

CarnegieMellon

# 15-826: Multimedia Databases and Data Mining

Lecture #31: Conclusions  
*C. Faloutsos*

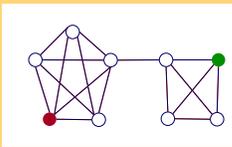
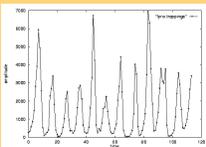
1

CarnegieMellon

## Problem



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



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

2

CarnegieMellon



## 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

15-826 Copyright (c) 2019 C. Faloutsos 3

3

CarnegieMellon

## Outline

Goal: ‘Find **similar / interesting** things’

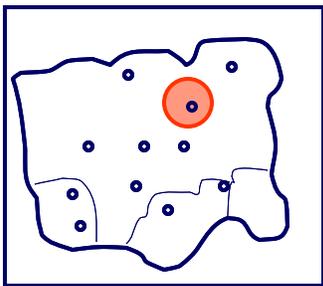
- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - Time sequences; images etc
  - Graphs

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

4

CarnegieMellon

## Indexing - similarity search



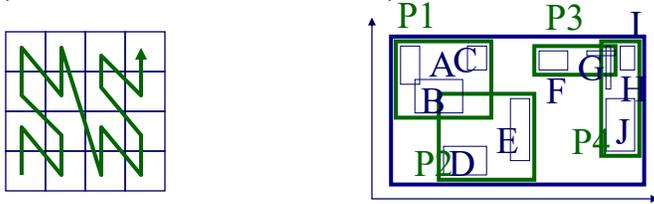
15-826 Copyright (c) 2019 C. Faloutsos 5

5

CarnegieMellon

## Indexing - similarity search

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



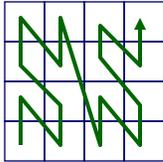
15-826 Copyright (c) 2019 C. Faloutsos 6

6

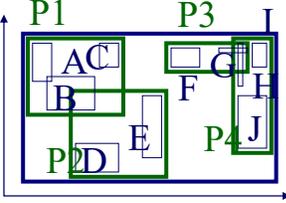
CarnegieMellon

## Indexing - similarity search

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



15-826



Copyright (c) 2019 C. Faloutsos

7

7

CarnegieMellon

## Outline

Goal: 'Find **similar / interesting things**'

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

15-826

Copyright (c) 2019 C. Faloutsos

8

8

CarnegieMellon

## Text searching

- ‘find all documents with word *bla*’



15-826 Copyright (c) 2019 C. Faloutsos 9

9

CarnegieMellon

## 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.)

15-826 Copyright (c) 2019 C. Faloutsos 10

10

CarnegieMellon

## Outline

Goal: 'Find similar / interesting things'

