SCS-CMU

Data Mining Tools

A crash course

C. Faloutsos

SCS-CMU

Subset of:

www.cs.cmu.edu/~christos/TALKS/SIGMETRICS03-tut/

SCS-CMU

High-level Outline

- [I - Traditional Data Mining tools
  - classification, CART trees; clustering
- II - Time series: analysis and forecasting
  - ARIMA; Fourier, Wavelets]
- III - New Tools: SVD
- IV - New Tools: Fractals & power laws

High-level Outline

- [I - Traditional Data Mining tools
- II - Time series: analysis and forecasting]
- III - New Tools: SVD
- IV - New Tools: Fractals & power laws

III - SVD - outline

- Introduction - motivating problems
- Definition - properties
- Interpretation / Intuition
- Solutions to posed problems
- Conclusions

SVD - Motivation

- problem #1: find patterns in a matrix
  - (e.g., traffic patterns from several IP-sources)
  - compression; dim. reduction

SCS-CMU

Problem#1

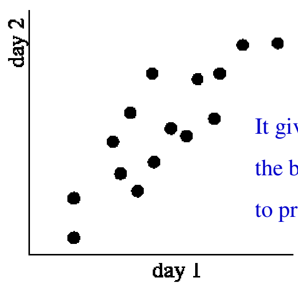
- ~10**6 rows; ~10**3 columns; no updates;
- Compress / find patterns

day / customer	We 7/10/96	Th 7/11/96	Fr 7/12/96	Sa 7/13/96	Su 7/14/96
ABC Inc.	1	1	1	0	0
DEF Ltd.	2	2	2	0	0
GHI Inc.	1	1	1	0	0
KLM Co.	5	5	5	0	0
Smith	0	0	0	2	2
Johnson	0	0	0	3	3
Thompson	0	0	0	1	1

15-744, S07 C. Faloutsos 7

SCS-CMU

SVD - in short:

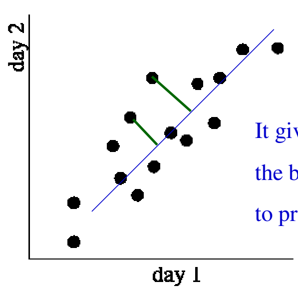


It gives
the best hyperplane
to project on

15-744, S07 C. Faloutsos 8

SCS-CMU

SVD - in short:



It gives
the best hyperplane
to project on

15-744, S07 C. Faloutsos 9

SCS-CMU

III - SVD - outline

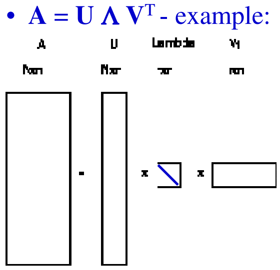
- Introduction - motivating problems
- Definition - properties
- Interpretation / Intuition
- Solutions to posed problems
- Conclusions

15-744, S07 C. Faloutsos 10

SCS-CMU

SVD - Definition

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



15-744, S07 C. Faloutsos 11

Skip

SCS-CMU

SVD - notation

Conventions:

- bold capitals -> matrix (eg. $\mathbf{A}$, $\mathbf{U}$, $\mathbf{\Lambda}$, $\mathbf{V}$)
- bold lower-case -> <u>column</u> vector (eg., $\mathbf{x}$, $\mathbf{v}_1$, $\mathbf{u}_3$)
- regular lower-case -> scalars (eg., λ_1 , λ_r)

15-744, S07 C. Faloutsos 12

SCS-CMU

SVD - Definition

$\boxed{A_{[n \times m]} = U_{[n \times r]} \Lambda_{[r \times r]} (V_{[m \times r]})^T}$

- **A**: n x m matrix (eg., n customers, m days)
- **U**: n x r matrix (n customers, r concepts)
- **Λ**: r x r diagonal matrix (strength of each 'concept') (r : rank of the matrix)
- **V**: m x r matrix (m days, r concepts)

SCS-CMU

SVD - Properties

THEOREM [Press+92]: always possible to decompose matrix **A** into $\mathbf{A} = \mathbf{U} \, \mathbf{\Lambda} \, \mathbf{V}^T$, where

- **U, Λ, V**: unique (*)
- **U**, **V**: column orthonormal (ie., columns are unit vectors, orthogonal to each other)
  - $\mathbf{U}^T \mathbf{U} = \mathbf{I}$; $\mathbf{V}^T \mathbf{V} = \mathbf{I}$ (**I:** identity matrix)
- **Λ:** eigenvalues are positive, and sorted in decreasing order

SCS-CMU

SVD - example

- Customers; days; #packets

	day / customer	We 7/10/96	Th 7/11/96	Fr 7/12/96	Sa 7/13/96	Su 7/14/96
Comm.	ABC Inc.	1	1	1	0	0
	DEF Ltd.	2	2	2	0	0
	GHI Inc.	1	1	1	0	0
	KLM Co.	5	5	5	0	0
Res.	Smith	0	0	0	2	2
	Johnson	0	0	0	3	3
	Thompson	0	0	0	1	1

SCS-CMU

SVD - Example

- **A = U Λ V**^{T} - example:

$$
\begin{array}{cc}
 & \begin{array}{ccccc} \text{We} & \text{Th.} & \text{Fr} & \text{Sa} & \text{Su} \end{array} \\
\begin{array}{c} \text{Com.} \\ \\ \\ \\ \text{Res.} \\ \\ \\ \end{array} &
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
0.18 & 0 \\
0.36 & 0 \\
0.18 & 0 \\
0.90 & 0 \\
0 & 0.53 \\
0 & 0.80 \\
0 & 0.27
\end{bmatrix}
\times
\begin{bmatrix}
9.64 & 0 \\
0 & 5.29
\end{bmatrix}
\times
\begin{bmatrix}
0.58 & 0.58 & 0.58 & 0 & 0 \\
0 & 0 & 0 & 0.71 & 0.71
\end{bmatrix}
$$

