# Chapter 12.

# Finding fragments of orders, partial orders, and total orders from 0-1 data



## Themes of the chapter

- Given a 0/1 a matrix
- Rows: observations, columns variables
- Can one find ordering information for the observations?
- Without additional assumptions, no; with some assumptions, yes
- Paleontological application:
  - find orders for subsets of fossil sites
  - a good ordering for (a subset of) the rows is one where the 1s are consecutive
- Also other applications

## Themes of the chapter

- Finding small total orders (fragments) from 0-1 data
  - Local models/patterns
- Finding partial orders from 0-1 data
  - A global model
- Find total orders for 0-1 data
  - A global model



## Finding small total orders (fragments) from 0-1 data

- Model: a subset of observations and a total order on the subset
- Task: find **all** such models fulfilling certain criteria
- Algorithm: a pattern discovery algorithm (levelwise search)

## Finding partial orders from 0-1 data

- Model: a partial order over all observations
- Loglikelihood: proportional to the number of cases the observed occurrence patterns violate the continuity of species
- Prior: prefer partial orders that are as specific as possible
- Task: find **a** model with high likelihood * prior
- Algorithm: Find fragments and use heuristic search to build a good partial order



## Find total orders for 0-1 data

- Model: a total order
- Loglikelihood: how many cases the observed occurrence patterns violate the continuity of species
- Task: find **the** best total order for the observations
- Algorithm: spectral method

# Type of data

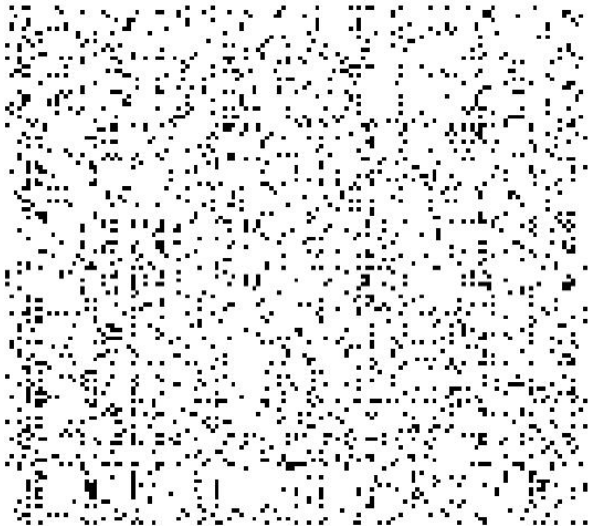
- 0-1 data, large number of variables
- Examples:
  - Occurrences of words in documents
  - Occurrences of species in paleontological sites
  - Occurrence of a particular motif in a promoter region of a gene
- Typically the data is sparse: only a few 1s
- Asymmetry between 0s and 1s
  - A ”1” means that there really was something
  - A ”0” has less information (in a way)



# Example

- Paleontological data from the NOW (Neogene Mammal Database)
- Fossil **sites** (one location, one layer)
- Each site contains fossils that are about the same age (+- 1 Ma)
- Variables: species/genera
- A ”1” is reasonably certain
- A ”0” might be due to several reasons
  - The species was not extant at that time
  - The remains did not fossilize
  - The tooth was overlooked
  - …

# Site-genus -matrix






# Background knowledge

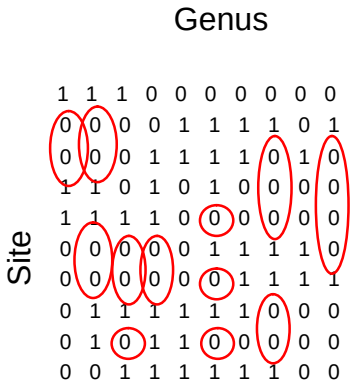
- Species do not vanish and return
- An ordering of the sites with a ”0” between ”1”s is improbable

0
0
1
1
1
0
1
0
0
1
1

time

# Example: seriation in paleontological data

- Given data about the occurrences of genera in fossil sites
- Want to find an ordering in which occurrences of a genus are consecutive
- **Lazarus count**: how many 0s are between 1s





