Carnegie Mellon

15-826: Multimedia Databases and Data Mining

Lecture #31: Conclusions

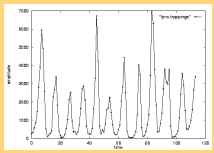
C. Faloutsos

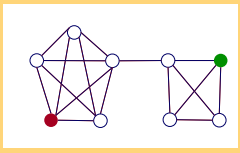
1

Carnegie Mellon

Problem

- Given a large dataset (points; text doc's; time series; images; nodes in a graph)
- Find similar/interesting things





2

Carnegie Mellon

Summary

- **T1: fractals / power laws** lead to startling discoveries
  - 'the mean may be meaningless'
  - Don't assume Gaussian (average, k-means, etc)
- **T2: SVD**: behind PageRank/HITS/tensors/…
- **T3: Wavelets**: Nature seems to prefer them
- **T4: RLS**: matrix inversion, without inverting

3

Carnegie Mellon

Outline

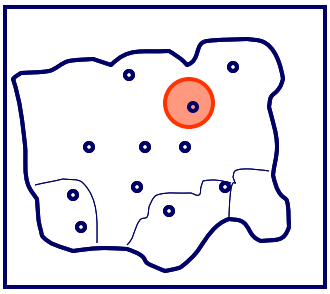
Goal: 'Find similar / interesting things'

- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - Time sequences; images etc
  - Graphs

4

Carnegie Mellon

Indexing - similarity search

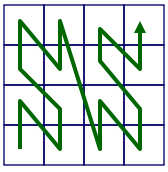


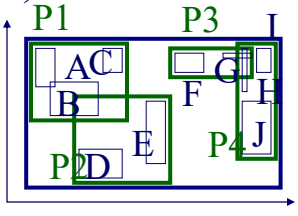
5

Carnegie Mellon

Indexing - similarity search

- R-trees
- z-ordering / hilbert curves
- M-trees
- (DON'T FORGET …)



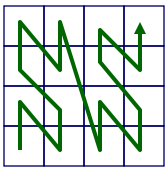


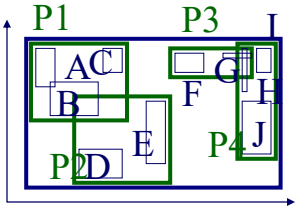
6

Carnegie Mellon

Indexing - similarity search

- R-trees
- z-ordering / hilbert curves
- M-trees
- **beware of high intrinsic dimensionality**





7

Carnegie Mellon

Outline

Goal: 'Find similar / interesting things'

- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - Time sequences; images etc
  - Graphs

8

Carnegie Mellon

Text searching

- 'find all documents with word *bla*'

9

Carnegie Mellon

Text searching

- Full text scanning ('grep')
- Inversion (B-tree or hash index)
- signature files – Bloom filters
- Vector space model
  - Ranked output
  - Relevance feedback
- String editing distance (-> dynamic prog.)

10

Carnegie Mellon

Outline

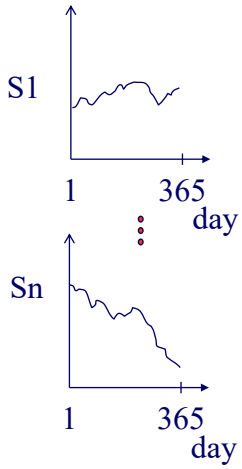
Goal: 'Find similar / interesting things'

- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - Time sequences; images etc
  - Graphs

11

Carnegie Mellon

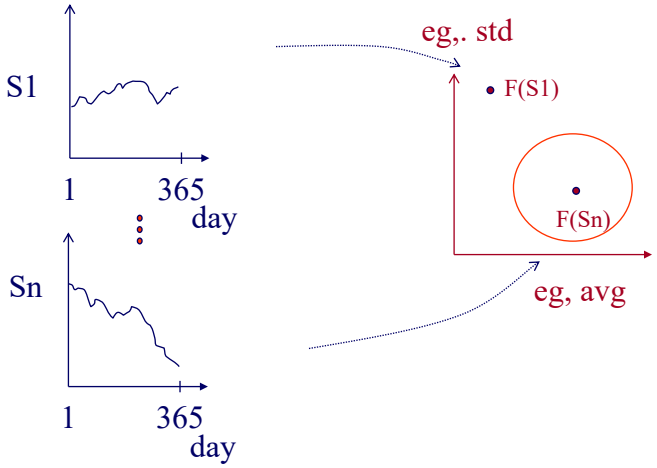
Multimedia indexing



12

Carnegie Mellon

'GEMINI' - Pictorially

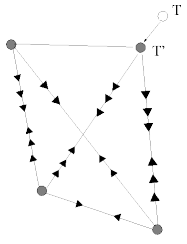


13

Carnegie Mellon

Multimedia indexing

- Feature extraction for indexing (GEMINI)
  - Lower-bounding lemma, to guarantee no false alarms
- MDS/FastMap