SCS-CMU

III - SVD - outline

- Introduction - motivating problems
- Definition - properties
- Interpretation / Intuition
  - #1: customers, days, concepts
  - #2: best projection - dimensionality reduction
- Solutions to posed problems
- Conclusions

SCS-CMU

SVD - Interpretation #1

'customers', 'days' and 'concepts'

- **U**: customer-to-concept similarity matrix
- **V**: day-to-concept sim. matrix
- **Λ**: its diagonal elements: 'strength' of each concept

SCS-CMU

SVD - Interpretation #1

- **A = U Λ V^T** - example:

Rank=2

2x2

$$\begin{array}{c|c} & \text{We} \quad \text{Th.} \quad \text{Fr} \quad \text{Sa} \quad \text{Su} \\ \begin{array}{c}\text{Com.}\\ \\ \\ \\ \text{Res.}\\ \\ \\ \end{array} & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \text{x} \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \text{x} \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-744, S07 C. Faloutsos 19

SCS-CMU

SVD - Interpretation #1

- **A = U Λ V^T** - example:

Rank=2

=2 'concepts'

$$\begin{array}{c|c} & \text{We} \quad \text{Th.} \quad \text{Fr} \quad \text{Sa} \quad \text{Su} \\ \begin{array}{c}\text{Com.}\\ \\ \\ \\ \text{Res.}\\ \\ \\ \end{array} & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \text{x} \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \text{x} \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-744, S07 C. Faloutsos 20

SCS-CMU

(reminder)

- Customers; days; #packets

	customer \ day	We 7/10/96	Th 7/11/96	Fr 7/12/96	Sa 7/13/96	Su 7/14/96
Comm.	ABC Inc.	1	1	1	0	0
	DEF Ltd.	2	2	2	0	0
	GHI Inc.	1	1	1	0	0
	KLM Co.	5	5	5	0	0
Res.	Smith	0	0	0	2	2
	Johnson	0	0	0	3	3
	Thompson	0	0	0	1	1

15-744, S07 C. Faloutsos 21

SCS-CMU

SVD - Interpretation #1

- **$A = U \Lambda V^T$** - example:

U: customer-to-concept similarity matrix

weekday-concept

W/end-concept

$$\begin{array}{c|c} & \begin{array}{ccccc} \text{We} & \text{Th.} & \text{Fr} & \text{Sa} & \text{Su} \end{array} \\ \begin{array}{c} \text{Com.} \\ \\ \\ \\ \text{Res.} \\ \\ \\ \end{array} & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-744, S07 C. Faloutsos 22

SCS-CMU

SVD - Interpretation #1

- **$A = U \Lambda V^T$** - example:

U: Customer to concept similarity matrix

weekday-concept

W/end-concept

$$\begin{array}{c|c} & \begin{array}{ccccc} \text{We} & \text{Th.} & \text{Fr} & \text{Sa} & \text{Su} \end{array} \\ \begin{array}{c} \text{Com.} \\ \\ \\ \\ \text{Res.} \\ \\ \\ \end{array} & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-744, S07 C. Faloutsos 23

Skip

SCS-CMU

SVD - Interpretation #1

- **$A = U \Lambda V^T$** - example:

unit

$$\begin{array}{c|c} & \begin{array}{ccccc} \text{We} & \text{Th.} & \text{Fr} & \text{Sa} & \text{Su} \end{array} \\ \begin{array}{c} \text{Com.} \\ \\ \\ \\ \text{Res.} \\ \\ \\ \end{array} & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} \boxed{\begin{matrix} 0.18 \\ 0.36 \\ 0.18 \\ 0.90 \\ 0 \\ 0 \\ 0 \end{matrix}} & \begin{matrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.53 \\ 0.80 \\ 0.27 \end{matrix} \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-744, S07 C. Faloutsos 24

SCS-CMU

SVD - Interpretation #1

- **A = U Λ V^T** - example:

Strength of 'weekday' concept

weekday-concept

$$\begin{array}{c} \\ \text{Com.} \\ \\ \text{Res.} \end{array} \overset{\text{We Th. Fr Sa Su}}{\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-744, S07 C. Faloutsos 25

SCS-CMU

SVD - Interpretation #1

- **A = U Λ V^T** - example:

weekday-concept

V: day to concept similarity matrix

$$\begin{array}{c} \\ \text{Com.} \\ \\ \text{Res.} \end{array} \overset{\text{We Th. Fr Sa Su}}{\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-744, S07 C. Faloutsos 26

SCS-CMU

III - SVD - outline

- Introduction - motivating problems
- Definition - properties
- Interpretation / Intuition
  - #1: customers, days, concepts
  - #2: best projection - dimensionality reduction
- Solutions to posed problems
- Conclusions