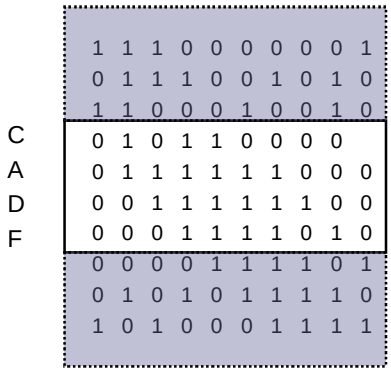

# A better ordering

A smaller Lazarus count

1 1 1 0 0 0 0 0 0 0
1 1 1 1 0 0 0 0 0 0
1 1 0 1 0 1 0 0 0 0
0 1 0 1 1 0 0 0 0 0
0 1 1 1 1 1 1 0 0 0
0 0 1 1 1 1 1 1 0 0
0 0 0 1 1 1 1 0 1 0
0 0 0 0 1 1 1 1 0 1
0 0 0 0 0 1 1 1 1 0
0 0 0 0 0 0 1 1 1 1

## Find small total orders (fragments) from 0-1 occurrence data

- Fragment: a total ordering of **a subset** of observations
- E.g., C<A<D<F
- Intuitive interpretation:
- For most variables the sequence of observations has no pattern of the form …1…0…1…






## Fragments of order

- 0/1 data set
- Fragment of order *f* is a sequence of observations
  $A_1 < A_2 < A_3 < \ldots < A_k$
- An variable *t* **disagrees** with fragment *f*, if for some $i<j<k$ we have $A_i=A_k=1$, but $A_j=0$
- Otherwise t **agrees** with f:
  $A_1 < A_2 < A_3 < \ldots < A_k$
- Then the column for *t* has the form
  0 0 …0 0 1 1 … 1 1 0 0 … 0 0
  for the observations in *f*

## Example

| | | | | |
|---|---|---|---|---|
| A | 1 | 0 | 0 | 1 |
| B | 1 | 1 | 1 | 0 |
| C | 0 | 0 | 0 | 1 |
| D | 1 | 0 | 1 | 0 |
| E | 1 | 0 | 1 | 1 |
| F | 1 | 1 | 1 | 1 |

| | | | | |
|---|---|---|---|---|
| A<B<C<D: | dis<br>1101 | ag<br>0100 | dis<br>0101 | dis<br>1010 |
| B<D<F<A: | ag<br>1111 | dis<br>1010 | ag<br>1110 | ag<br>0011 |



## What is a good fragment of order?

- A sequence f of rows, say, A < D < B < C
- Da(f): the number of variables disagreeing with the ordering
- Fr(f): the number of variables having at least 2 ones in the rows of f
- A good fragment has high Fr(f) and low Da(f)

## What is a good fragment of order?

We want to find orders such that most variables agree with the order

| | | | |
|---|---|---|---|
| A | 1 | 0 | 0 |
| B | 1 | 1 | 1 |
| C | 0 | 0 | 1 |
| D | 1 | 0 | 0 |
| E | 1 | 1 | 1 |
| F | 1 | 1 | 1 |

A < B < C < D
2 variables agree
1 disagrees

variable disagrees with an ordering: a Lazarus event



## What is a good fragment of order?

A < B < C < D
5 observations agree,
0 disagree

| | | | | |
|---|---|---|---|---|
| A | 1 | 0 | 1 | 0 |
| B | 1 | 1 | 0 | 1 |
| C | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 0 |
| E | 1 | 1 | 1 | 1 |
| F | 1 | 1 | 1 | 1 |

But this ordering is not very informative

Only 1 observation could have disagreed!

