

15-826: Multimedia (Databases) and Data Mining

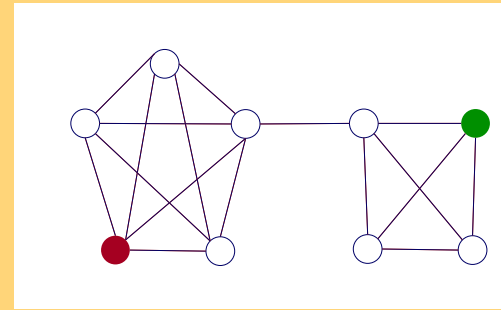
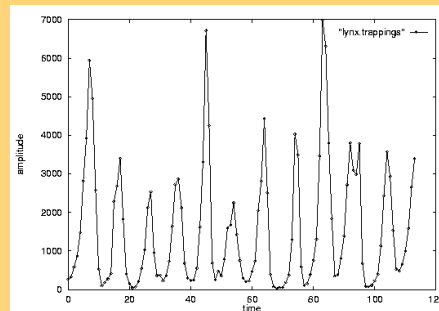
Lecture #31: Conclusions

C. Faloutsos



Problem

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





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

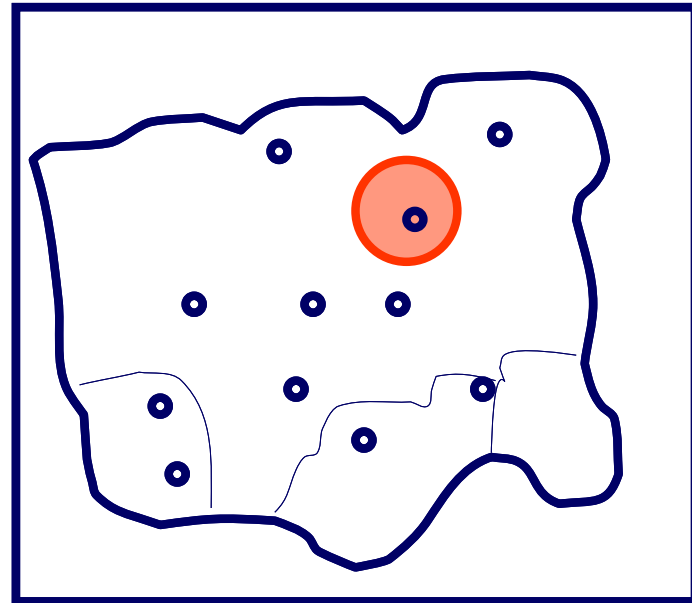


Outline

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

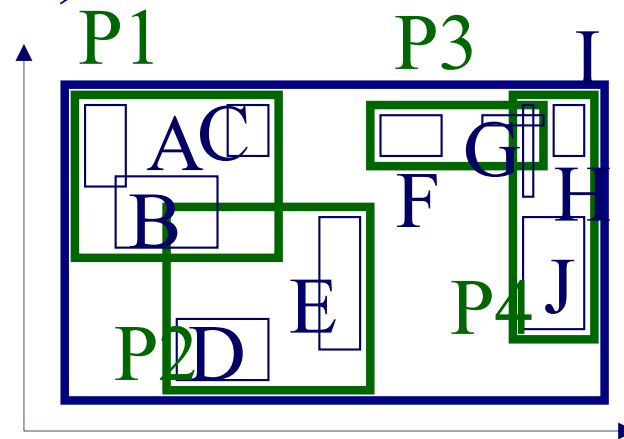
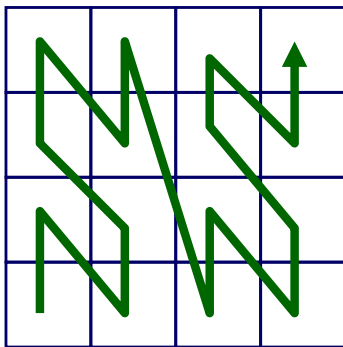
- Intro to DB
- Indexing - similarity search
 - Points
 - Text
 - Time sequences; images etc
 - Graphs

Indexing - similarity search



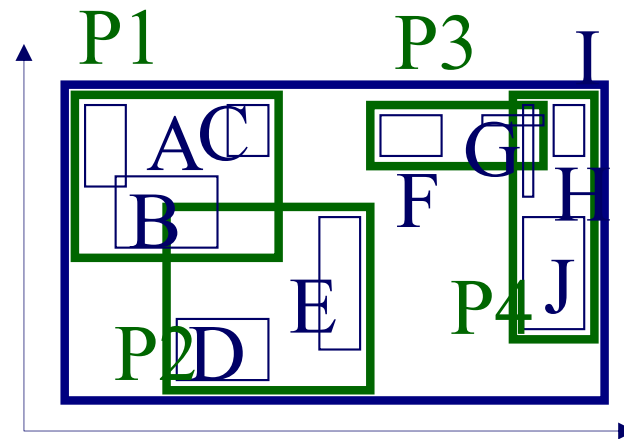
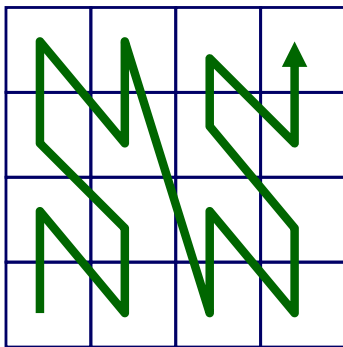
Indexing - similarity search

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



Indexing - similarity search

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





Outline

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

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

Text searching

- ‘find all documents with word *bla*’



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.)