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

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

11

CarnegieMellon

## Multimedia indexing

S1

1 365 day

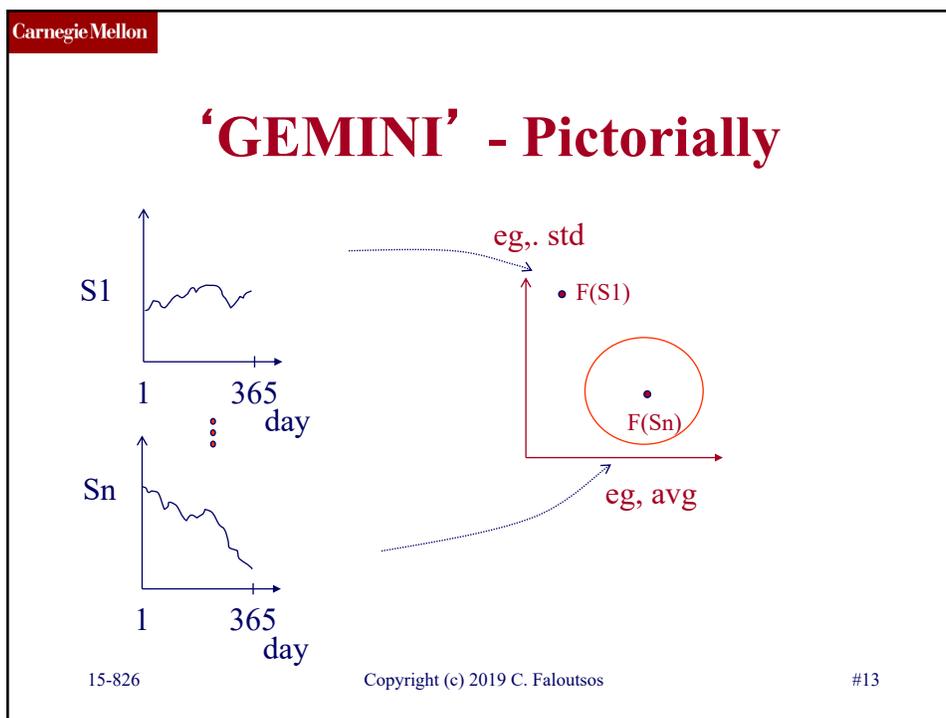
⋮

Sn

1 365 day

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

12



13

CarnegieMellon

## Multimedia indexing

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

The diagram illustrates MDS/FastMap. It shows a network of nodes connected by edges, with a specific path highlighted by arrows. The nodes are arranged in a roughly circular pattern, and the path starts from a node at the top right and moves through several nodes in a clockwise direction.

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

14

CarnegieMellon

## Outline

Goal: 'Find similar / interesting things'

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

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

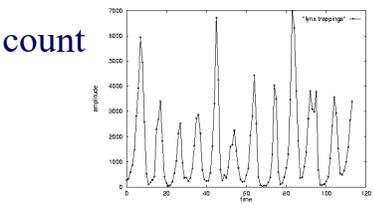
15

CarnegieMellon

## Time series & forecasting

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

Find: patterns and/or compress



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

16

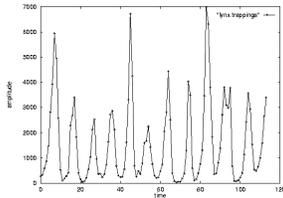
CarnegieMellon

## Time series & forecasting

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

Find: patterns and/or compress

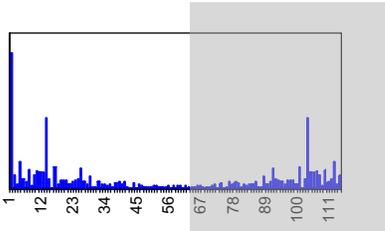
count



year

DFT





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

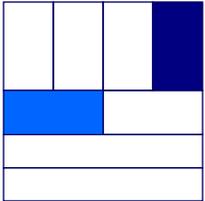
17

CarnegieMellon

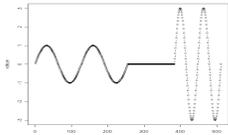
## Wavelets

- Q: baritone/silence/soprano - DWT?

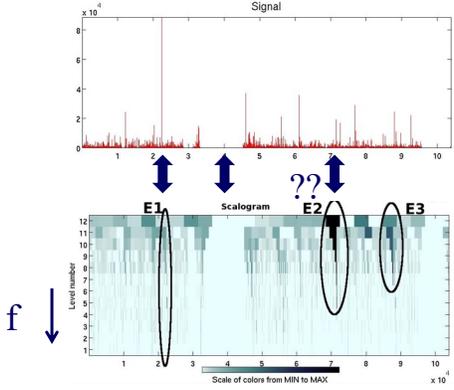
f ↑



value



f ↓



Scale of colors from MIN to MAX

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

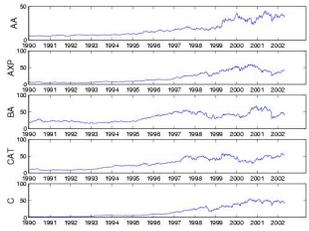
18

CarnegieMellon



## Problem:

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



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

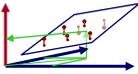
19

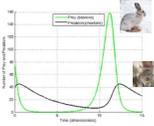
CarnegieMellon



## 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**





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

20

CarnegieMellon

## Outline

Goal: 'Find similar / interesting things'

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

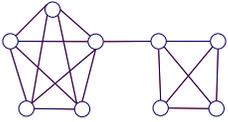
15-826 Copyright (c) 2019 C. Faloutsos 21

21

CarnegieMellon

## Graphs

- Real graphs: surprising patterns
  - ??