# What is a good fragment of order?

**Frequency** Fr(f) of a fragment: number of variables such that there are at least 2 ones in the variables of the ordering

| | | | | |
|---|---|---|---|---|
| A | 1 | 0 | 1 | 0 |
| B | 1 | 1 | 0 | 1 |
| C | 0 | 0 | 1 | 0 |
| D | 0 | 1 | 1 | 0 |
| E | 1 | 1 | 1 | 1 |
| F | 1 | 1 | 1 | 1 |

Fr(A<B<C) = 2

Fr(A<B<C) = Fr(A<C<B)

Fr(E<C<D) = 3

Fr(A<C<D) = 1

Frequency Fr(f) does not depend on the order of the observations in the fragment



# What is a good fragment of order?

Da(f): number of variables that disagree with fragment f

| | | | | |
|---|---|---|---|---|
| A | 1 | 0 | 1 | 1 |
| B | 1 | 1 | 1 | 1 |
| C | 0 | 1 | 0 | 1 |
| D | 1 | 0 | 0 | 0 |
| E | 1 | 1 | 1 | 1 |
| F | 1 | 1 | 0 | 0 |

Da(A<B<C) = 0

Da(A<C<D<F) = 2

## What is a good fragment of order?

A good fragment f has

high Fr(f)
low Da(f)

| A | B | C | D |
|---|---|---|---|
| 0 | 1 | 1 | 0 |
| 1 | 0 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 0 | 1 | 1 | 0 |
| 0 | 0 | 1 | 1 |



## Problem statement

- Given thresholds $\sigma$ and $\gamma$
- Find all fragments of order f such that in the data

  $Fr(f) > \sigma$

  $Da(f) < \gamma$

- Find all submatrices that contain at least $\sigma$ rows and are within $\gamma$ of having the consecutive ones property

## The definition has problems

Any fragment f of {A,B,C,D,E,F} has

Fr(f)=3 and
Da(f)=0.

A good ordering has to stand above its peers.

| | | | |
|---|---|---|---|
| A | 1 | 1 | 1 |
| B | 1 | 1 | 1 |
| C | 1 | 1 | 1 |
| D | 1 | 1 | 1 |
| E | 1 | 1 | 1 |
| F | 1 | 1 | 1 |



## Peers of a fragment

- Peers of a fragment f = A<B<C<D:
  - all permutations of the observations occurring in f
- The fragment f and its peer g have Fr(f)=Fr(g)
- A good fragment f has smaller Da(f) than its peers
- Da(f) = Da($f^R$): a fragment and its reverse have the same number of disagreeing variables

## Problem statement

- Given thresholds $\sigma$ and $\gamma$
- Find all fragments of order f such that in the data

  $Fr(f) > \sigma$

  $Da(f) < \gamma$

- and the fragment has smaller Da value than its peers



## How to avoid noninformative rows

D does not contribute in any way to A<B<C<D

| | | | | |
|---|---|---|---|---|
| A | 1 | 1 | 1 | 1 |
| B | 1 | 1 | 1 | 1 |
| C | 1 | 1 | 1 | 1 |
| D | 0 | 0 | 0 | 0 |

Require that all subfragments h satisfy $Fr(h) > \sigma$ and $Da(h) > \gamma$

h=C<D does not satisfy $Fr(h) > \sigma$

## Problem statement

- Given thresholds $\sigma$ and $\gamma$
- Find all fragments of order f such that in the data

  $Fr(f) > \sigma$

  $Da(f) < \gamma$

- and all subfragments of f satisfy these
- and the fragment has smaller Da value than its peers



## Algorithm

- How to find fragments with the specific properties?
- Start from fragments of length 2
  - No disagreements are possible
  - Only the bound $Fr(f) > \sigma$ to be tested
- Iteration:
- Assume fragments of length k-1 are known
- Then we can build candidate fragments of length k
- Continue until no new patterns are found
- A complete algorithm: all fragments will be found

## Monotonicity property

- Fragment $A_1 < A_2 < A_3 < \ldots < A_k$ can satisfy the requirements only if all subfragments of length k-1 satisfy them:

  $A_1 < A_2 < A_3 < \ldots < A_{k-2} < A_{k-1}$
  $A_1 < A_2 < A_3 < \ldots < A_{k-2} < A_k$
  $\ldots$
  $A_2 < A_3 < \ldots < A_{k-1} < A_k$