15-744, S07 C. Faloutsos 27

SCS-CMU

SVD - Interpretation #2

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

15-744, S07 C. Faloutsos 28

SCS-CMU

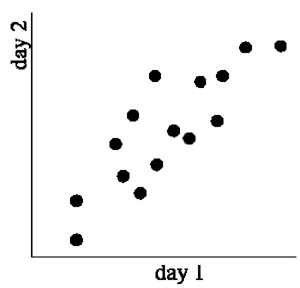
SVD - Interpretation #2

day / customer	We 7/10/96	Th 7/11/96	Fr 7/12/96	Sa 7/13/96	Su 7/14/96
ABC Inc.	1	1	1	0	0
DEF Ltd.	2	2	2	0	0
GHI Inc.	1	1	1	0	0
KLM Co.	5	5	5	0	0
Smith	0	0	0	2	2
Johnson	0	0	0	3	3
Thompson	0	0	0	1	1

15-744, S07 C. Faloutsos 29

SCS-CMU

SVD - Interpretation#2

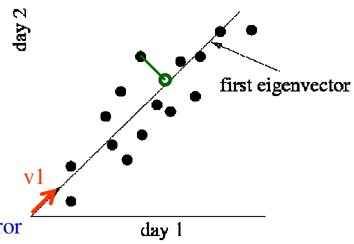


15-744, S07 C. Faloutsos 30

SCS-CMU

SVD - interpretation #2

SVD: gives
best axis to project



- minimum RMS error

15-744, S07 C. Faloutsos 31

SCS-CMU

SVD - Interpretation #2

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

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{array}{l} \text{v1} \\ \begin{bmatrix} \boxed{0.58 \;\; 0.58 \;\; 0.58 \;\; 0 \;\; 0} \\ 0 \;\; 0 \;\; 0 \;\; 0.71 \;\; 0.71 \end{bmatrix} \end{array}$$

15-744, S07 C. Faloutsos 32

SCS-CMU

SVD - Interpretation #2

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

variance ('spread') on the v1 axis

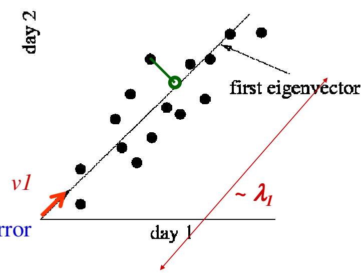
$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-744, S07 C. Faloutsos 33

SCS-CMU

SVD - interpretation #2

SVD: gives best axis to project



- minimum RMS error

SCS-CMU

SVD, PCA and the v vectors

- how to 'read' the **v** vectors (= principal components)

SCS-CMU

SVD

- Recall**:** $\mathbf{A} = \mathbf{U}\, \mathbf{\Lambda}\, \mathbf{V}^T$ - example:

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} \boxed{0.58 \;\; 0.58 \;\; 0.58 \;\; 0 \;\; 0} \\ 0 \;\; 0 \;\; 0 \;\; 0.71 \;\; 0.71 \end{bmatrix}$$

SCS-CMU

SVD

- First Principal component = **v1** -> weekdays are correlated positively
- similarly for **v2**
- (we'll see negative correlations later)

$$\begin{array}{c} \\ \text{We} \\ \text{Th} \\ \text{Fr} \\ \text{Sa} \\ \text{Su} \end{array} \begin{array}{c} \begin{array}{cc} \mathbf{v1} & \mathbf{v2} \end{array} \\ \begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix} \end{array}$$

SCS-CMU

Skip

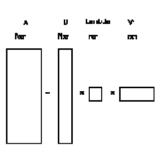
SVD - Complexity

- O(n * m * m) or O(n * n * m) (whichever is less)
- less work, if we just want eigenvalues
- ... or if we want first k eigenvectors
- ... or if the matrix is sparse [Berry]
- Implemented: in *any* linear algebra package (LINPACK, matlab, Splus, mathematica ...)

SCS-CMU

SVD - conclusions so far

- SVD: **A= U Λ V$^{\mathbf{T}}$** : unique (*)
- **U**: row-to-concept similarities
- **V**: column-to-concept similarities
- **Λ**: strength of each concept



(*) see [Press+92]

SCS-CMU

SVD - conclusions so far

- dim. reduction: keep the first few strongest eigenvalues (80-90% of 'energy' [Fukunaga])
- SVD: picks up linear correlations

SCS-CMU

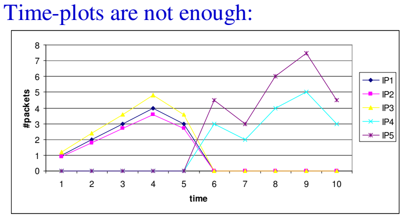
III - SVD - outline

- Introduction - motivating problems
- Definition - properties
- Interpretation / Intuition
- Solutions to posed problems
  - P1: **patterns** in a matrix; compression
- Conclusions

SCS-CMU

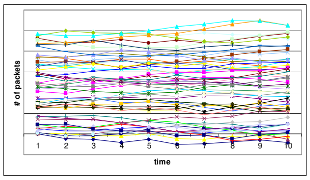
SVD & visualization:

- Visualization for free!
  - Time-plots are not enough:



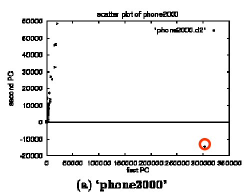
SVD & visualization:

- Visualization for free!
  - Time-plots are not enough:



SVD & visualization

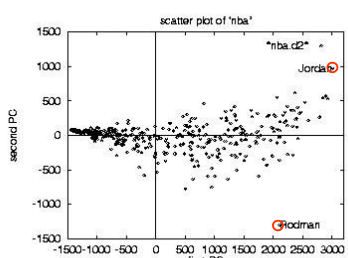
- SVD: project 365-d vectors to best 2 dimensions, and plot:
- no Gaussian clusters; Zipf-like distribution



phonecalls

SVD and visualization

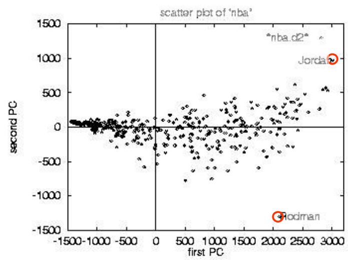
NBA dataset
~500 players;
~30 attributes
(#games, #points, #rebounds,…)



SCS-CMU

SVD and visualization

could be network dataset:
- *N* IP sources
- *k* attributes (#http bytes, #http packets)



SCS-CMU

Moreover, PCA/rules for free!