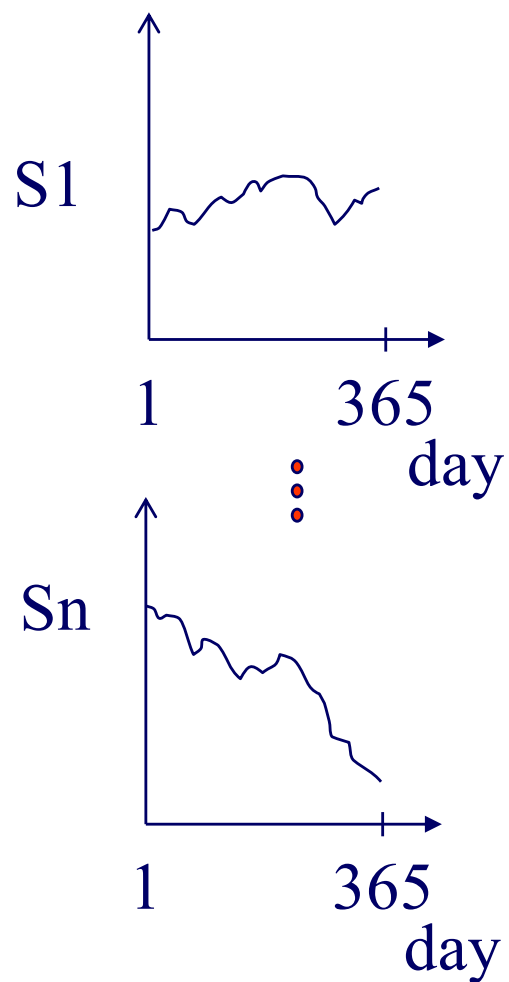


Outline

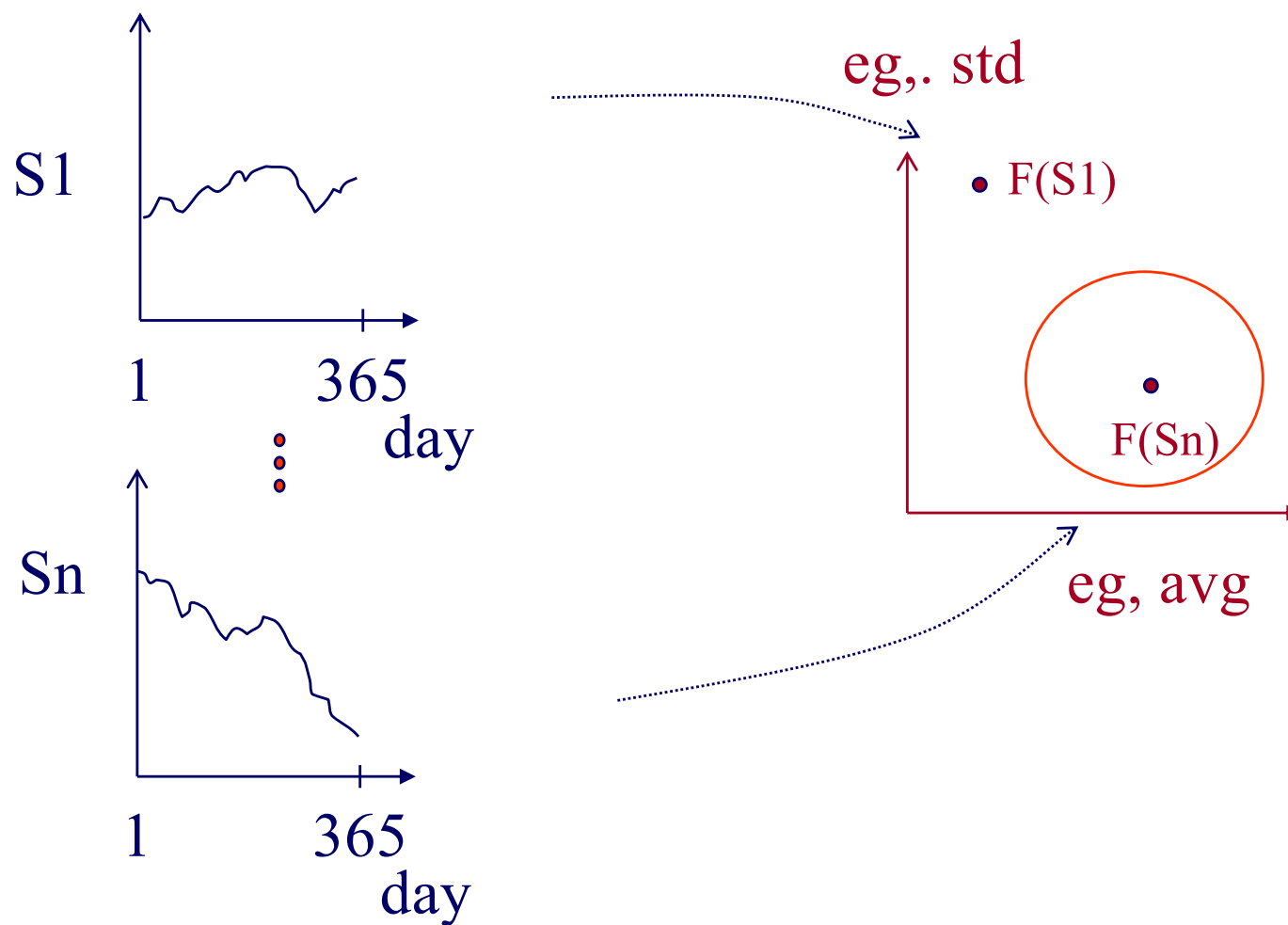
Goal: ‘Find **similar / interesting** things’

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

Multimedia indexing

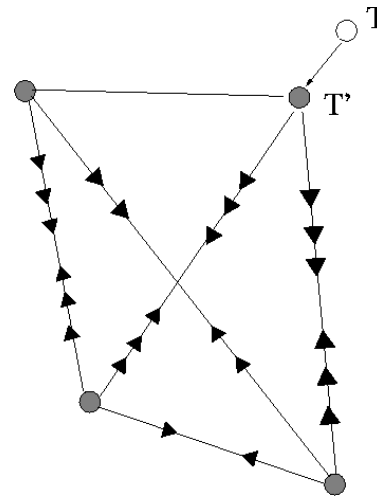


'GEMINI' - Pictorially



Multimedia indexing

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





Outline

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

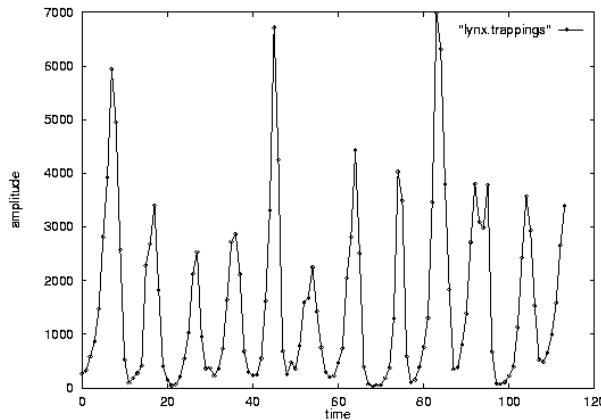
- Intro to DB
- Indexing - similarity search
 - Points
 - Text
 - ➔ – Time sequences; images etc – **DFT/DWT**
 - Graphs

