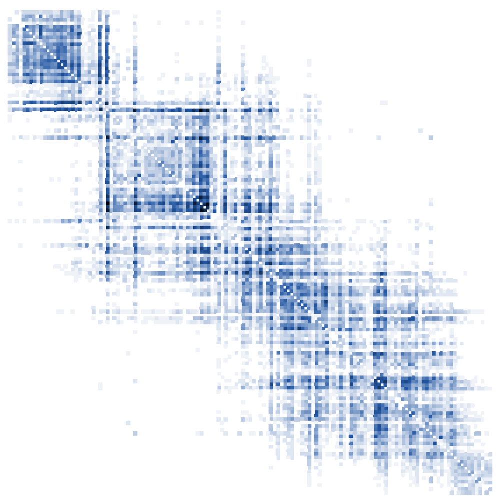
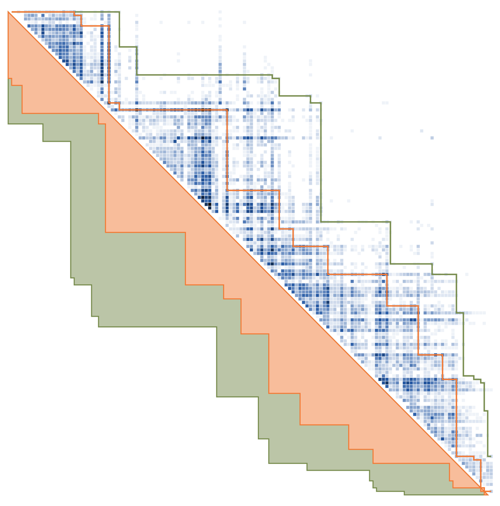
# DISCOVERING BANDS FROM GRAPHS

NIKOLAJ.TATTI@AALTO.FI
Aalto University, Helsinki Institute of Information Technology

## DISCOVERING BANDS

Many datasets have a band around the diagonal:

PROBLEM Given a(n adjacency) matrix, order entries and find $K-1$ bands

$$\emptyset = B_0 \subsetneq B_1 \subsetneq \cdots \subsetneq B_K = A$$

such that

- inner bands are more dense,

$$a(B_i) > a(B_{i+1})$$

- segments are homogenous; they minimize some score

$$\sum_{i=1}^{K} q(B_i \setminus B_{i-1}) \qquad \text{(e.g., } q = L_2\text{)}$$

## ALGORITHM IN A NUTSHELL

1. Find order
   - spectral heuristic
   - hill-climb refinement
2. Find all the borders
   - exactly with grid isotonic regression
   - approximate by iterating total order isotonic regression
3. Select $K$ borders that optimize the score

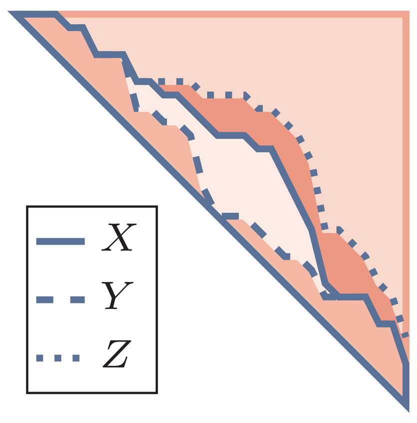
## BORDERS

$X$ is *not* a border if there are bands $Y$ and $Z$ such that

$$Y \subsetneq X \subsetneq Z$$

and

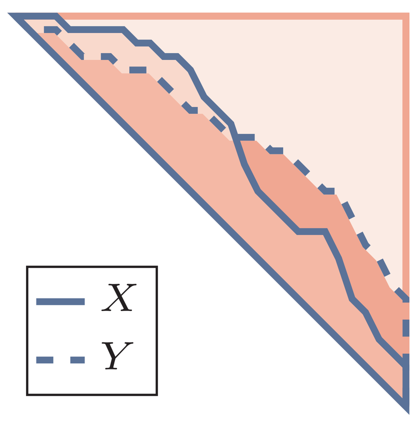
$$a(X \setminus Y) \le a(Z \setminus X)$$




otherwise, $X$ *is* a border

THEOREM There is an optimal solution that contains only borders

THEOREM
Given borders $X$ and $Y$, either $X \subseteq Y$ or $Y \subseteq X$.




COROLLARY All borders form a chain, $\emptyset = B_1 \subsetneq B_2 \subsetneq \cdots \subsetneq B_L = A$.
The density is decreasing $a(B_i) > a(B_{i+1})$.

COROLLARY There are at most $n(n-1)/2$ borders.