- All these have to be in the collection of fragments of size k-1



## Algorithm

- Find F2, fragments of size 2
- C = all triples A<B<C such that A<B, A<C, and B<C are in F2
- k ß 3

- While C is not empty
  compute Da(f) for all f in C
  Fk ß {f in C | Fr(f)> σ and Da(f)< γ }
  k ß k+1
  C ß all fragments of length k such that all the subfragments of length k-1 are in Fk

## Complexity of the algorithm

- Potentially exponential in the number of variables
- |F+C| = the size of the answer + all the candidates
- Proportional to
  $$|F+C| \; n \; m$$
  for a matrix with n rows and m columns
- Too low values of $\sigma$ or too high values of $\gamma$ will lead to huge outputs

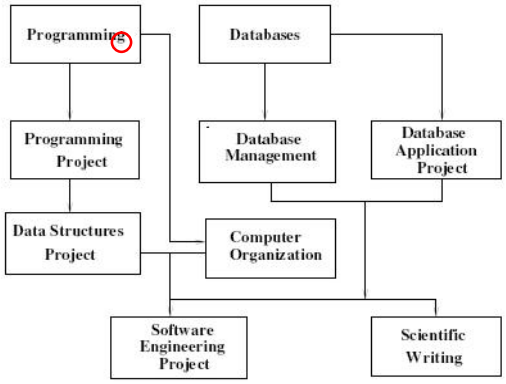


## Experimental results

- Data about students and courses
- Observations: students
- Variables: courses
- D(s,c)=1 if student s has taken course c
- Here we know the true ordering
  - Or actually two: official ordering
  - Real order in which the student took the courses

# Part of the recommendations

Discovered fragment f
Fr(f)=1361, Da(f)=3.2%




⟨Programming, Computer Organization, Programming Project, Data Structures Project, Scientific Writing⟩



# Results

| $\sigma$ (in %) | $\tau$ (in %) | $Max\ l$ | $\lvert\mathcal{T}\rvert$ | $\alpha$ (in %) | $\beta$ (in %) |
|---|---|---|---|---|---|
| 20 | 0 | 3 | 2 | 96.3 | 99.5 |
| 20 | 2.5 | 5 | 578 | 48.6 | 70.5 |
| 20 | 5 | 6 | 1528 | 40.0 | 66.0 |
| 15 | 0 | 3 | 28 | 89.9 | 98.6 |
| 15 | 2.5 | 6 | 1934 | 46.8 | 78.2 |
| 15 | 5 | 7 | 5158 | 38.9 | 72.3 |

## Results (paleontological data)

- Fragments for sites
- Or transpose the matrix: fragments for species
- Sequences of sites such that there are very few Lazarus events
- Provide ways of looking at projections of the data
- Can be used to find partial orders



## Results (cont.)

- What does one do with the results?
- Given a single fragment f, are Fr(f) and Da(f) somehow unusual?
- Approaches like DuMouchel & Pregibon, ”Empirical Bayes Screening for Multi-Item Associations”, 2001

## Example: words in documents

- Represent collections of documents as term vectors
- Which words occur (1) in the document or not (0)
- Very large dimensionality, lots of observations



## Example from Citeseer

| "database system" | "query" | "selectivity estimation" | Hits |
|---|---|---|---|
| 1 | 1 | 1 | 49 |
| 1 | 1 | 0 | 1930 |
| 0 | 1 | 1 | 221 |
| 1 | 0 | 1 | 4 |

What does this tell us about these terms?
*Databases* and *selectivity estimation* together do not occur without *queries*

*Databases < queries < selectivity estimation*

## Example from Google Scholar

- prior distribution – MCMC
  151,000 documents
- prior distribution MCMC
  2950 documents
- – prior distribution MCMC
  1050 documents
- prior – distribution MCMC
  165 documents

prior < distribution < MCMC



## Themes of the talk

- Find small total orders from 0-1 data
- Finding partial orders from 0-1 data
- Find total orders for 0-1 data

