



## Lecture #30: Conclusions

*C. Faloutsos*



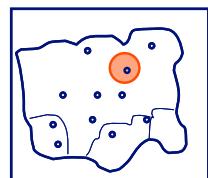
Goal: ‘Find similar / interesting things’

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

• D  
15-826

(c) 2013, C. Faloutsos

2



15-826

(c) 2013, C. Faloutsos

3

CMU SCS

## Indexing - similarity search

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

15-826 (c) 2013, C. Faloutsos 4

---



---



---



---



---



---



---



---

CMU SCS

## Indexing - similarity search

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

15-826 (c) 2013, C. Faloutsos 5

---



---



---



---



---



---



---



---

CMU SCS

## Outline

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

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

15-826 (c) 2013, C. Faloutsos 6

---



---



---



---



---



---



---



---



## Text searching

- ‘find all documents with word *bla*’

15-826

(c) 2013, C. Faloutsos

7



## Text searching

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

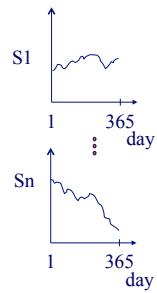
15-826

(c) 2013, C. Faloutsos

8



## Multimedia indexing

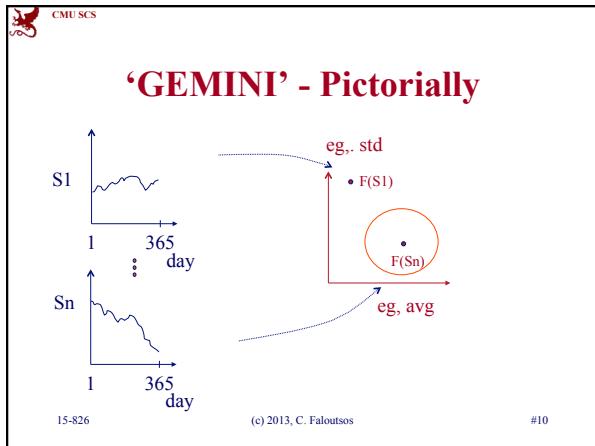


15-826

(c) 2013, C. Faloutsos

9






CMU SCS

## Multimedia indexing

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

15-826

(c) 2013, C. Faloutsos

11



CMU SCS

## Outline

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

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

15-826

(c) 2013, C. Faloutsos

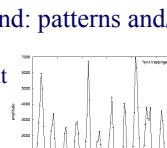
12


 CMU SCS

# Time series & forecasting

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

Find: patterns and/or compress



lynx caught per year

15-826

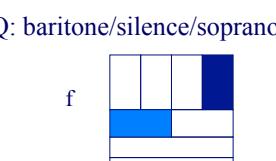
(c) 2013, C. Faloutsos

13


 CMU SCS

# Wavelets

- Q: baritone/silence/soprano - DWT?



The figure illustrates the Discrete Wavelet Transform (DWT) of a signal  $f$  over time  $t$ . The signal  $f$  is represented by a blue waveform with three distinct regions: a baritone section (low frequency), a silence section (no signal), and a soprano section (high frequency). The corresponding wavelet coefficients are shown as a 4x4 grid of blocks. The first three columns of the grid have 4 white blocks each, while the fourth column has 3 blue blocks. Below the grid, the signal  $f$  is plotted against time  $t$ , with the soprano section having a dotted line.

15-826

(c) 2013, C. Faloutsos

14

 CMU SCS

# Time series + forecasting

- Fourier; Wavelets
- Box/Jenkins and AutoRegression
- non-linear/chaotic forecasting (fractals again)
  - ‘Delayed Coordinate Embedding’  $\sim$  nearest neighbors

15-826

(c) 2013, C. Faloutsos

15

 CMU SCS

# Outline

 CMU SCS

# Graphs

- Real graphs: surprising patterns
  - ??

15-826

(c) 2013, C. Faloutsos

17

 CMU SCS

# Graphs

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

15-826

(c) 2013, C. Faloutsos

18



## Graphs - SVD

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

15-826

(c) 2013, C. Faloutsos

19



## Outline

Goal: ‘Find similar / interesting things’

- Intro to DB
- Indexing - similarity search
- • Data Mining

15.826

(c) 2013 C. Faloutsos

20



Data Mining - DB

15.926

© 2012 S. E. J.

21





## Data Mining - DB

- Association Rules ('diapers' -> 'beer')
- [ ~~OLAP (DataCubes, roll-up, drill-down)~~ ]
- [ ~~Classifiers~~ ]

15-826

(c) 2013, C. Faloutsos

22



## Taking a step back:

We saw some fundamental, recurring concepts and tools:

15.826

(c) 2013 C. Faloutsos

23



## Powerful, recurring tools

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



15-826

(c) 2013, C. Faloutsos

24




 CMU SCS

# Powerful, recurring tools

- Fractals/ self similarity
  - Zipf, Korcak, B
  - (Kronecker graphs)



15-826

(c) 2013, C. Faloutsos

25


 CMU SCS

 CMU SCS

# Powerful, recurring tools

- Discrete Fourier Transform
- Wavelets

15-826

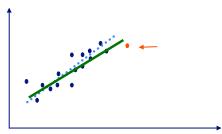
(c) 2013, C. Faloutsos

27

CMU SCS

## Powerful, recurring tools

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



15-826 (c) 2013, C. Faloutsos 28

---



---



---



---



---



---



---



---

CMU SCS

## Summary

- **fractals / power laws** probably lead to the most startling discoveries ('the mean may be meaningless')
- **SVD**: behind PageRank/HITS/tensors/...
- **Wavelets**: Nature seems to prefer them
- **RLS**: matrix inversion, without inverting
- approximate counting (do the impossible!)

15-826 (c) 2013, C. Faloutsos 29

---



---



---



---



---



---



---



---

CMU SCS

## Thank you!

- Feel free to contact me:
  - christos@cs GHC 8019
- Reminder: faculty course eval's:
  - [www.cmu.edu/hub/fce/](http://www.cmu.edu/hub/fce/)
- Final: Tue, Dec. 10, 1:00-4:00p.m. WEH7500  
(double-check with
  - [www.cmu.edu/hub/docs/final-exams.pdf](http://www.cmu.edu/hub/docs/final-exams.pdf)
- Have a great break!

15-826 (c) 2013, C. Faloutsos 30

---



---



---



---



---



---



---



---