Time series & forecasting

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

Find: patterns and/or compress

count



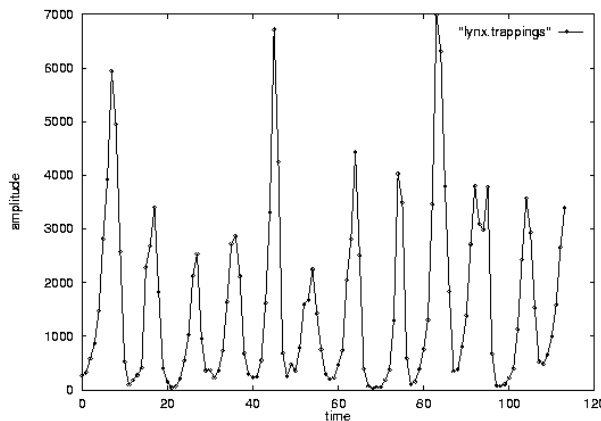
year

Time series & forecasting

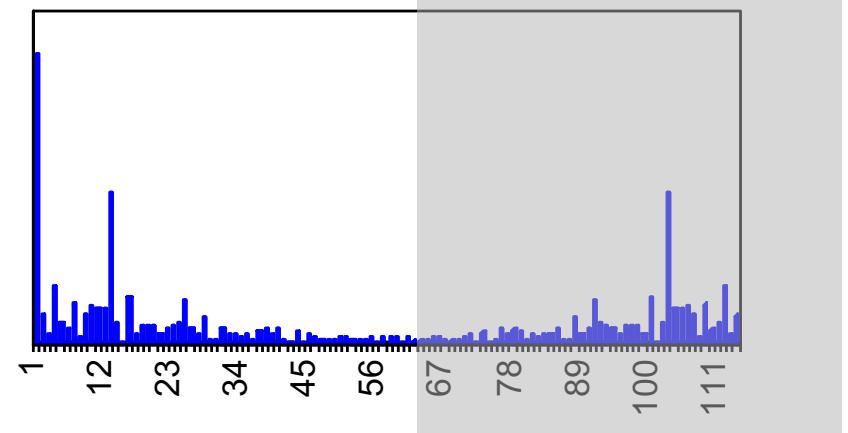
Goal: given a signal (eg., sales over time and/or space)

Find: patterns and/or compress

count

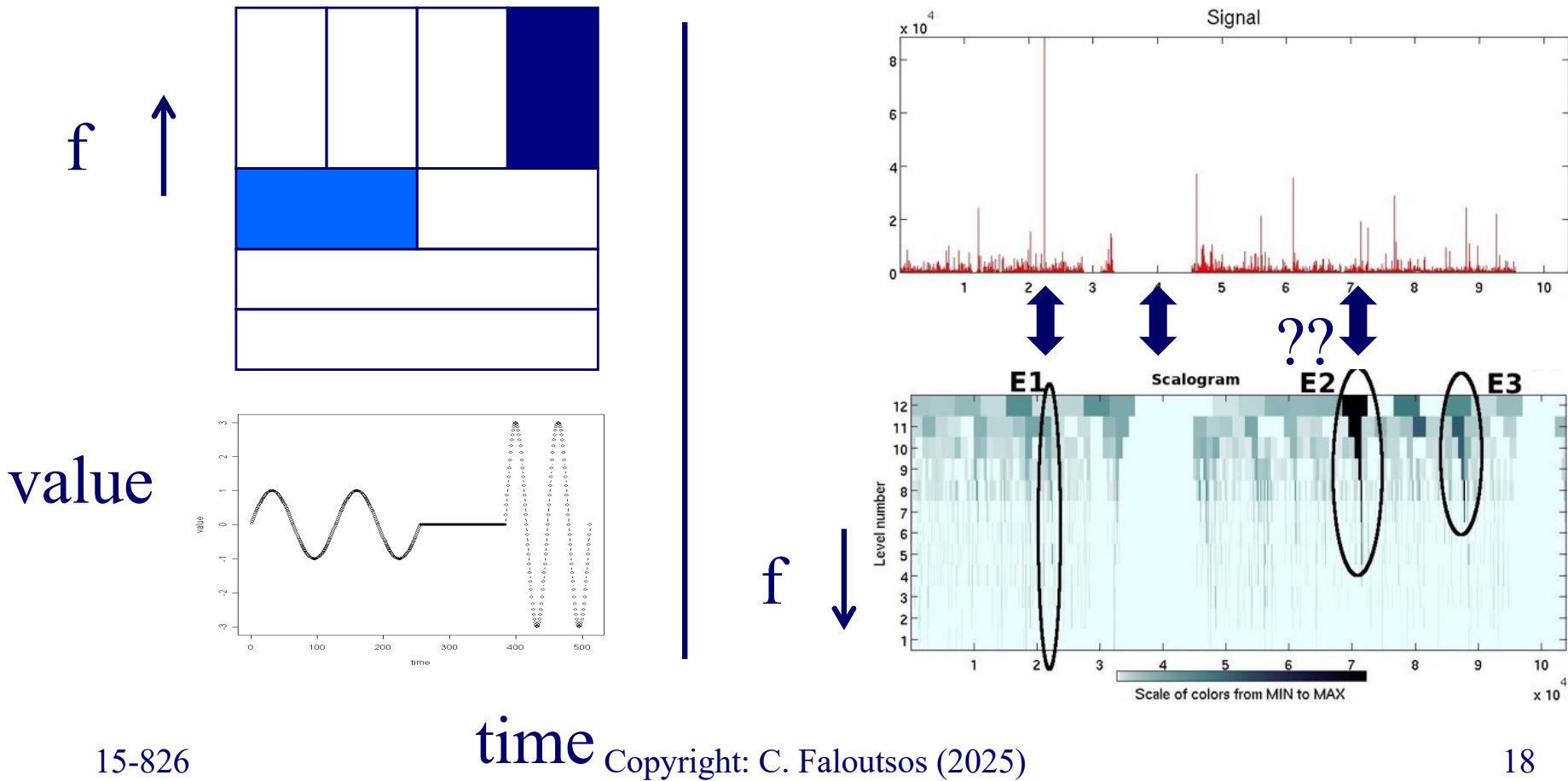


DFT



Wavelets

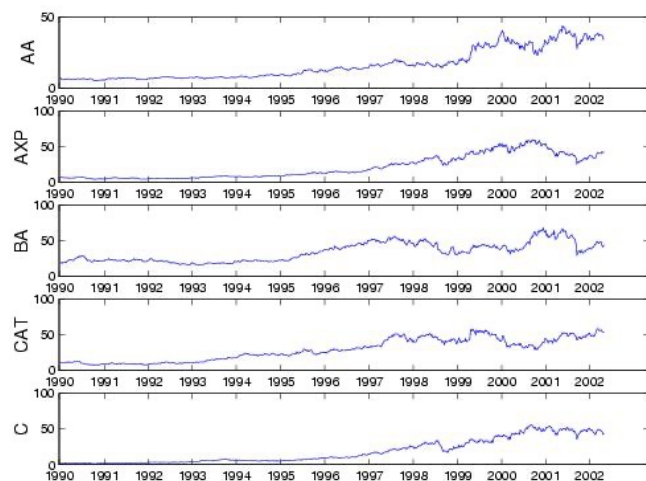
- Q: baritone/silence/soprano - DWT?





