CMU SCS

Graph partitioning

- Documents x terms
- Customers x products
- Users x web-sites



- Q: HOW MANY PIECES?
- A: MDL/ compression

15-826 Copyright: C. Faloutsos (2012) #21

CMU SCS

Cross-association

Desiderata:

- ✓ Simultaneously discover row and column groups
- ✓ Fully Automatic: No “magic numbers”
- ✓ Scalable to large matrices

Reference:

- Chakrabarti et al. Fully Automatic Cross-Associations, KDD'04

CMU SCS

What makes a cross-association “good”?

versus

Why is this better?

15-826 Copyright: C. Faloutsos (2012) -23

CMU SCS

What makes a cross-association “good”?

versus

Why is this better?

simpler; easier to describe
easier to compress!

15-826 Copyright: C. Faloutsos (2012) -24

CMU SCS

What makes a cross-association “good”?

Problem definition: given an encoding scheme

- decide on the # of col. and row groups k and l
- and reorder rows and columns,
- to achieve best compression

15-826 Copyright: C. Faloutsos (2012) -25

CMU SCS

details

Main Idea

Good Compression

→

Better Clustering

Total Encoding Cost = $\sum_i \text{size}_i * H(x_i)$ + Cost of describing cross-associations

Code Cost
Description Cost

Minimize the total cost (# bits)
for lossless compression

15-826 Copyright: C. Faloutsos (2012) -26

CMU SCS

Algorithm

Iteration 1: $k=1, l=2$

Iteration 2: $k=2, l=2$

Iteration 3: $k=2, l=3$

Iteration 4: $k=3, l=3$

Iteration 5: $k=3, l=4$

Iteration 6: $k=4, l=4$

Iteration 7: $k=4, l=5$

15-826 Copyright: C. Faloutsos (2012) -27

CMU SCS

Experiments

Documents

Words

“CLASSIC”

- 3,893 documents
- 4,303 words
- 176,347 “dots”

Combination of 3 sources:

- MEDLINE (medical)
- CISI (info. retrieval)
- CRANFIELD (aerodynamics)

15-826 Copyright: C. Faloutsos (2012) 28

CMU SCS

Experiments

Documents

Words

“CLASSIC” graph of documents & words:
k=15, l=19

29

CMU SCS

Experiments

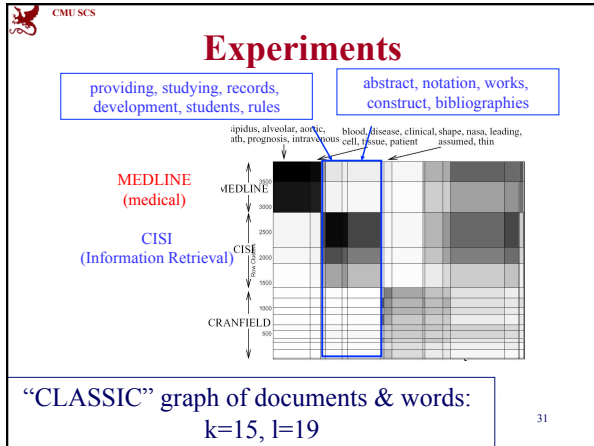
insipidus, alveolar, aortic, death, prognosis, intravenous

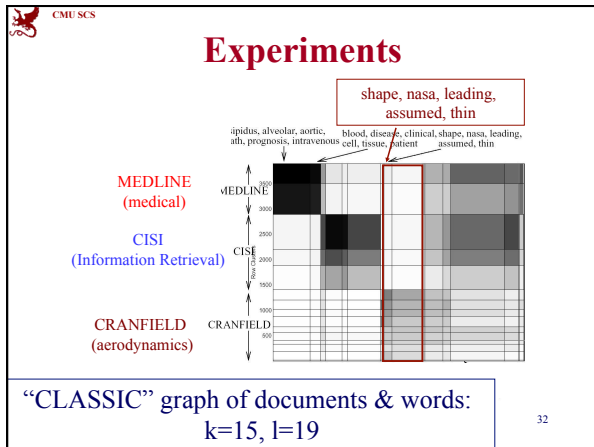
blood, disease, clinical, cell, tissue, patient