- SVD ~ PCA = Principal component analysis
- PCA: get eigenvectors **v1**, **v2**, ...
- ignore entries with small abs. value
- try to interpret the rest

SCS-CMU

Skip

PCA & Rules

NBA dataset - **V** matrix (term to ‘concept’ similarities)

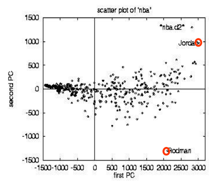
field	RR_1	RR_2	RR_3
minutes played	.808	−.4	
field goals			
goal attempts			
points	.406	.199	
total rebounds		−.489	.602
assists			−.486
steals			−.07

v1

Skip

PCA & Rules

- (Ratio) Rule#1: minutes:points = 2:1
- corresponding concept?



v1

Skip

PCA & Rules

- RR1: minutes:points = 2:1
- corresponding concept?
- A: 'goodness' of player
- (in a systems setting, could be 'volume of traffic' generated by this IP address)

Skip

PCA & Rules

- RR2: points: rebounds negatively correlated(!)

field	RR_1	RR_2	RR_3
minutes played	.808	−.4	
field goals			
goal attempts			
points	.406	.199	
total rebounds		−.489	.602
assists			−.486
steals			−.07

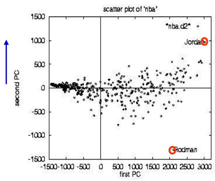
SCS-CMU

Skip

PCA & Rules

- RR2: points: rebounds negatively correlated(!) - concept?

v2



15-744, S07 C. Faloutsos 52

SCS-CMU

Skip

PCA & Rules

- RR2: points: rebounds negatively correlated(!) - concept?
- A: position: offensive/defensive
- (in a network setting, could be e-mailers versus gnutella-users)

15-744, S07 C. Faloutsos 53

SCS-CMU

III - SVD - outline

- Introduction - motivating problems
- Definition - properties
- Interpretation / Intuition
- Solutions to posed problems
  - P1: patterns in a matrix; compression
- Conclusions

15-744, S07 C. Faloutsos 54

SVD - conclusions

SVD: a **valuable** tool , whenever we have a matrix, e.g.

- many time sequences
- many feature vectors
- graph (-> adjacency matrix)

SVD - conclusions

SVD: a **valuable** tool , whenever we have a matrix, e.g.

- many time sequences
  - SVD finds groups
  - principal components
  - dim. reduction

	#packets on day1	#packets on day2	...		
IP address1	1	1	1	0	0
IP address2	2	2	2	0	0
IP address3	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

SVD - conclusions

SVD: a **valuable** tool , whenever we have a matrix, e.g.

- feature vectors
  - SVD finds groups
  - principal components
  - (Ratio) Rules
  - visualization

	#packets sent	#bytes sent	#packets lost	...	
IP address1	1	1	1	0	0
IP address2	2	2	2	0	0
IP address3	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

SVD - conclusions

SVD: a **valuable** tool , whenever we have a matrix, e.g.

- adjacency matrix
  - source, dest, bandwidth
  - SVD -> ‘most central node’

	Dest. router1	Dest. router2	Dest. router3		...
Source router1	1	1	1	0	0
Source router2	2	2	2	0	0
Source router3	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

SVD - conclusions - cont’d

Has been used/re-invented **many times**:

- LSI (Latent Semantic Indexing) [Foltz+92]
- PCA (Principal Component Analysis) [Jolliffe86]
- KL (Karhunen-Loeve Transform)
- Mahalanobis distance
- ...

Resources: Software and urls

- SVD packages: in **many** systems (matlab, mathematica, LINPACK, LAPACK)
- stand-alone, free code: SVDPACK from Michael Berry

  http://www.cs.utk.edu/~berry/projects.html

SCS-CMU

Books

- Faloutsos, C. (1996). *Searching Multimedia Databases by Content*, Kluwer Academic Inc.
- Jolliffe, I. T. (1986). *Principal Component Analysis*, Springer Verlag.

SCS-CMU

Books

- [Press+92] William H. Press, Saul A. Teukolsky, William T. Vetterling and Brian P. Flannery: *Numerical Recipes in C*, Cambridge University Press, 1992, 2nd Edition. (Great description, intuition and code for SVD)

SCS-CMU

Additional Reading

- Berry, Michael: http://www.cs.utk.edu/~lsi/
- Brin, S. and L. Page (1998). *Anatomy of a Large-Scale Hypertextual Web Search Engine*. 7th Intl World Wide Web Conf.

Additional Reading

- [Foltz+92] Foltz, P. W. and S. T. Dumais (Dec. 1992). "*Personalized Information Delivery: An Analysis of Information Filtering Methods*." Comm. of ACM (CACM) 35(12): 51-60.

Additional Reading

- Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, Academic Press.
- Kleinberg, J. (1998). *Authoritative sources in a hyperlinked environment*. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms.