Forecasting:

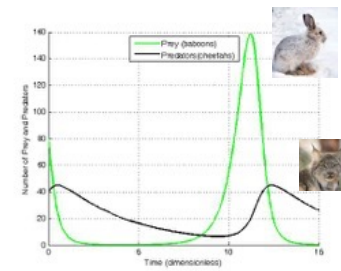
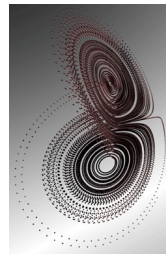
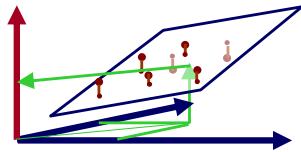
Q: mine/forecast (one, or more)
time sequences





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**





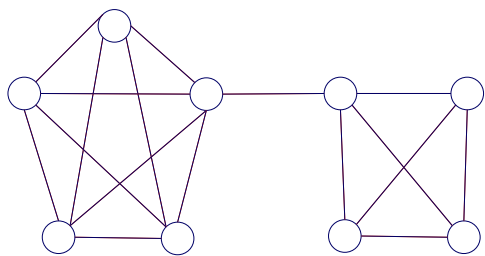
Outline

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

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

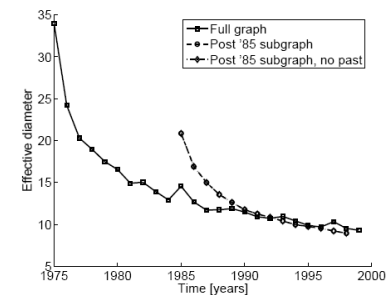
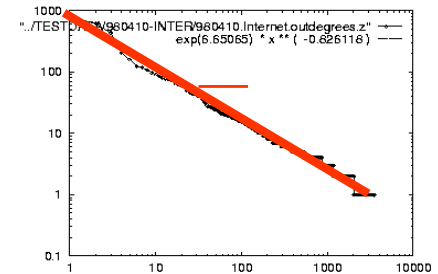
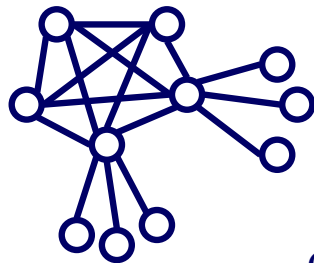
Graphs

- Real graphs: surprising patterns
 - ??



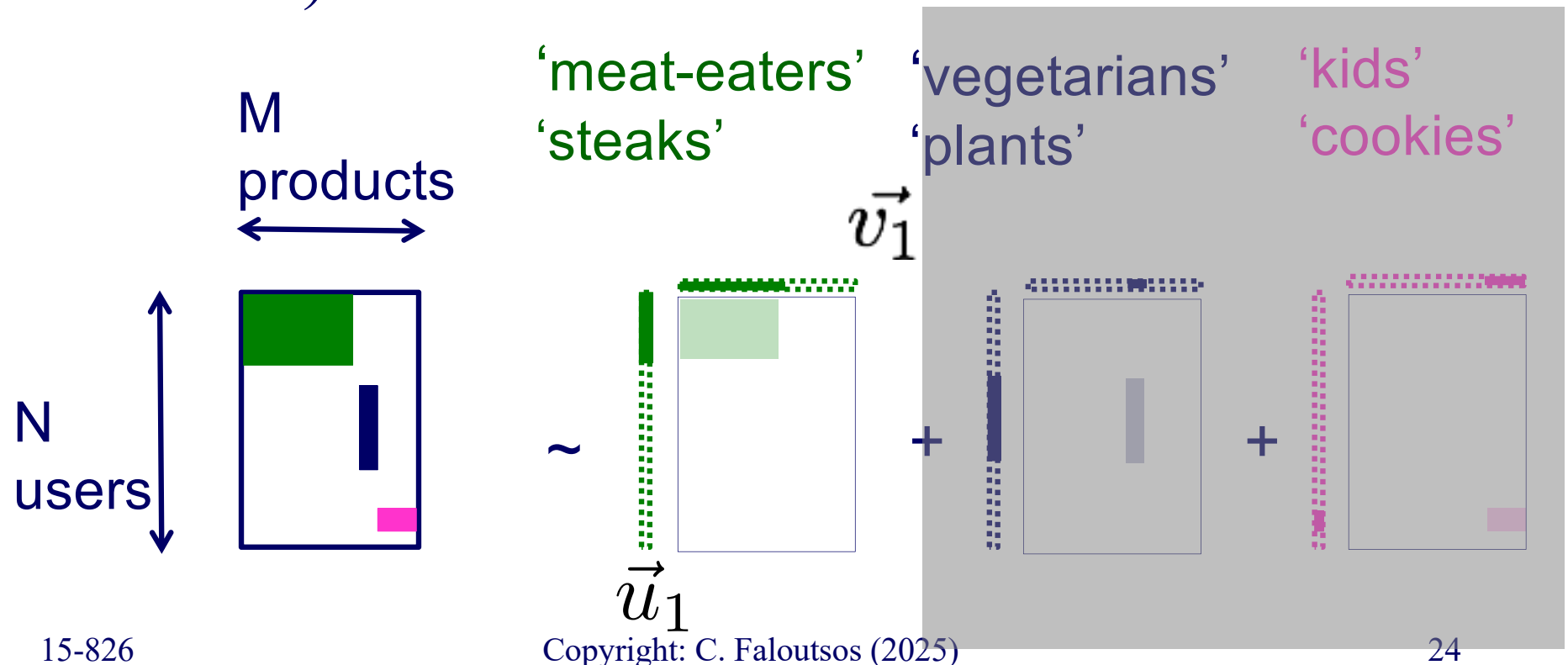
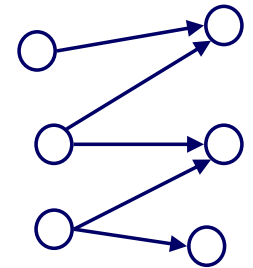
Graphs