## DISCOVERING $K$ BANDS

Dynamic programming:

$$opt(i,k) = \text{optimal solution covering } B_i \text{ with } k \text{ bands} \quad .$$

Update equation:

$$opt(i,k) = \max_{j<i} q(B_i \setminus B_j) + opt(j, k-1) \quad .$$

## DISCOVERING BORDERS

Can be done with isotonic regression

PROBLEM For a DAG $G = (V, E, f)$ with vertex weights, find $g$ such that

$$g(v) \ge g(w) \quad \text{for every} \quad (v,w) \in E$$

and

$$\sum_{v \in V} |f(v) - g(v)|^2$$

is minimized.

To find borders, let

- vertices to be cells $V = \{(i,j) \mid i < j\}$.
- edges to be to the cells away from the diagonal,

$$E = \{(i,j) \to (i,j+1)\} \cup \{(i,j) \to (i-1,j)\} \quad .$$

- weights are values in adjacency matrix

THEOREM Given a border $B$, there is $t$ s.t.

$$B = \{(i,j) \mid g(i,j) \ge t\} \quad .$$

Needs $O(n^4)$ time, much less in practice.

## APPROXIMATING BORDERS

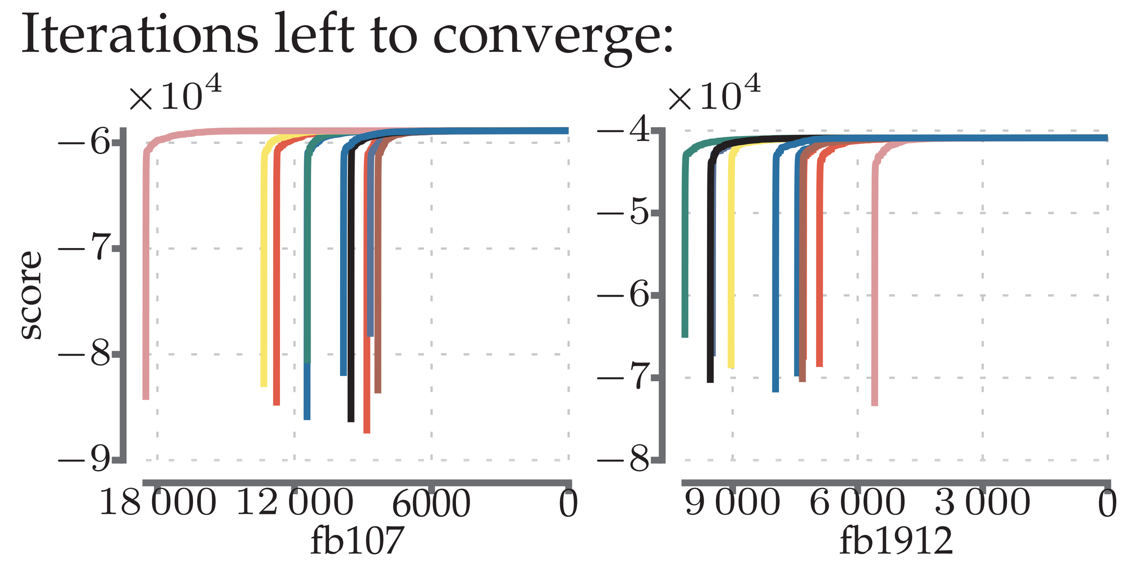
THEOREM There is an order for cells, that given a border $B$, there is $t$ such that

$$B = \{(i,j) \mid g(i,j) \ge t\},$$

where $g$ is a solution for total order isotonic regression.

- guess an order for cells
- solve total order isotonic regression
  - needs $O(m)$ time
- permute the cells within the borders
- repeat

Iterations left to converge:



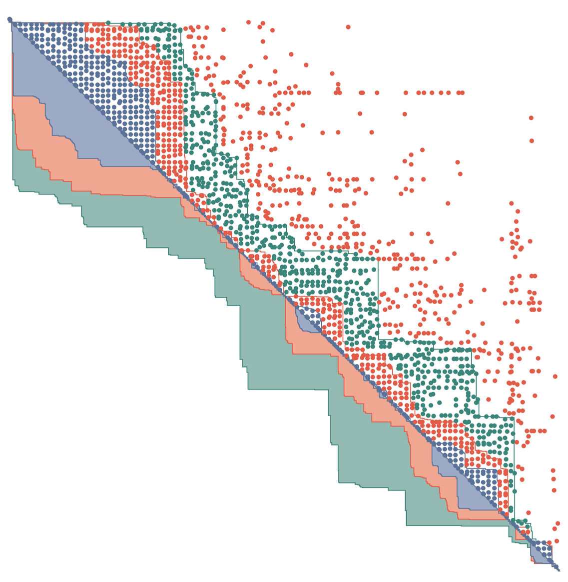
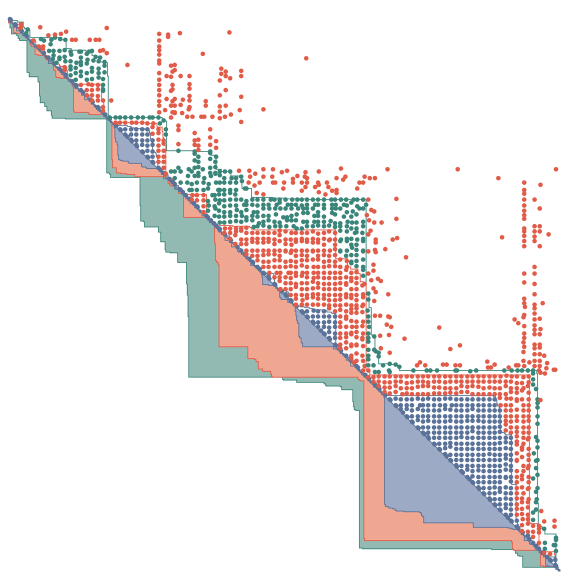
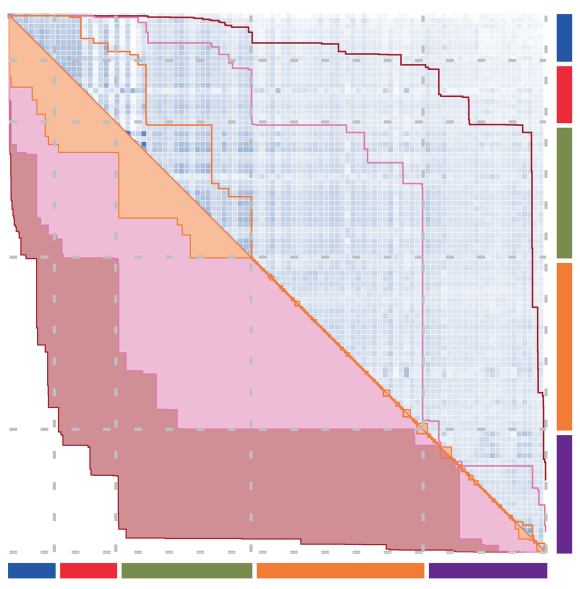
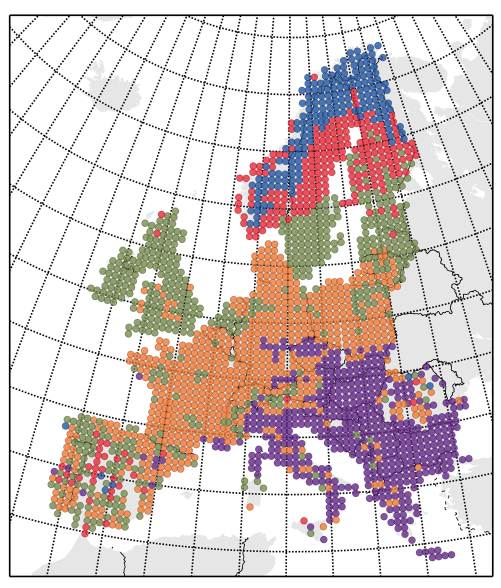

## FINDING ORDER

No fast approach to our knowledge.

Use heuristics:

- Fielder order:
  - order nodes using the 2nd smallest eigenvector of Lagrangian.
- refine order with hill climbing by swapping entries

## EXPERIMENTS

| | Approximate borders | | | | | | | Exact borders | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Name | time | brd | iter | rnd | ref | initial | final | time | initial | final |
| DblpCF | 0.2s | 55 | 235 | 48 | 3 | 945 | 908 | .02s | 945 | 905 |
| DblpCP | 0.4s | 53 | 701 | 135 | 3 | 966 | 927 | .05s | 966 | 918 |
| Fb107 | 12m | 476 | 7217 | 676 | 7 | 61 734 | 60 444 | 20s | 61 723 | 60 427 |
| Fb1912 | 5m | 375 | 7357 | 813 | 4 | 43 212 | 42 909 | 3.2s | 43 212 | 42 930 |
| Paleo | 4s | 201 | 423 | 51 | 4 | −8645 | −8906 | .13s | −8645 | −8906 |
| Mammals | 33m | 2975 | 2000 | 40 | | 19 798 | | 2m | 19 798 | 19 798 |
| Synthetic | 37m | 625 | 2000 | 40 | | 6 956 048 | | | | |