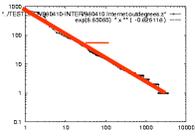
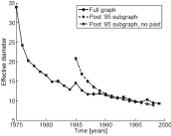
15-826 Copyright (c) 2019 C. Faloutsos 22

22

CarnegieMellon

## 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

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

23

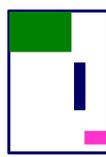
CarnegieMellon

## Graphs - SVD

- Hubs/Authorities (SVD on adjacency matrix)

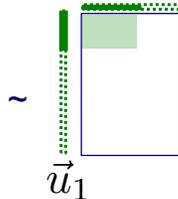
M  
products

N  
users

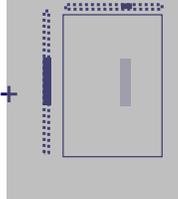


~

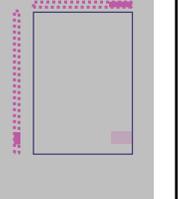
$\vec{u}_1$



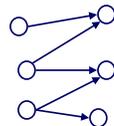
+



+



$\vec{v}_1$



‘meat-eaters’  
‘steaks’

‘vegetarians’  
‘plants’

‘kids’  
‘cookies’

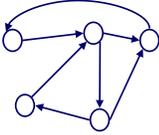
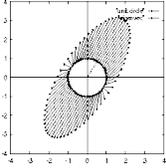
15-826
Copyright (c) 2019 C. Faloutsos
24

24

CarnegieMellon

## Graphs - PageRank

- Hubs/Authorities (SVD on adjacency matrix)
- PageRank (fixed point  $\rightarrow$  eigenvector)

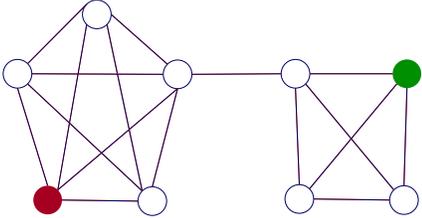
15-826
Copyright (c) 2019 C. Faloutsos
25

25

CarnegieMellon

## Belief Propagation

- What color, for the rest?

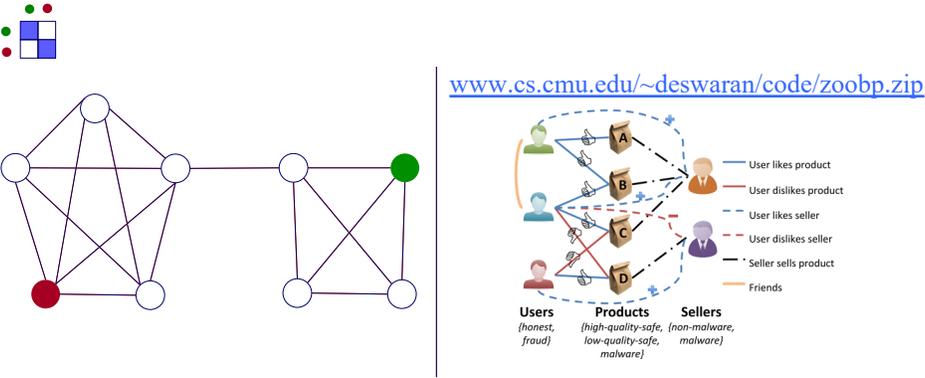
15-826
Copyright (c) 2019 C. Faloutsos
26

26

CarnegieMellon

## Belief Propagation

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



[www.cs.cmu.edu/~deswaran/code/zoobp.zip](http://www.cs.cmu.edu/~deswaran/code/zoobp.zip)

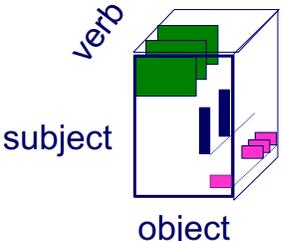
15-826 Copyright (c) 2019 C. Faloutsos 27