- 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~~



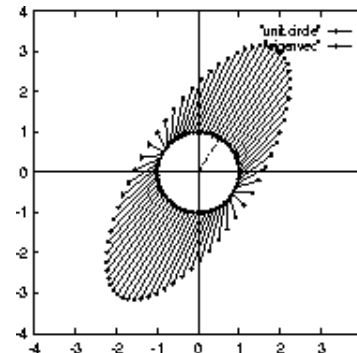
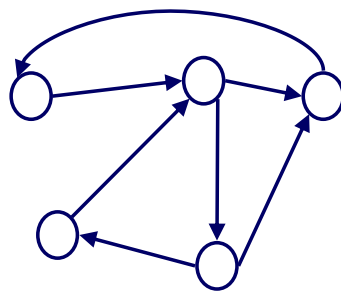
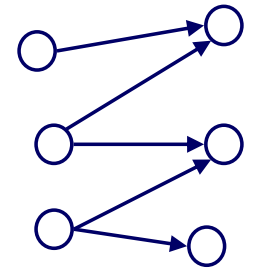
Graphs - SVD

- Hubs/Authorities (SVD on adjacency matrix)



Graphs - PageRank

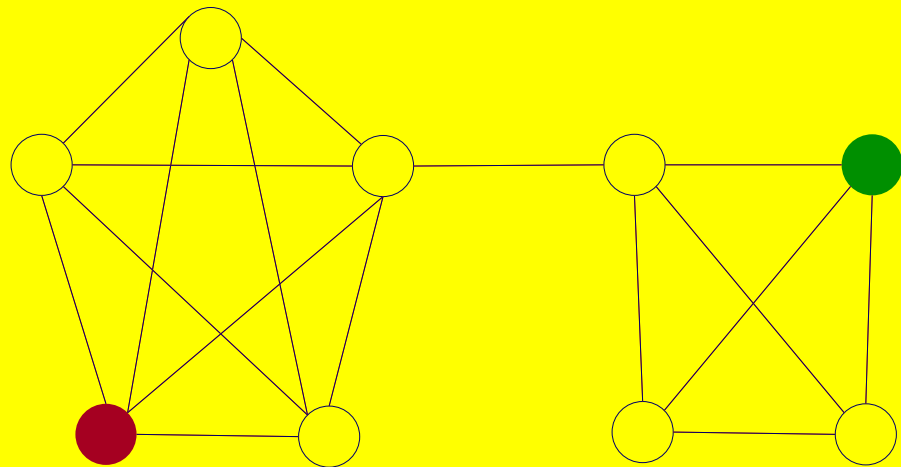
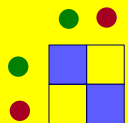
- Hubs/Authorities (SVD on adjacency matrix)
- PageRank (fixed point $\rightarrow$ eigenvector)



Not in the final exam

Belief Propagation

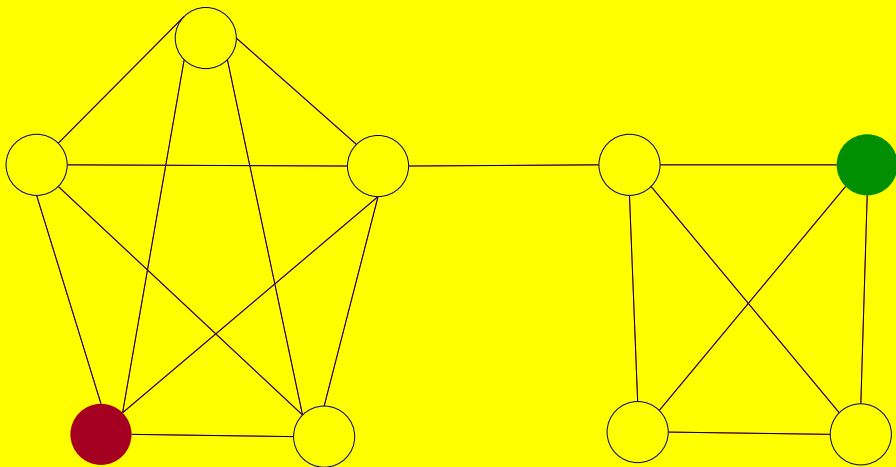
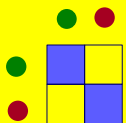
- What color, for the rest?