Additional Reading

- Korn, F., H. V. Jagadish, et al. (May 13-15, 1997). *Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences*. ACM SIGMOD, Tucson, AZ.
- Korn, F., A. Labrinidis, et al. (2000). "*Quantifiable Data Mining Using Ratio Rules*." VLDB Journal 8(3-4): 254-266.

SCS-CMU

Fractals

15-744, S07 C. Faloutsos 67

SCS-CMU

High-level Outline

- I - Traditional Data Mining tools
- II - Time series: analysis and forecasting
- III - New Tools: SVD
- IV - New Tools: Fractals & power laws

15-744, S07 C. Faloutsos 68

SCS-CMU

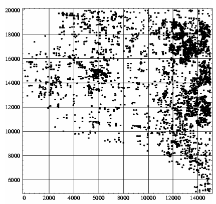
IV - Fractals - outline

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Fast Estimation of fractal dimension
- Solutions to posed problems
- More examples and tools
- Conclusions – practitioner's guide

15-744, S07 C. Faloutsos 69

SCS-CMU

Problem #0: GIS - points



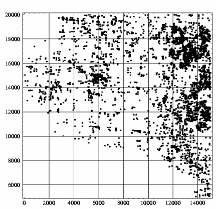
Road end-points of Montgomery county:

- Q1: # neighbors(*r*)?
- Q2 : distribution?
  - not uniform
  - not Gaussian
  - no rules??

15-744, S07 C. Faloutsos 70

SCS-CMU

Problem #0: GIS - points



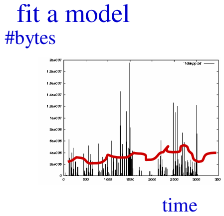
(could be: geo-locations of IP addresses launching DDoS attack)

15-744, S07 C. Faloutsos 71

SCS-CMU

Problem #1: traffic

- disk trace (from HP - J. Wilkes); Web traffic - fit a model



- how many explosions to expect?

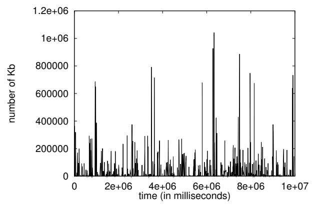
- queue length distr.?

15-744, S07 C. Faloutsos 72

SCS-CMU

Problem #1': traffic

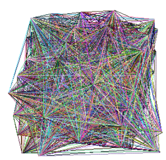
- Kb per unit time (requests on a web server)
  http://repository.cs.vt.edu/ lbl-conn-7.tar.Z



Skip SCS-CMU

Problem #2 - topology

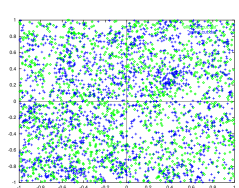
How does the Internet look like?



Skip SCS-CMU

Problem #3 - spatial d.m.

Galaxies (Sloan Digital Sky Survey w/ B. Nichol)



- 'spiral' and 'elliptical' galaxies

- patterns?

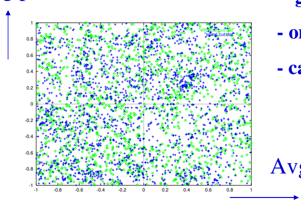
- attraction/repulsion?

- separable?

SCS-CMU

Problem #3 - spatial d.m.

Avg packet rate



Avg packet size

- 'good' and 'bad' IP addresses
- or 'read' and 'write' requests
- can we separate them?

15-744, S07 C. Faloutsos 76

SCS-CMU

Common answer:

Fractals / self-similarities / power laws

15-744, S07 C. Faloutsos 77

SCS-CMU

IV - Fractals - outline

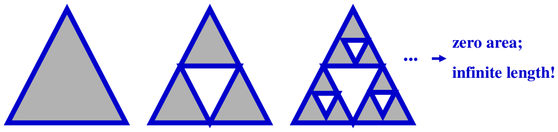
- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Fast Estimation of fractal dimension
- Solutions to posed problems
- More examples and tools
- Conclusions – practitioner's guide

15-744, S07 C. Faloutsos 78

SCS-CMU

What is a fractal?

= self-similar point set, e.g., Sierpinski triangle:



SCS-CMU

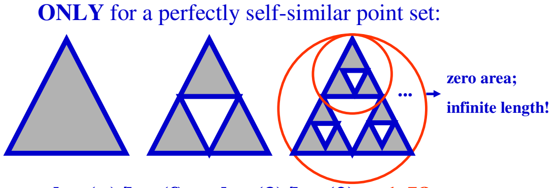
Definitions (cont'd)

- Paradox: Infinite perimeter ; Zero area!
- 'dimensionality': between 1 and 2
- actually: Log(3)/Log(2) = 1.58...

SCS-CMU

Dfn of fd:

ONLY for a perfectly self-similar point set:

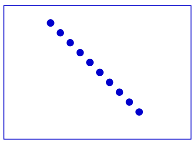


=log(n)/log(f) = log(3)/log(2) = 1.58

SCS-CMU

Intrinsic ('fractal') dimension

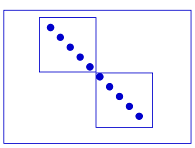
- Q: fractal dimension of a line?
- A: 1 (= log(2)/log(2)!)



SCS-CMU

Intrinsic ('fractal') dimension

- Q: fractal dimension of a line?
- A: 1 (= log(2)/log(2)!)