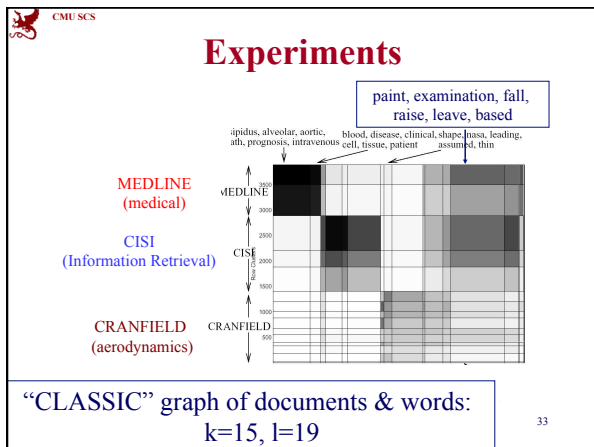
MEDLINE
(medical)

“CLASSIC” graph of documents & words:
k=15, l=19

30







CMU SCS


Algorithm

Code for cross-associations (matlab):

www.cs.cmu.edu/~deepay/mywww/software/CrossAssociations-01-27-2005.tgz

Variations and extensions:

- ‘Autopart’ [Chakrabarti, PKDD’04]
- www.cs.cmu.edu/~deepay





15-826 Copyright: C. Faloutsos (2012) 34

CMU SCS

Algorithm

- Hadoop implementation [ICDM’08]

Spiros Papadimitriou, Jimeng Sun: DisCo: Distributed Co-clustering with Map-Reduce: A Case Study towards Petabyte-Scale End-to-End Mining. ICDM 2008: 512-521

15-826 Copyright: C. Faloutsos (2012) 36

CMU SCS

Detailed outline

- Motivation
- Hard clustering – k pieces
- Hard co-clustering – (k, l) pieces
- Hard clustering – optimal # pieces
- ➔ • (Soft clustering – matrix decompositions
 - PCA, ICA, non-negative matrix factorization, ...)
- Observations

15-826 Copyright: C. Faloutsos (2012) 36

CMU SCS

Detailed outline


- Motivation
- Hard clustering – k pieces
- Hard co-clustering – (k, l) pieces
- Hard clustering – optimal # pieces
- (Soft clustering)
- ➡ • Observations

15-826 Copyright: C. Faloutsos (2012) 37

CMU SCS

Observation #1

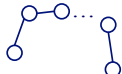
- Skewed degree distributions – there are nodes with huge degree ($>O(10^4)$, in facebook/linkedin popularity contests!)
- TRAP: ‘find all pairs of nodes, within 2 steps from each other’



IM
15-826 Copyright: C. Faloutsos (2012) 38

CMU SCS

Observation #2



- TRAP: *shortest-path between two nodes*
- (cheat: look for 2, at most 3-step paths)
- Why:
 - If they are close (within 2-3 steps): solved
 - If not, after ~6 steps, you’ll have ~ the whole graph, and the path won’t be very meaningful, anyway.

15-826 Copyright: C. Faloutsos (2012) 39

CMU SCS

Observation #3

- Maybe there are no good cuts: ``jellyfish'' shape [Tauro+'01], [Siganos+', '06], strange behavior of cuts [Chakrabarti+'04], [Leskovec+', '08]

15-826 Copyright: C. Faloutsos (2012) 40

CMU SCS

Observation #3

- Maybe there are no good cuts: ``jellyfish'' shape [Tauro+'01], [Siganos+', '06], strange behavior of cuts [Chakrabarti+'04], [Leskovec+', '08]

15-826 Copyright: C. Faloutsos (2012) 41

CMU SCS

Jellyfish model [Tauro+]

A Simple Conceptual Model for the Internet Topology, L. Tauro, C. Palmer, G. Siganos, M. Faloutsos, Global Internet, November 25-29, 2001

Jellyfish: A Conceptual Model for the AS Internet Topology G. Siganos, Sudhir L Tauro, M. Faloutsos, J. of Communications and Networks, Vol. 8, No. 3, pp 339-350, Sept. 2006.

CMU SCS

Strange behavior of min cuts

- ‘negative dimensionality’ (!)