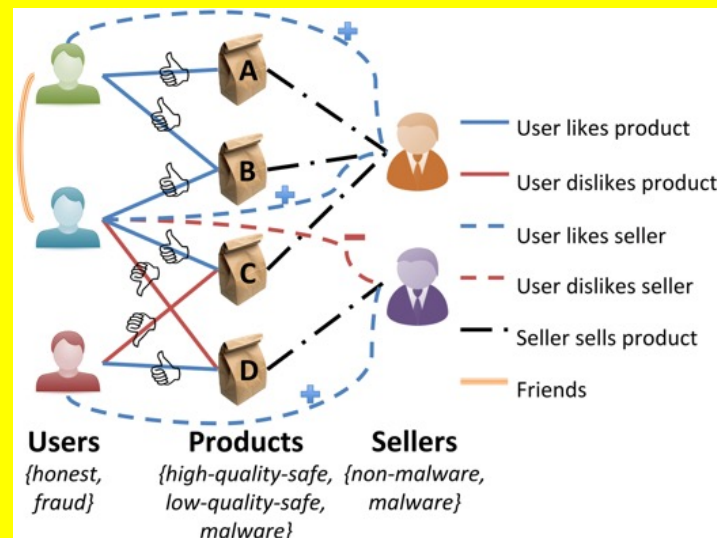
Not in the final exam

Belief Propagation

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

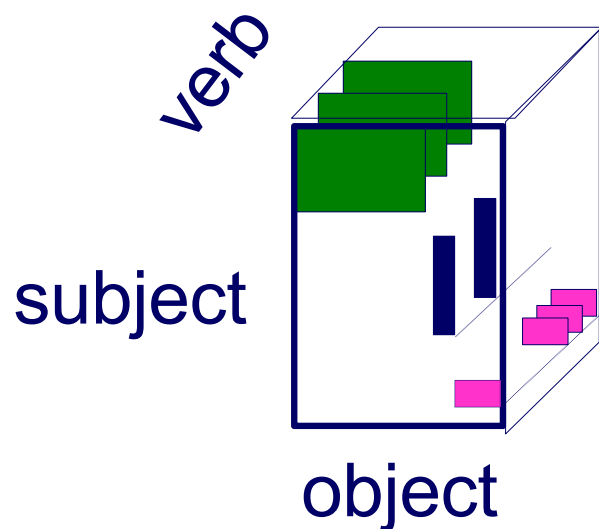


www.cs.cmu.edu/~deswaran/code/zoobp.zip



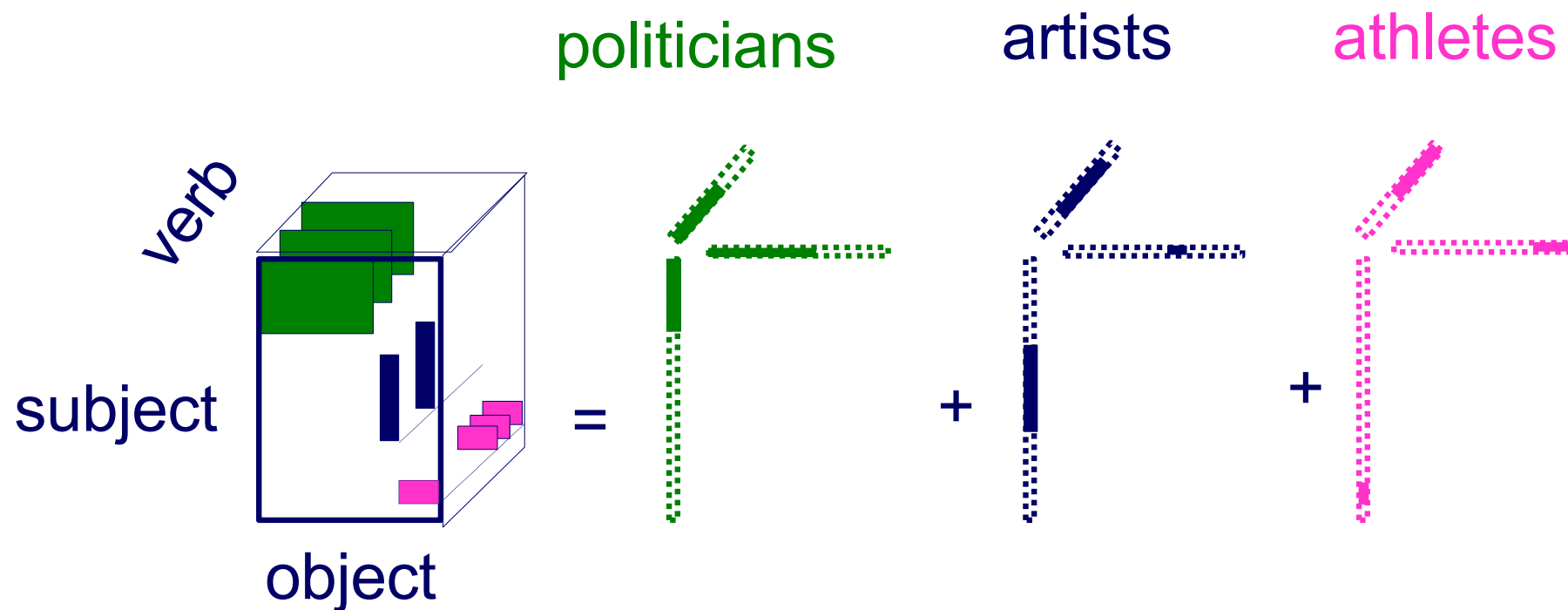
Tensors

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



Tensors

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

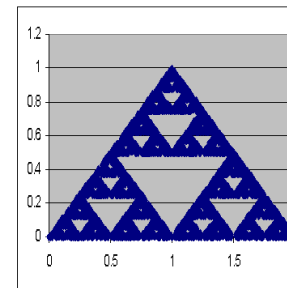


Taking a step back:

We saw some fundamental, recurring concepts and tools:

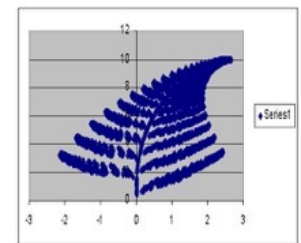
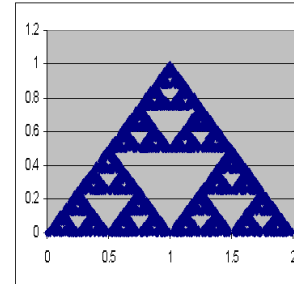
T1: Powerful, recurring tools

- Fractals/ self similarity

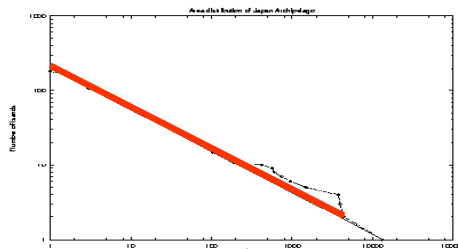


T1: Powerful, recurring tools

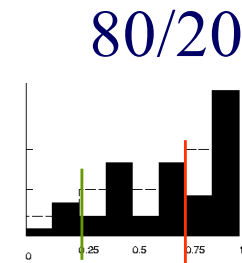
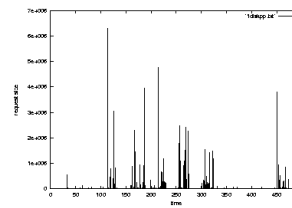
- Fractals/ self similarity $\leftrightarrow$ Power laws
 - Zipf, Korcak, Pareto's laws
 - intrinsic dimension (Sierpinski triangle)
 - correlation integral
 - Barnsley's IFS compression
 - ~~Kronecker graphs~~



15-826



Copyright: C. Faloutsos (2025)



