

Self-organized soft clustering, feature selection, and network inference using Gaussian processes



Christoph Best, Ralf Zimmer, Joannis Apostolakis
Institute for Informatics, LMU Munich, Germany

<http://www.bio.ifi.lmu.de/~best>

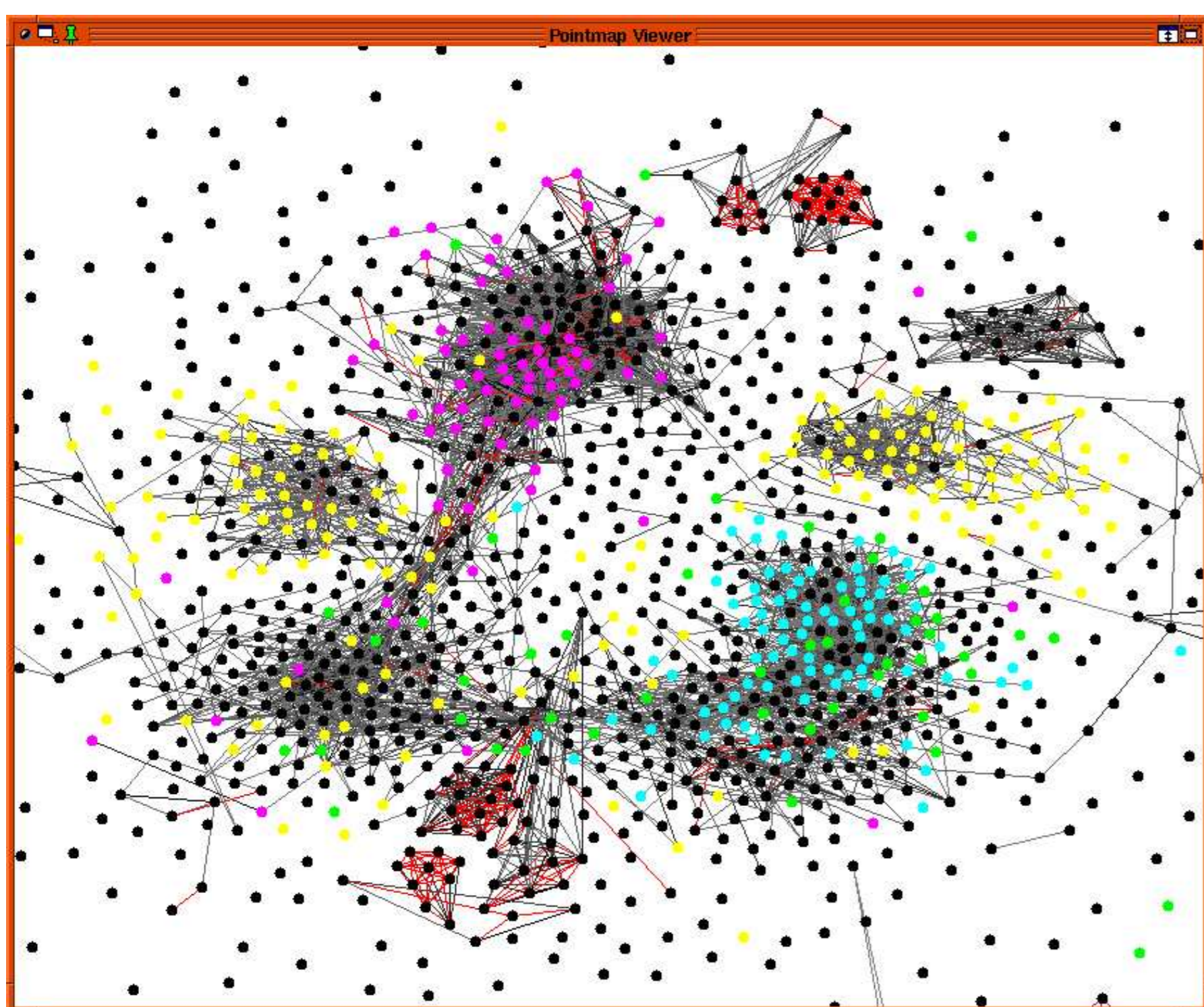
Soft clustering and feature selection for large data sets

METHOD

CLUSTERING = Mapping of objects into disjoint subsets such that similar objects appear in the same subset

MULTIDIMENSIONAL SCALING = Mapping of objects into a low-dimensional space (plane, cube) such that similar objects appear close to each other

- ▷ Data is mapped into a low-dimensional space by optimizing a target function that favors conserving similarity relations
- ▷ Reveals clusters, but in a continuous way \Rightarrow **VISUALIZATION**
- ▷ Proximity between clusters indicate similarities
- ▷ Different choices of distance and weighing functions allow different views of the data



Two-dimensional map of a subset of 1352 yeast genes from a knock-out compendium data set [Hughes et al., Cell 2000]. Lines connect similar genes. The colors represent different function assignments in the GO database.