14

Carnegie Mellon

Outline

Goal: 'Find similar / interesting things'

- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - Time sequences; images etc – **DFT/DWT**
  - Graphs

15

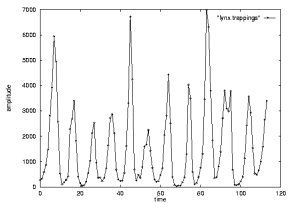
Carnegie Mellon

Time series & forecasting

Goal: given a signal (eg., sales over time and/or space)

Find: patterns and/or compress

count



year

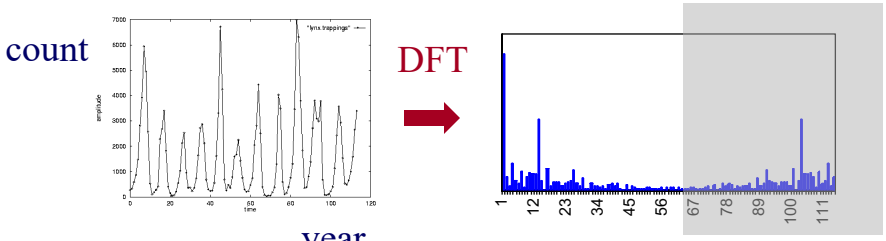
16

Carnegie Mellon

Time series & forecasting

Goal: given a signal (eg., sales over time and/or space)

Find: patterns and/or compress

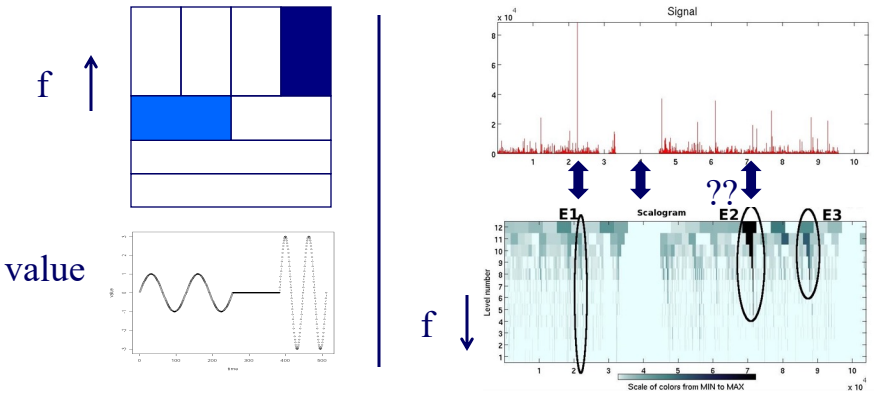


17

Carnegie Mellon

Wavelets

- Q: baritone/silence/soprano - DWT?

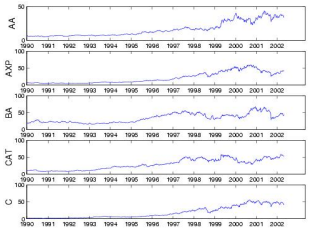


18

Carnegie Mellon

Problem:

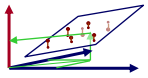
Q: mine/forecast (one, or more) time sequences

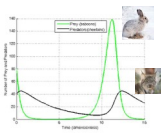


Carnegie Mellon

Answers

- Similarity search: **Euclidean**/time-warping; **feature extraction** and **SAMs**
- Linear Forecasting: **AR** (Box-Jenkins)
- Non-linear forecasting: **lag-plots**
- Gray-box modeling: **Lotka-Volterra**





Carnegie Mellon

Outline

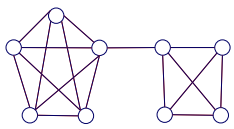
Goal: 'Find similar / interesting things'

- Intro to DB
- Indexing - similarity search
  - – Points
  - – Text
  - – Time sequences; images etc
  - – Graphs

Carnegie Mellon

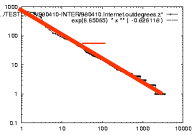
Graphs

- Real graphs: surprising patterns
  - – ??