27

CarnegieMellon

## Tensors

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



15-826 Copyright (c) 2019 C. Faloutsos 28

28

CarnegieMellon

## Tensors

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

verb

subject

object

=

politicians

+

artists

+

athletes

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

29

CarnegieMellon

## Association Rules

- Given many market baskets
- Find which products sell together

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

30

CarnegieMellon

## Association Rules

- Given many market baskets
- Find which products sell together 
- Association rules [‘large itemsets’]
  - {Milk, bread} -> butter [milk,bread,butter: often ]

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

31

CarnegieMellon

## Taking a step back:

We saw some fundamental, recurring concepts and tools:

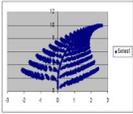
15-826 Copyright (c) 2019 C. Faloutsos 32

32

CarnegieMellon

# T1: Powerful, recurring tools

- Fractals/ self similarity



15-826

Copyright (c) 2019 C. Faloutsos

33

33

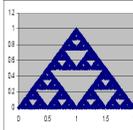
CarnegieMellon

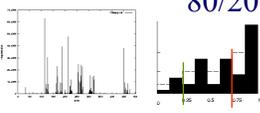
# T1: Powerful, recurring tools

- Fractals/ self similarity  $\leftrightarrow$  Power laws
  - Zipf, Korcak, Pareto’ s laws
  - intrinsic dimension (Sierpinski triangle)
  - correlation integral
  - Barnsley’ s IFS compression
  - Kronecker graphs

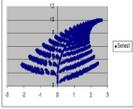


15-826





80/20



15-826

Copyright (c) 2019 C. Faloutsos

34

34

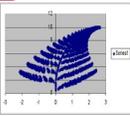
CarnegieMellon

## T1: Powerful, recurring tools

- Fractals/ self similarity
  - Zipf, Kor

- ‘Take logarithms’
- mean → meaningless
- Gaussian trap

(wecker graphs)



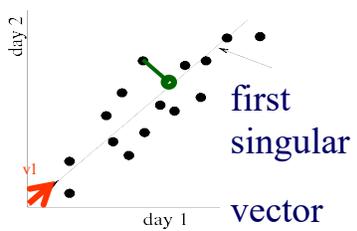
15-826 Copyright (c) 2019 C. Faloutsos 35

35

CarnegieMellon

## T2: Powerful, recurring tools

- SVD (optimal L2 approx)



15-826 Copyright (c) 2019 C. Faloutsos 36

36

CarnegieMellon

## T2: Powerful, recurring tools

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

day 2

day 1

first singular vector

$\vec{u}_1$

$\vec{v}_1$

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

37

CarnegieMellon

## T3: powerful, recurring tools

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

count

year

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

38

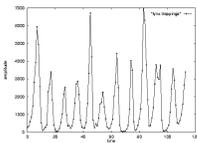
CarnegieMellon

## T3: powerful, recurring tools

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

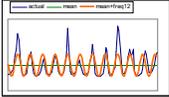


count



year

A1: Fourier (DFT)



A2: Wavelets (DWT)



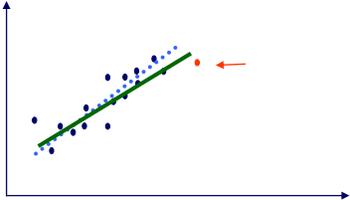

15-826 Copyright (c) 2019 C. Faloutsos

39

CarnegieMellon

## T4: Powerful, recurring tools

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



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

40

CarnegieMellon 

## 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

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

41

CarnegieMellon 

## 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

**• ‘Take logarithms’**  
**• mean -> meaningless**  
**• Gaussian trap**

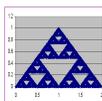
15-826 Copyright (c) 2019 C. Faloutsos 42

42

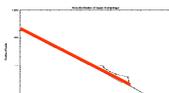
CarnegieMellon

## Thank you!

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



15-826



Copyright (c) 2019 C. Faloutsos

• 'Take logarithms'  
• mean  $\rightarrow$  meaningless  
• Gaussian trap

43