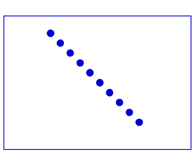


SCS-CMU

Intrinsic ('fractal') dimension

- Q: dfn for a given set of points?

x	y
5	1
4	2
3	3
2	4



SCS-CMU

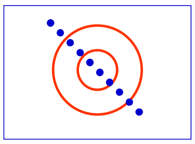
Intrinsic ('fractal') dimension

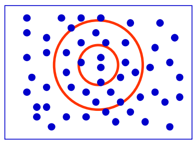
- Q: fractal dimension of a line?
- A: nn (<= r) ~ r^1

('power law': y=x^a)

- Q: fd of a plane?
- A: nn (<= r) ~ r^2

fd== slope of (log(nn) vs log(r))





15-744, S07 C. Faloutsos 85

SCS-CMU

Intrinsic ('fractal') dimension

- Algorithm, to estimate it?

Notice

- *avg nn(<=r)* is exactly *tot#pairs(<=r) / (N)*

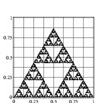
15-744, S07 C. Faloutsos 86

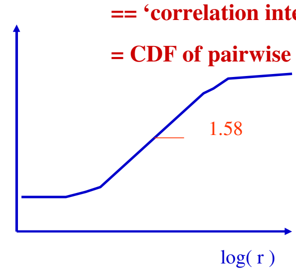
SCS-CMU

Sierpinsky triangle

== 'correlation integral'

= CDF of pairwise distances



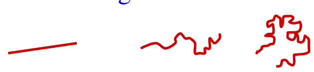


15-744, S07 C. Faloutsos 87

SCS-CMU

Observations:

- Euclidean objects have **integer** fractal dimensions
  - point: 0
  - lines and smooth curves: 1
  - smooth surfaces: 2
- fractal dimension -> roughness of the periphery



15-744, S07 C. Faloutsos 88

SCS-CMU

IV - Fractals - outline

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Fast Estimation of fractal dimension
- Solutions to posed problems
- More examples and tools
- Conclusions – practitioner's guide

15-744, S07 C. Faloutsos 89

SCS-CMU

Skip

Fast estimation

- Bad news: There are more than one fractal dimensions
  - Minkowski fd; Hausdorff fd; Correlation fd; Information fd
- Great news:
  - they can all be computed fast! (O(N); O(N logN))
  - Code is on the web (`www.cs.cmu.edu/~christos`)
  - they usually have nearby values

15-744, S07 C. Faloutsos 90

SCS-CMU

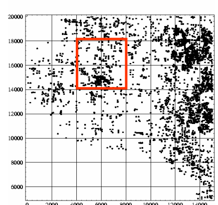
IV - Fractals - outline

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Fast Estimation of fractal dimension
- Solutions to posed problems: P#0 - points
- More examples and tools
- Conclusions – practitioner's guide

15-744, S07 C. Faloutsos 91

SCS-CMU

Problem #0: GIS points



Cross-roads of Montgomery county:

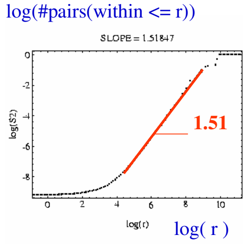
•any rules?

15-744, S07 C. Faloutsos 92

SCS-CMU

Solution #0

log(#pairs(within <= r))



log(r)

A: self-similarity ->

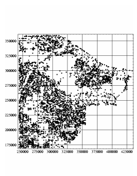
- **<=> fractals**
- **<=> scale-free**
- **<=> power-laws** (y=x^a, F=C*r^(-2))

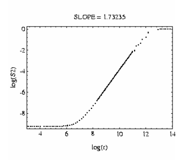
15-744, S07 C. Faloutsos 93

SCS-CMU

Examples:LB county

- Long Beach county of CA (road end-points)





15-744, S07 C. Faloutsos 94

SCS-CMU

IV - Fractals - outline

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Fast Estimation of fractal dimension
- Solutions to posed problems: P#1- traffic
- More examples and tools
- Conclusions – practitioner's guide

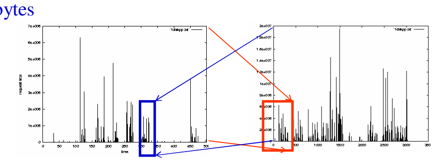
15-744, S07 C. Faloutsos 95

SCS-CMU

Solution #1: traffic

- disk traces: self-similar: (also: [Leland+94])
- How to generate such traffic?

#bytes



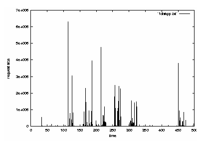
time

15-744, S07 C. Faloutsos 96

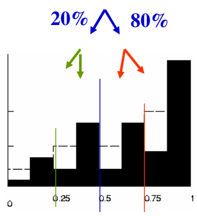
Solution #1: traffic

- disk traces (80-20 'law' = '**multifractal**') [Riedi+99], [Wang+02]

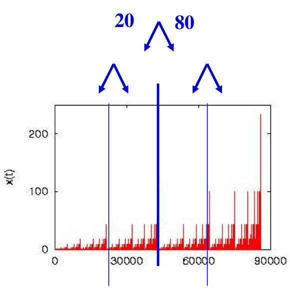
#bytes

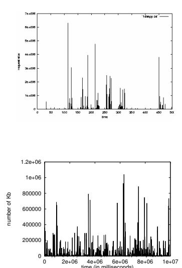


time

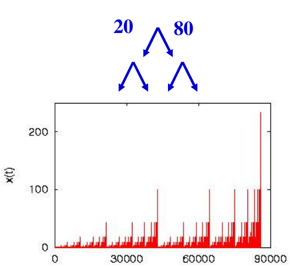


80-20 / multifractals





80-20 / multifractals



- p ; (1-p) in general
- **yes, there are dependencies**

SCS-CMU

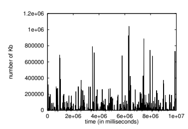
How to estimate *p*?