## Finding partial orders from 0-1 data

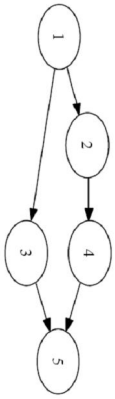
- Model: a partial order over all observations
- Loglikelihood: proportional to the number of cases the observed occurrence patterns violate the continuity of species
- Prior: prefer partial orders that are as specific as possible
- Task: find a model with high likelihood * prior
- Algorithm: Find fragments and use heuristic search to build a good partial order



## Why partial orders?

- Determining the ages of sites is difficult
- Radioisotope methods apply only to few sites
- In paleontology the so-called MN system: 18 classes for the last 25 Ma
- Classes are assigned by ad hoc methods
- Searching for a total order might not be a good idea
- The MN system is a partial order

## Finding partial orders from data

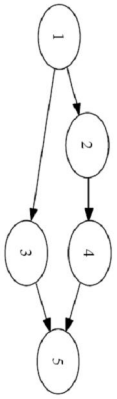



- How to find a partial order that fits well with the data?
- What does this mean?



## What is a good partial order?

- The Lazarus count of a species with respect to a partial order P:
  - For how many sites the species was extinct at the site, but extant before and after it (as determined by P)
  - The same definition as for total orders
- A good partial order has small Lazarus count
- Can be formulated as a likelihood (a Lazarus event is a false positive)




| | | | |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 1 | 1 | 0 |
| 3 | 0 | 0 | 1 |
| 4 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 |

Laz No Laz No Laz



# What is a good partial order?

- Find a partial order that has a low Lazarus count
- The trivial partial order has Lazarus count 0
- Want to find a partial order that is specific (close to a total order) and agrees with the data
- Measures of specificity:
  - the number of linear extensions of P (hard to compute)
  - number of edges in P
- Find a partial order that has high specificity * likelihood

## Algorithm for finding partial orders

- Compute fragments from the unordered data
- E.g., A < D < B < E < F and B < E < C
- Form a precedence matrix: in what fraction of the fragments does A precede B
- Form a partial order that approximates the precedence matrix (heuristic search)



## Fragments and reverse fragments

- The fragment generation will produce for each fragment f also its reverse $f^R$
- The pairwise precedence matrix would be useless
- Divide the fragments into two classes (graph cutting)
- Discard one class
- Build the precedence matrix

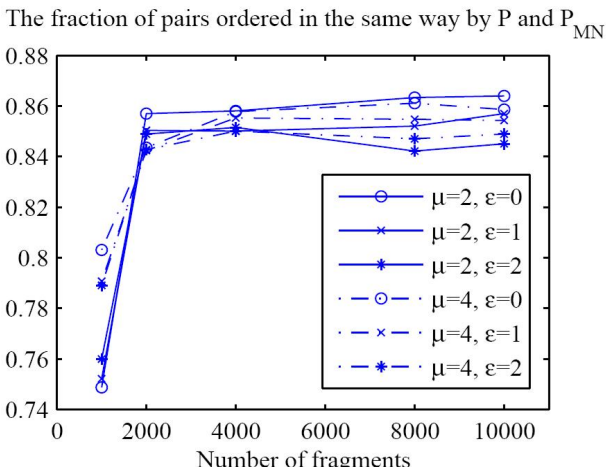
# From precedence matrix to partial order

- Heuristic search
- Add edges to the partial order so that the match with the precedence matrix improves
- Keep track of transitivity

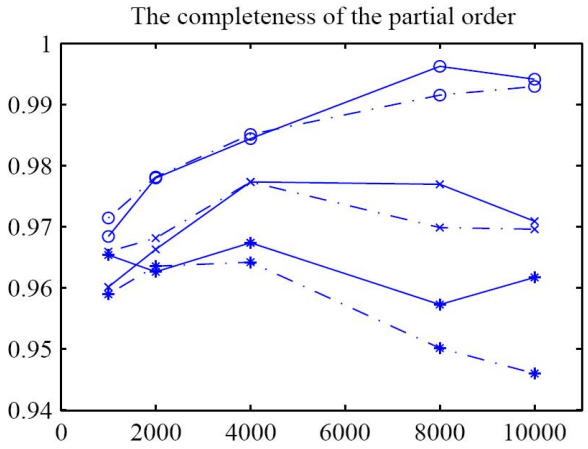
- Difficult (and interesting) algorithmic problem
- Empirical results look good