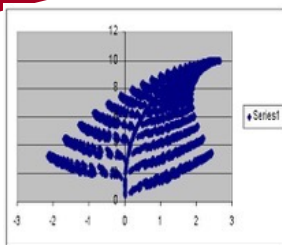
34

T1: Powerful, recurring tools

- Fractals/ self similarity

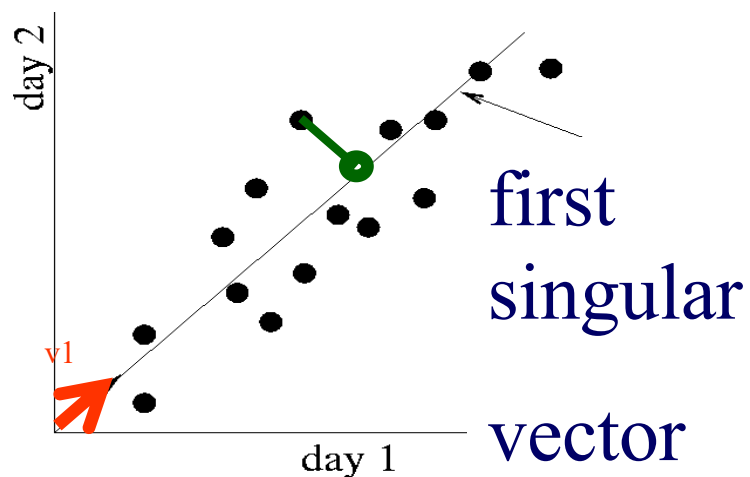
- Zipf, Ker

- **‘Take logarithms’**
- **mean $\rightarrow$ meaningless**
- **Gaussian trap**



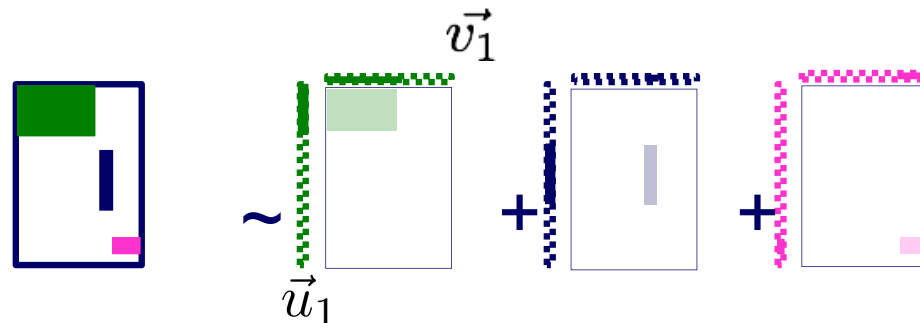
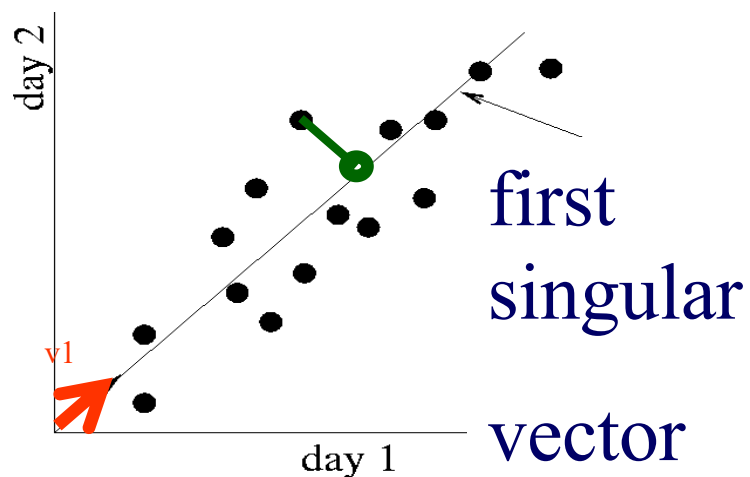
T2: Powerful, recurring tools

- SVD (optimal L2 approx)



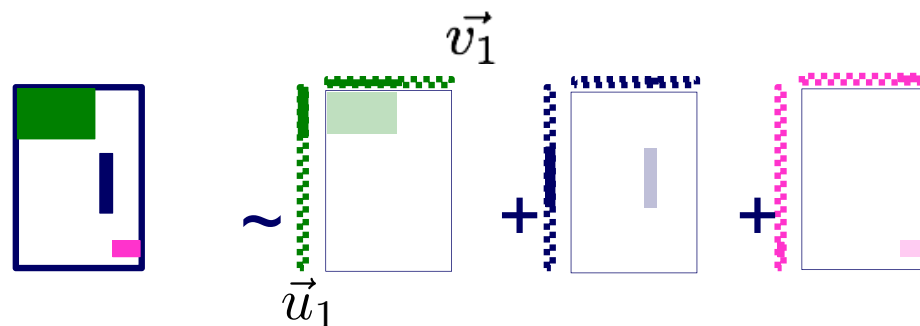
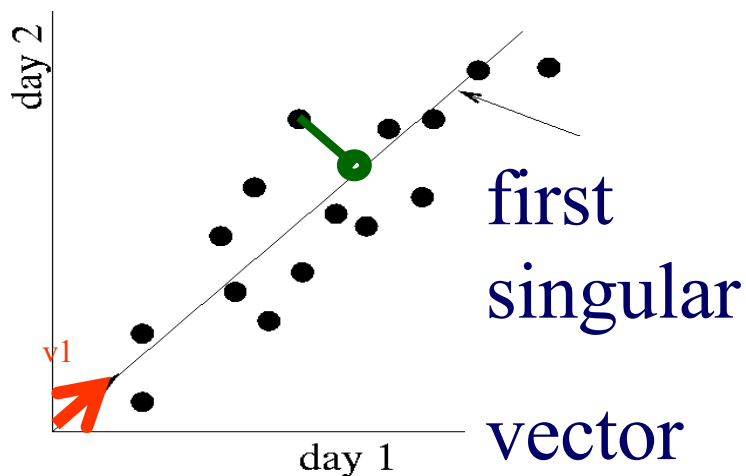
T2: Powerful, recurring tools

- Q: Cases we have a matrix as input?
- A: ...



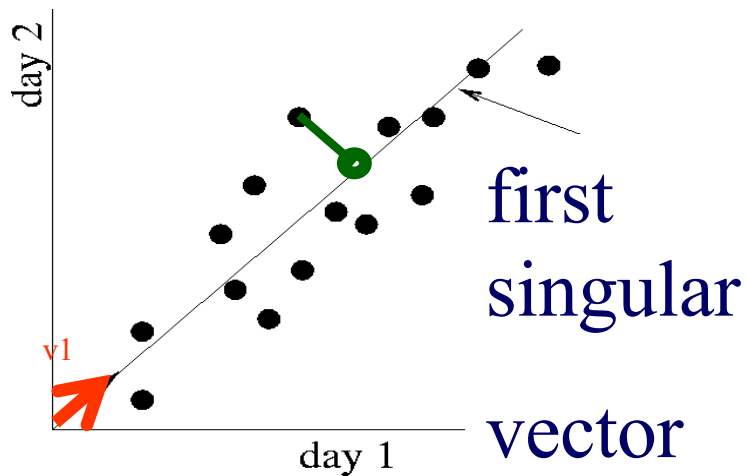
T2: Powerful, recurring tools

- Q: Cases we have a matrix as input?
- A1: graphs
- A2: co-evolving time sequences
- A3: entities in feature space