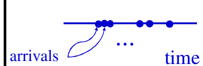
- A: entropy plot [Wang+'02]
- [~ correlation integral]

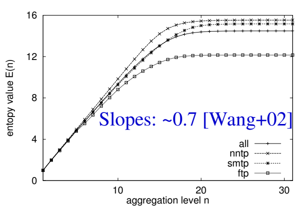
SCS-CMU

Example: traffic

- Kb per unit time (requests on a web server)



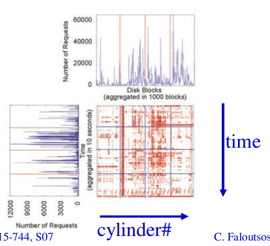


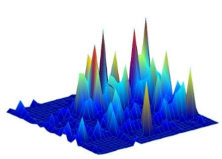


SCS-CMU

More on 80/20: PQRS

- Part of 'self-* storage' project

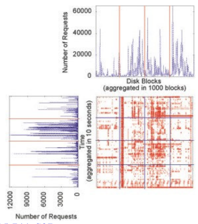


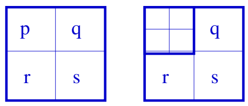


SCS-CMU

More on 80/20: PQRS

- Part of ‘self-* storage’ project





15-744, S07 C. Faloutsos 103

Skip SCS-CMU

IV - Fractals - outline

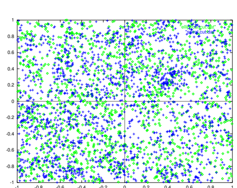
- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Fast Estimation of fractal dimension
- Solutions to posed problems: P#3: spatial d.m.
- More examples and tools
- Conclusions – practitioner’s guide

15-744, S07 C. Faloutsos 104

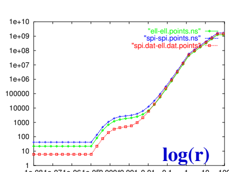
Skip SCS-CMU

Solution#3: spatial d.m.

Galaxies (‘BOPS’ plot - [sigmod2000])



log(#pairs(<=r))



15-744, S07 C. Faloutsos 105

SCS-CMU

IV - Fractals - outline

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Fast Estimation of fractal dimension
- Solutions to posed problems
- More examples and tools
- Conclusions – practitioner's guide

15-744, S07 C. Faloutsos 106

SCS-CMU

Fractals and power laws

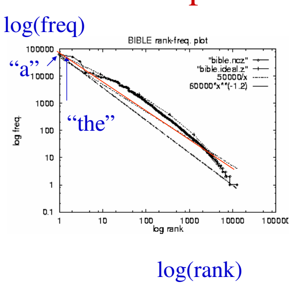
Recall that they are related concepts:

- fractals <=>
- self-similarity <=>
- scale-free <=>
- power laws ($y= x^a$)

15-744, S07 C. Faloutsos 107

SCS-CMU

A famous power law: Zipf's law



- Bible - **rank** vs **frequency** (log-log)

15-744, S07 C. Faloutsos 108

Power laws, cont'ed

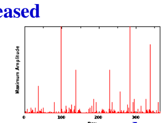
- In- and out-degree distribution of web sites [Barabasi], [IBM-CLEVER]
- length of file transfers [Bestavros+]
- Click-stream data [Montgomery+01]
- web hit counts [Huberman]

Skip

More power laws

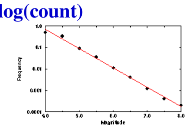
- duration of UNIX jobs; of UNIX file sizes
- Energy of earthquakes (Gutenberg-Richter law) [simscience.org]

Energy released



day

log(count)



Magnitude = log(energy)

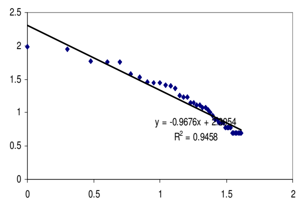
Skip

Even more power laws:

- Income distribution (Pareto's law)
- publication counts (Lotka's law)

Skip

Olympic medals (Sidney):



Skip

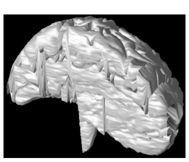
Fractals

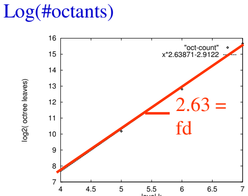
Let’s see some fractals, in real settings:

Skip

Fractals: Brain scans

- Oct-trees; brain-scans





Skip

Fractals: Medical images

[Burdett et al, SPIE '93]:

- benign tumors: fd ~ 2.37
- malignant: fd ~ 2.56

Skip

More fractals:

- cardiovascular system: 3 (!) 
- stock prices (LYCOS) - random walks: 1.5

1 year

2 years

- Coastlines: 1.2-1.58 (Norway!)