- Very recent theoretical results

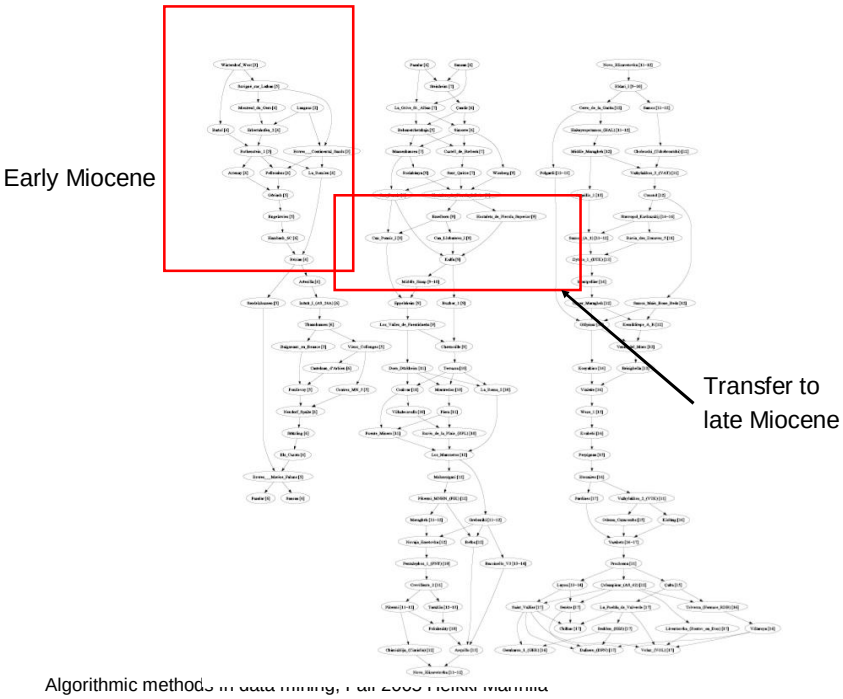


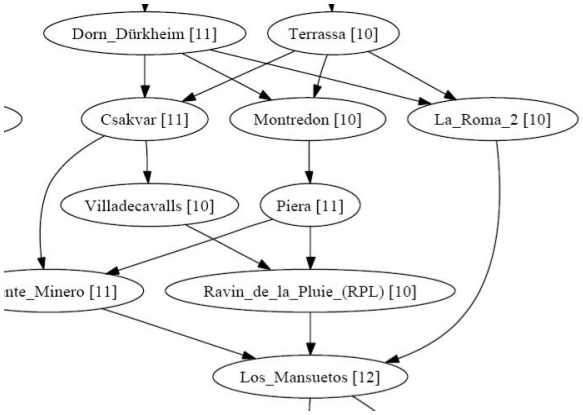



Algorithmic methods in data mining, Fall 2005 Heikki Mannila


The completeness of the partial order
1
0.99
0.98
0.97
0.96
0.95
0.94
0
2000
4000
6000
8000
10000




Early Miocene
Transfer to
late Miocene




approximately 6–7 MN classes of sites will be re-evaluated on the basis of the partial order.