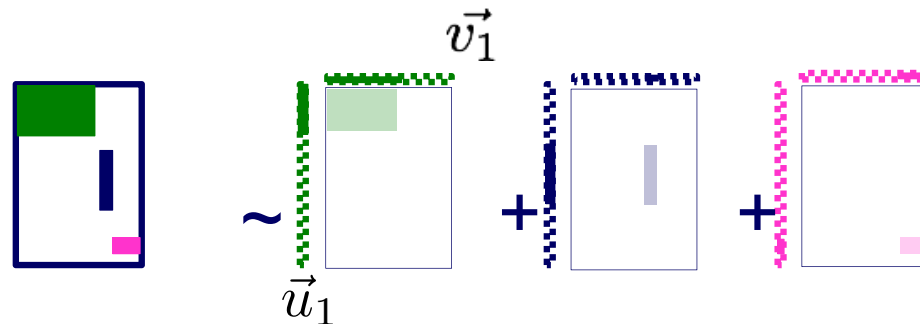
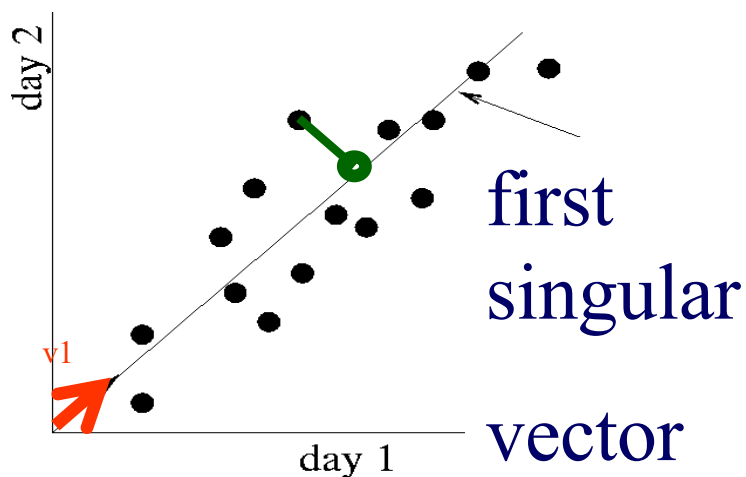
T2: Powerful, recurring tools

- SVD (optimal L2 approx)
- Algorithms in course, where SVD worked?



T2: Powerful, recurring tools

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



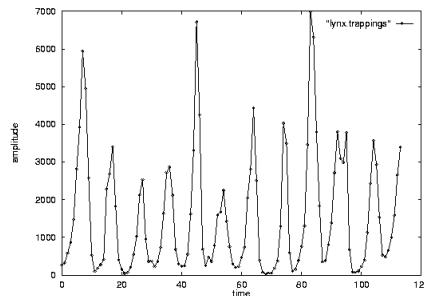
T3: powerful, recurring tools

DFT (Discrete Fourier Transform)

DWT (Discrete Wavelet Transform)



count



year

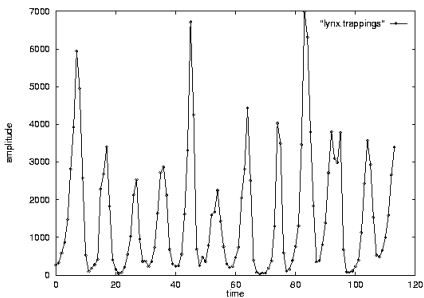
T3: powerful, recurring tools

DFT (Discrete Fourier Transform)

DWT (Discrete Wavelet Transform)

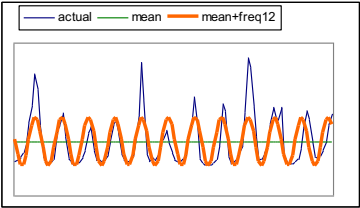


count

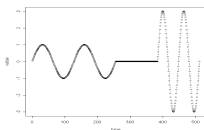
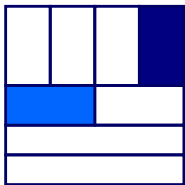


year

A1: Fourier (DFT)

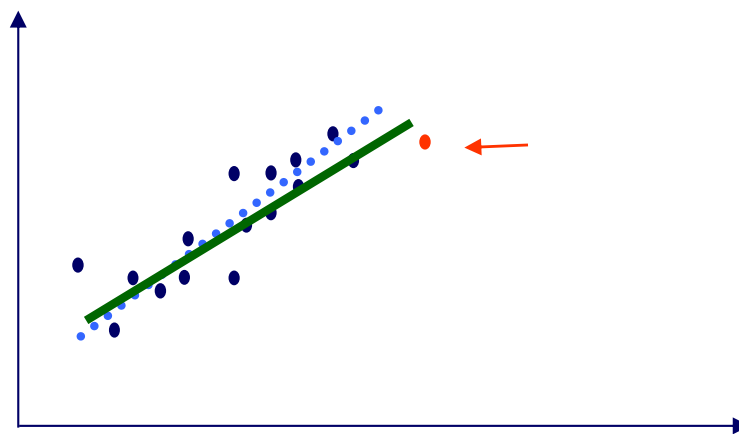


A2: Wavelets (DWT)



T4: Powerful, recurring tools

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





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

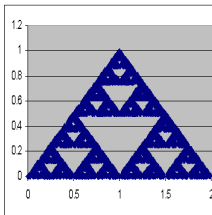


Summary of summary

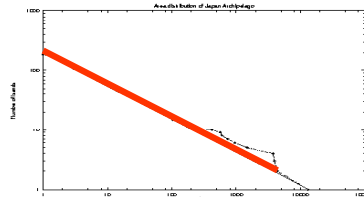
- **T1: fractals / power laws** lead to surprising discoveries
 - ‘the mean is meaningless’
- **‘Take logarithms’**
 - **mean $\rightarrow$ meaningless**
 - **Gaussian trap**
- **T2: low-rank/HITS/tensors/...**
- **T3: wavelets:** Nature seems to prefer them
- **T4: RLS:** matrix inversion, without inverting

Thank you!

- Feel free to contact me:
 - Cell#; christos@cs; GHC 7003
- Reminder: faculty course eval's:
 - <http://www.cmu.edu/hub/fce/>
- Have a great holiday!



15-826



Copyright: C. Faloutsos (2025)

• ‘Take logarithms’
• mean $\rightarrow$ meaningless
• Gaussian trap

46