Skip

SCS-CMU

IV - Fractals - outline

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Fast Estimation of fractal dimension
- Solutions to posed problems
- More examples and tools
- Conclusions – practitioner's guide

15-744, S07 C. Faloutsos 118

SCS-CMU

Conclusions

- Real data often **disobey** textbook assumptions (Gaussian, Poisson, uniformity, independence)
  - avoid 'mean' - use median, or even better, use:
- fractals, self-similarity, and power laws, to find patterns

15-744, S07 C. Faloutsos 119

SCS-CMU

Practitioner's guide:

- Fractals: help characterize a (non-uniform) set of points
- Detect non-homogeneous regions (eg., legal login time-stamps may have different fd than intruders')

15-744, S07 C. Faloutsos 120

SCS-CMU

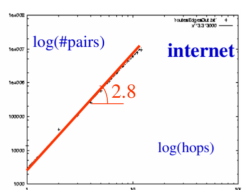
Practitioner's guide

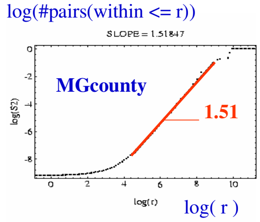
- **tool#1: (for points) 'correlation integral'**: (#pairs within <= *r*) vs (distance *r*)
  - ~ entropy plot
- **tool#2: (for categorical values) rank-frequency** plot (a'la Zipf)

SCS-CMU

Practitioner's guide:

- **tool#1**: correlation integral, for a **set of objects**, with a distance function (slope = intrinsic dimensionality)

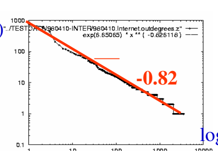


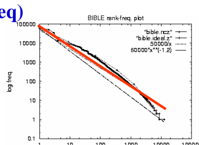


SCS-CMU

Practitioner's guide:

- **tool#2**: rank-frequency plot (for **categorical attributes**)





SCS-CMU

High-level Outline

- [I - Traditional Data Mining tools
- II - Time series: analysis and forecasting]
- III - New Tools: SVD
- IV - New Tools: Fractals & power laws
- 'Take-home' messages:

15-744, S07 C. Faloutsos 124

SCS-CMU

OVERALL CONCLUSIONS

- WEALTH of powerful, scalable tools in data mining (classification, clustering, SVD, fractals)
- traditional assumptions (uniformity, iid, Gaussian, Poisson) are often violated, when fractals/self-similarity/power-laws deliver.

15-744, S07 C. Faloutsos 125

SCS-CMU

Resources: Software & urls

- Fractal dimensions: Software
  - `www.cs.cmu.edu/~christos`

15-744, S07 C. Faloutsos 126

SCS-CMU

References

- (SVD – Ratio Rules): Flip Korn, Alexandros Labrinidis, Yannis Kotidis, Christos Faloutsos Ratio Rules: *A New Paradigm for Fast, Quantifiable Data Mining,* in VLDB 1998, New York, NY.
  www.cs.cmu.edu/~christos/PUBLICATIONS/ratioRules.ps.gz
- (Fractals and bursty traffic): Mengzhi Wang, Anastassia Ailamaki and Christos Faloutsos, *Capturing the spatio-temporal behavior of real traffic data,* Performance 2002 (IFIP Int. Symp. on Computer Performance Modeling, Measurement and Evaluation), Rome, Italy, Sept. 2002
  www.cs.cmu.edu/~christos/PUBLICATIONS/performance02.ps.gz

15-744, S07 C. Faloutsos 127

SCS-CMU

Books

- Fractals: Manfred Schroeder: *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* W.H. Freeman and Company, 1991 (Probably the BEST book on fractals!)

15-744, S07 C. Faloutsos 128

SCS-CMU

Further reading:

- [Barabasi+] Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi, *Diameter of the World Wide Web*, Nature 401 130-131 (1999).
- [Kumar+99] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. *Extracting large scale knowledge bases from the web*. (VLDB) , September 1999.

15-744, S07 C. Faloutsos 129

SCS-CMU

Further reading:

- [sigcomm99] Michalis Faloutsos, Petros Faloutsos and Christos Faloutsos, *What does the Internet look like? Empirical Laws of the Internet Topology,* SIGCOMM 1999
- [sigmod2000] Christos Faloutsos, Bernhard Seeger, Agma J. M. Traina and Caetano Traina Jr., *Spatial Join Selectivity Using Power Laws*, SIGMOD 2000
- [ieeeTN94] W. E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, *On the Self-Similar Nature of Ethernet Traffic,* IEEE Transactions on Networking, 2, 1, pp 1-15, Feb. 1994.

SCS-CMU

Further reading

- [Montgomery+01] A. Montgomery and C. Faloutsos, *Identifying Web Browsing Trends and Patterns*, IEEE Computer, 2001
- [Palmer+01] Chris Palmer, Georgios Siganos, Michalis Faloutsos, Christos Faloutsos and Phil Gibbons: *The connectivity and fault-tolerance of the Internet topology* Workshop on Network Related Data Management (NRDM 2001), Santa Barbara, CA, May 25, 2001.

SCS-CMU

Further reading

- [Riedi+99] R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk, *A Multifractal Wavelet Model with Application to Network Traffic*, IEEE Special Issue on Information Theory, 45. (April 1999), 992-1018.
- [Wang+02] Mengzhi Wang, Tara Madhyastha, Ngai Hang Chang, Spiros Papadimitriou and Christos Faloutsos, *Data Mining Meets Performance Evaluation: Fast Algorithms for Modeling Bursty Traffic*, ICDE 2002, San Jose, CA, 2/26/2002 - 3/1/2002.

SCS-CMU

THANK YOU!

christos@cs.cmu.edu
www.cs.cmu.edu/~christos
Wean Hall 7107

15-744, S07 C. Faloutsos 133