# Themes of the talk

- Find small total orders from 0-1 data
- Finding partial orders from 0-1 data
- Find total orders for 0-1 data

## Finding good total orders for a matrix

- Given a site-genus matrix
- What is a good total ordering for the rows?
- One in which there are as few Lazarus events as possible
- Model class: total orders
- Loglikelihood proportional to the number of Lazarus events



## How to find such an ordering of the rows?

- If there is an ordering that has no Lazarus events, it can be found in linear time (Booth & Lueker)
  - consecutive ones property
- But normally there are (lots of) Lazarus events

## Finding good total orders for a matrix

- The problem of finding the best ordering of the matrix is NP-hard
- Finding whether there is a submatrix of size k that has no Lazarus events is NP-hard
- The fragment method finds such submatrices
- Local search, traveling salesperson approaches
- Spectral methods



## Spectral ordering for finding good total orders for a matrix

- Spectral ordering
- Compute a similarity measure *s(i,j)* between sites (e.g., dot product)
- Laplacian L(i,j)

$$L(i,j) = \begin{cases} -s(i,j), & i \neq j \\ \sum_k s(i,k), & i = j \end{cases}$$

- The eigenvector *v* corresponding to the second smallest eigenvalue of L satisfies

$$\sum_i v_i = 0, \quad \sum_i v_i^2 = 1, \text{ and } \sum_i s(i,j)(v_i - v_j)^2 = 1 \text{ is minimized.}$$

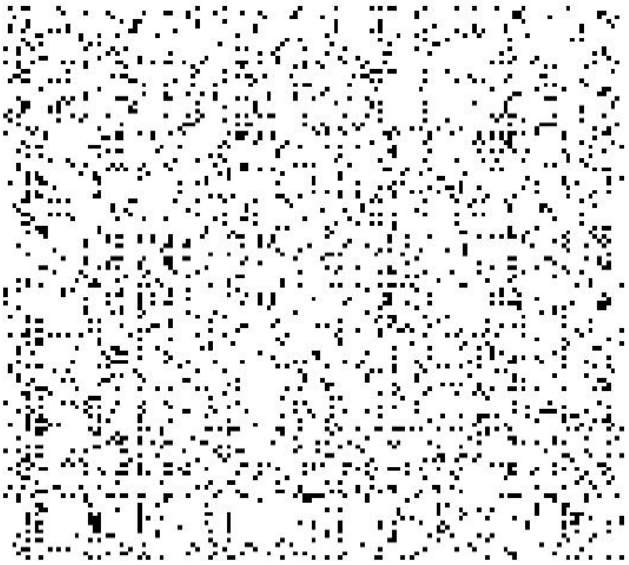
- Maps the points to 1-d, keeping similar points close to each other
- The values $v_i$ can be used to order the points



# Empirical observation

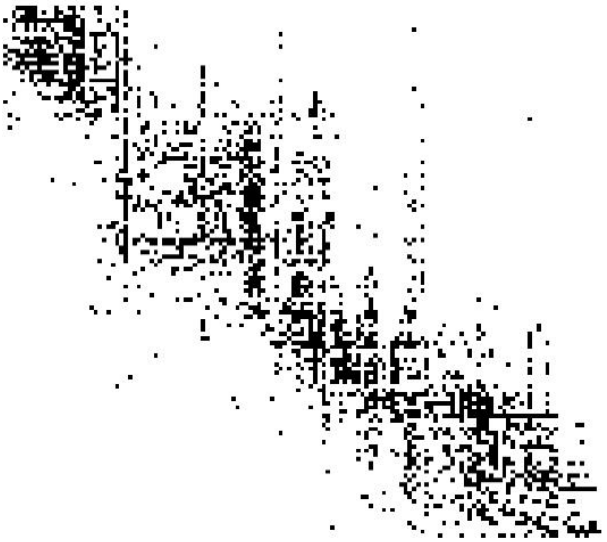
- The eigenvector seems to minimize also Lazarus events
- Even better than some combinatorial algorithms
- Why?
- No real theoretical understanding