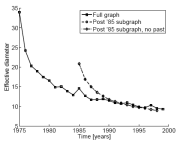


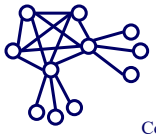
Carnegie Mellon

Graphs



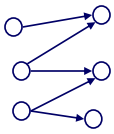
- Real graphs: surprising patterns
  - '**six degrees**'
  - **Skewed** degree distribution ('rich get richer')
  - Super-linearities (2x nodes -> 3x edges)
  - Diameter: **shrinks** (!)
  - Might have **no** good cuts



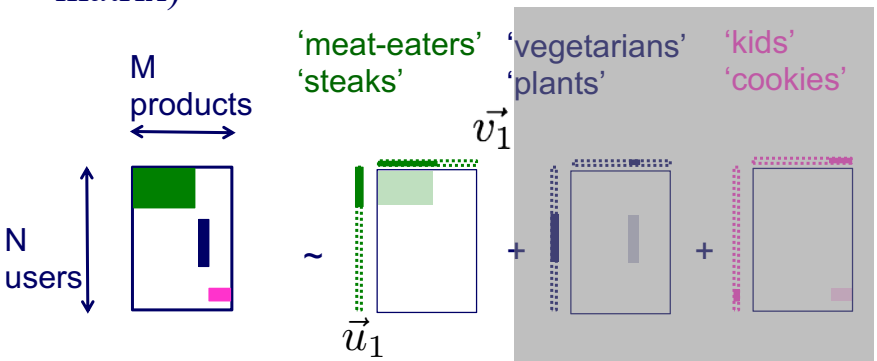


Carnegie Mellon

Graphs - SVD

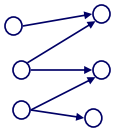


- Hubs/Authorities (SVD on adjacency matrix)

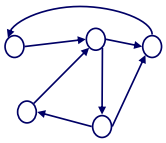


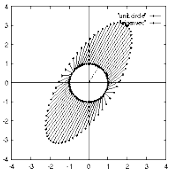
Carnegie Mellon

Graphs - PageRank



- Hubs/Authorities (SVD on adjacency matrix)
- PageRank (fixed point -> eigenvector)

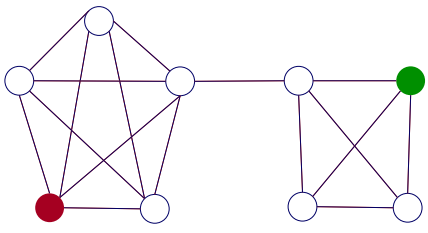




Carnegie Mellon

Belief Propagation

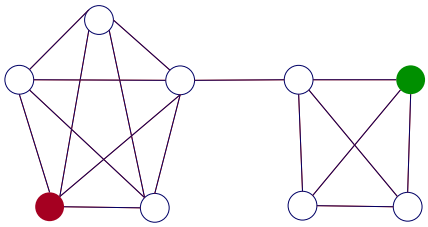
- What color, for the rest?



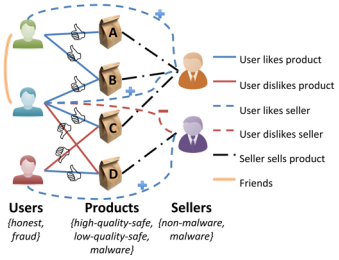
Carnegie Mellon

Belief Propagation

- What color, for the rest?
- A: Belief Propagation ('zooBP')



www.cs.cmu.edu/~deswaran/code/zoobp.zip

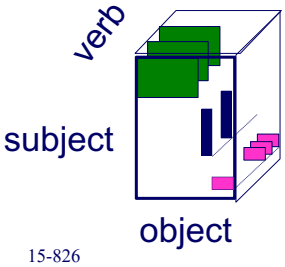


27

Carnegie Mellon

Tensors

- Eg., time evolving graphs; Subject-verb-object triplets; etc

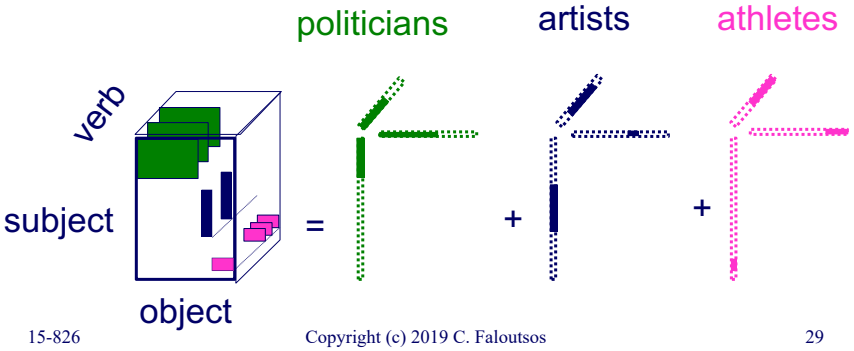


28

Carnegie Mellon

Tensors

- Eg., time evolving graphs; Subject-verb-object triplets; etc



29

Carnegie Mellon

Association Rules

- Given many market baskets
- Find which products sell together

30

Carnegie Mellon

Association Rules

- Given many market baskets
- Find which products sell together

- Association rules ['large itemsets']
  - {Milk, bread} -> butter [milk,bread,butter: often]

Carnegie Mellon

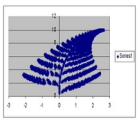
Taking a step back:

We saw some fundamental, recurring concepts and tools:

Carnegie Mellon

T1: Powerful, recurring tools

- Fractals/ self similarity

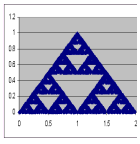


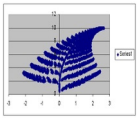
33

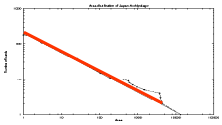
Carnegie Mellon

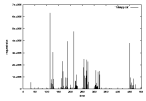
T1: Powerful, recurring tools

- Fractals/ self similarity <-> Power laws
  - Zipf, Korcak, Pareto's laws
  - intrinsic dimension (Sierpinski triangle)
  - correlation integral
  - Barnsley's IFS compression
  - Kronecker graphs

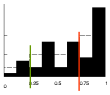








80/20

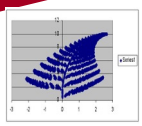


34

Carnegie Mellon

T1: Powerful, recurring tools

- Fractals/ self similarity
  - Zipf, Ko
- ‘Take logarithms’
- mean-> meaningless
- Gaussian trap
- pression
- ecker graphs)

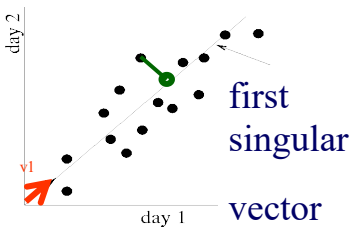


35

Carnegie Mellon

T2: Powerful, recurring tools

- SVD (optimal L2 approx)

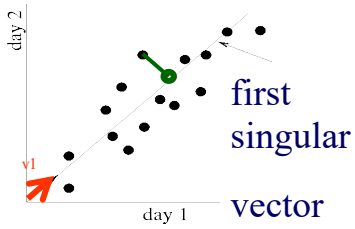


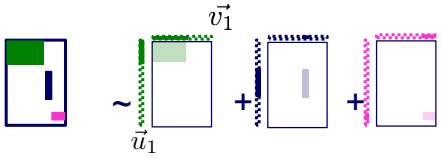
36

Carnegie Mellon

T2: Powerful, recurring tools

- SVD (optimal L2 approx)
  - LSI, KL, PCA, 'eigenSpokes', (& in ICA)
  - HITS (PageRank)





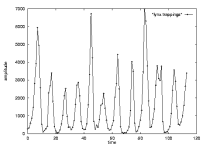
Carnegie Mellon

T3: powerful, recurring tools

DFT (Discrete Fourier Transform)

DWT (Discrete Wavelet Transform)

count

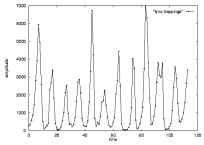


year

Carnegie Mellon

T3: powerful, recurring tools

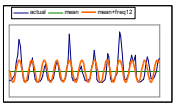
DFT (Discrete Fourier Transform)
DWT (Discrete Wavelet Transform)



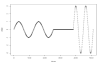
count

year

A1: Fourier (DFT)



A2: Wavelets (DWT)

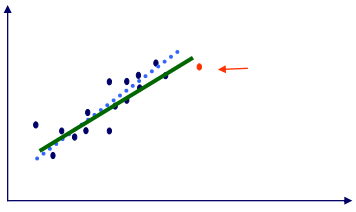


39

Carnegie Mellon

T4: Powerful, recurring tools

- Matrix inversion lemma
  - Recursive Least Squares
  - Sherman-Morrison(-Woodbury)



40

Carnegie Mellon

Summary of summary

- **T1: fractals / power laws** lead to startling discoveries
  - 'the mean may be meaningless'
  - Don't assume Gaussian (average, k-means, etc)
- **T2: SVD**: behind PageRank/HITS/tensors/…
- **T3: Wavelets**: Nature seems to prefer them
- **T4: RLS**: matrix inversion, without inverting

41

Carnegie Mellon

Summary of summary

- **T1: fractals / power laws** lead to st g discoveries
  - 'the me
  - ans, etc)
- T ankHITS/tensors/…
- T **avelets**: Nature seems to prefer them
- **T4: RLS**: matrix inversion, without inverting

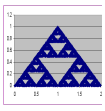
'Take logarithms'
- **mean -> meaningless**
- **Gaussian trap**

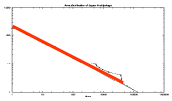
42

Carnegie Mellon

Thank you!

- Feel free to contact me:
  - Cell#; christos@cs; GHC 8019
- Reminder: faculty course eval's:
  - www.cmu.edu/hub/fce/
- Final: as announced in Hub
  - www.cmu.edu/hub/docs/final-exams.pdf
- Have a great holiday!





- 'Take logarithms'
- mean -> meaningless
- Gaussian trap