NetMine: New Mining Tools for Large Graphs, by D. Chakrabarti, Y. Zhan, D. Blandford, C. Faloutsos and G. Blleloch, in the SDM 2004 Workshop on Link Analysis, Counter-terrorism and Privacy

Statistical Properties of Community Structure in Large Social and Information Networks, J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. WWW 2008.

CMU SCS

“Min-cut” plot

- Do min-cuts recursively.

Mincut size = \sqrt{N}

N nodes

15-826 Copyright: C. Faloutsos (2012) 44

CMU SCS

“Min-cut” plot

- Do min-cuts recursively.

New min-cut

N nodes

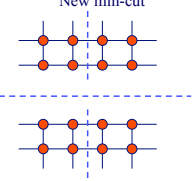
15-826 Copyright: C. Faloutsos (2012) 45

CMU SCS

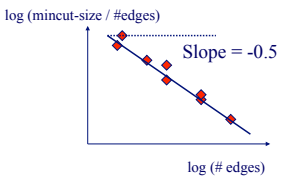
“Min-cut” plot

- Do min-cuts recursively.

New min-cut



N nodes



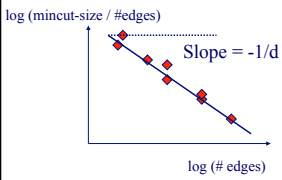
Slope = -0.5

For a d-dimensional grid, the slope is $-1/d$

15-826 Copyright: C. Faloutsos (2012) 46

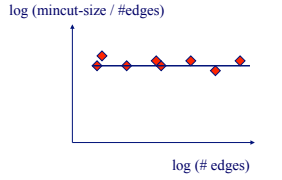
CMU SCS

“Min-cut” plot



Slope = $-1/d$

For a d-dimensional grid, the slope is $-1/d$



Slope = 0

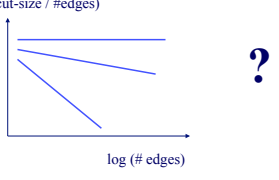
For a random graph, the slope is 0

15-826 Copyright: C. Faloutsos (2012) 47

CMU SCS

“Min-cut” plot

- What does it look like for a real-world graph?



15-826 Copyright: C. Faloutsos (2012) 48

CMU SCS

Experiments

- Datasets:
 - **Google Web Graph**: 916,428 nodes and 5,105,039 edges
 - **Lucent Router Graph**: Undirected graph of network routers from www.isi.edu/scan/mercator/maps.html; 112,969 nodes and 181,639 edges
 - **User → Website Clickstream Graph**: 222,704 nodes and 952,580 edges

NetMine: New Mining Tools for Large Graphs, by D. Chakrabarti, Y. Zhan, D. Blandford, C. Faloutsos and G. Blelloch, in the SDM 2004 Workshop on Link Analysis, Counter-terrorism and Privacy

CMU SCS

Experiments

- Used the METIS algorithm [Karypis, Kumar, 1995]

- Google Web graph
- Values along the y-axis are averaged
- We observe a “lip” for large edges
- Slope of -0.4, corresponds to a 2.5-dimensional grid!

15-826 Copyright: C. Faloutsos (2012) 50

CMU SCS

Experiments

- Used the METIS algorithm [Karypis, Kumar, 1995]


- Similarly, for
 - Lucent routers
 - clickstream

15-826 Copyright: C. Faloutsos (2012) 51

CMU SCS

Conclusions – Practitioner’s guide

- Hard clustering – k pieces **METIS**
- Hard co-clustering – (k, l) pieces **Co-clustering**
- Hard clustering – optimal # pieces **Cross-associations**
- Observations **‘jellyfish’:**
Maybe, there are no good cuts



15-826 Copyright: C. Faloutsos (2012) 52

CMU SCS

Overall conclusions

- Patterns in real graphs
 - Small, and shrinking **diameter**
 - Skewed **degree** distribution – power-law, log-normal, log-logistic
 - Super-linearities (power-laws)
 - Densification; fortification
 - Triangle law
 - Oscillating connected component sizes

15-826 Copyright: C. Faloutsos (2012) #53

CMU SCS

Overall conclusions

- Patterns in real graphs
 - Small, and shrinking **diameter**

**CAREFUL what algo to use;
what question to ask;
and what results to expect.**

- Triangle law
- Oscillating connected component sizes

15-826 Copyright: C. Faloutsos (2012) #54