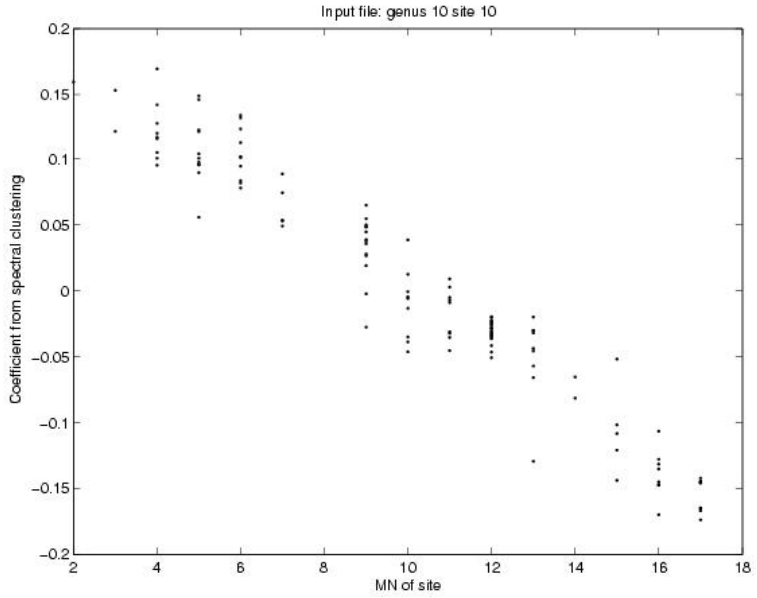
## Site-genus -matrix





## After spectral ordering






Fortelius, Jernvall, Gionis, Mannila, to appear in Paleobiology



| gl | sl | gn | sn | c | Nh | ch | NMN | cMN |
|---|---|---|---|---|---|---|---|---|
| 10 | 10 | 139 | 124 | 0.97 | 21 | 0.98 | 119 | 0.96 |
| 10 | 5 | 139 | 259 | 0.96 | 35 | 0.97 | 230 | 0.95 |
| 5 | 10 | 198 | 136 | 0.97 | 22 | 0.99 | 125 | 0.97 |
| 5 | 5 | 201 | 273 | 0.96 | 35 | 0.98 | 240 | 0.96 |
| 2 | 10 | 281 | 147 | 0.97 | 22 | 0.99 | 132 | 0.97 |
| 2 | 2 | 285 | 512 | 0.94 | 46 | 0.97 | 444 | 0.94 |

| gl | sl | Ls | LMN | Lage | Lazs | LazMN | Lazage |
|---|---|---|---|---|---|---|---|
| 10 | 10 | -4881 | -5153 | -4998 | 3792 | 4174 | 3974 |
| 10 | 5 | -9038 | -9573 | -9416 | 9728 | 10906 | 10563 |
| 5 | 10 | -6008 | -6455 | -6275 | 5220 | 5901 | 5622 |
| 5 | 5 | -10723 | -11340 | -11132 | 13003 | 14638 | 14147 |
| 2 | 10 | -6904 | -7429 | -7234 | 6398 | 7314 | 6969 |
| 2 | 2 | -16660 | -17610 | -17323 | 30568 | 34886 | 33621 |



# Questions

- Computational
  - Why does it work so well?
  - How well does it actually work (what is the smallest number of Lazarus events for this data?)
  - How to interpret the coefficients?
- Paleontological
  - Fully based on the occurrence matrix (excellent and bad)
  - Site-species data is only one type of data; how to use other types of data for the ordering?
  - …

## Rough estimates of the sizes of the model classes

- N observations
- Fragments of size at most k
  - $N^k$ individual fragments
  - $2^{N^k}$ sets of fragments
- Partial orders $2^{O(N^2)}$
- Total orders $N!$



## Concluding remarks

- General task: finding order from unordered data
- Here using species continuity as the additional information
- Other applications are possible

- Model classes
  - Fragments
  - Partial orders
  - Total orders

## Lots of open questions

- The unreasonable effectiveness of spectral methods on discrete optimization task
- Approximation guarantees
- Fragments from other applications
- MDL description of sequences via partial orders
- Etc.



## References


- A. Gionis, T. Kujala and H. Mannila: Fragments of order. *ACM SIGKDD 2003*, p. 129-136.
- A. Ukkonen, M. Fortelius, H. Mannila: Finding partial orders from unordered 0-1 data. *ACM SIGKDD 2005*, p. 285-293.
- M. Fortelius, A. Gionis, J. Jernvall, H. Mannila, Spectral Ordering and Biochronology of European Fossil Mammals, to appear in *Paleobiology*.