ALGORITHM

We numerically minimize a target function Q to place genes and experiments on the plane:

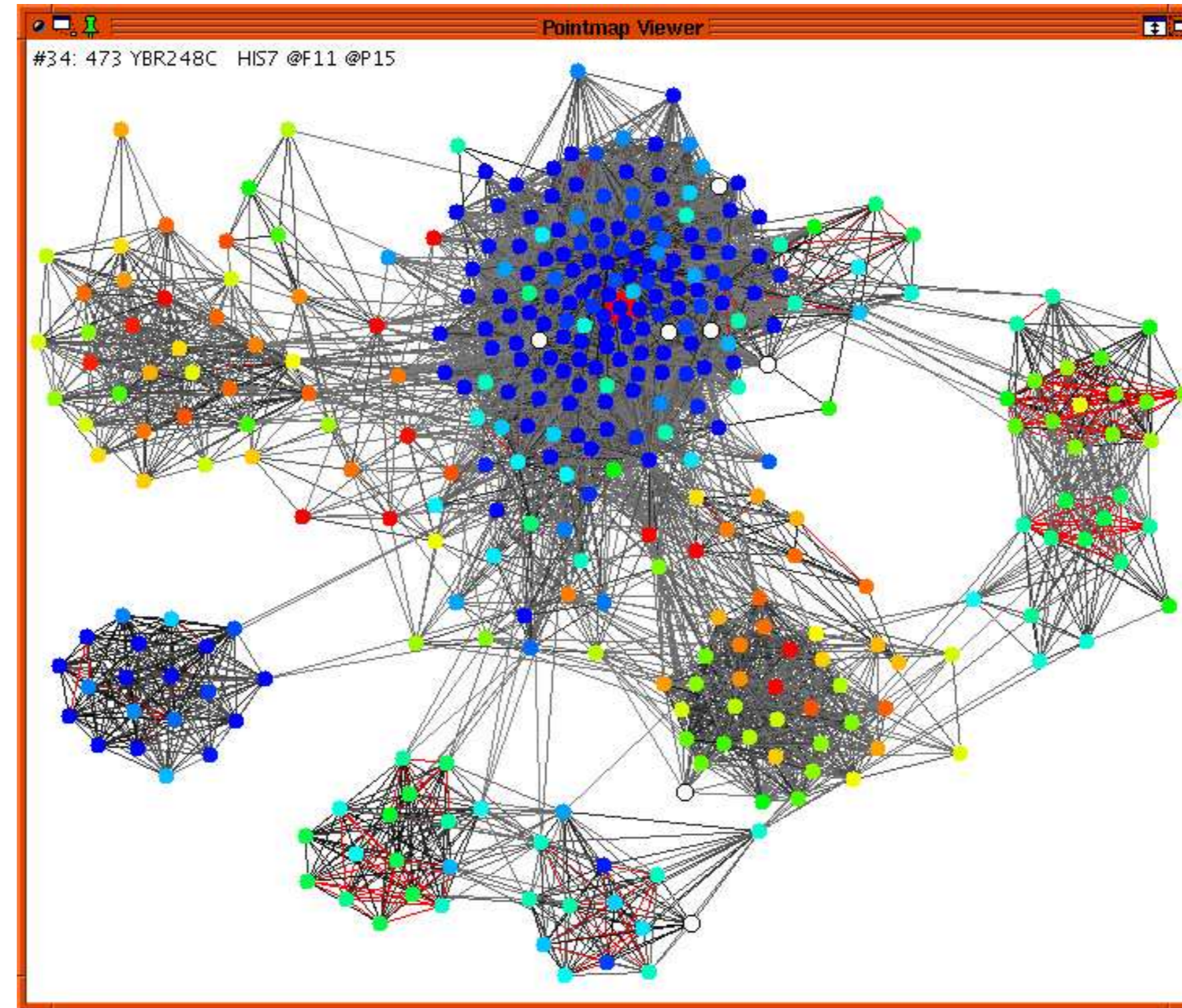
$$Q(\mathbf{x}, \mathbf{y}) = \sum_{i,j} w (|y_i - y_j| - f(|x_i - x_j|))^2$$

\mathbf{x}_i = Feature vector (high-dim.)
 \mathbf{y}_i = Representation vector (low-dim.)
 $f(d)$ = Distance kernel
 $w(d)$ = Weight kernel

The result is similar to a Self-Organized Map (SOM) without explicit prototypes.

JAMMING: Local sub-optimal minima occur frequently in low-dimensional problems.

- ▷ **DIMENSIONAL REDUCTION**: Optimization is first performed in higher-dimensional subspaces and successively reduced
- ▷ **MONTE CARLO SAMPLING**: The space of possible solutions is importance-sampled weighted by the target function (e.g. Langevin equations, molecular dynamics)



An alternative representation of the same subset of 352 genes using different distance kernels.

Network inference using Gaussian processes

Gaussian graphical models specify correlated expression between genes:

$$p(\mathbf{x}|M) = \sqrt{\frac{\det M}{(2\pi)^N}} \exp\left(-\frac{1}{2}\mathbf{x}^T M \mathbf{x}\right)$$

M_{ij} = sparse connectivity matrix of the network

Observed correlations can be direct (corresponding to a non-zero M_{ij}) or indirect (when mediated by a third object).

The challenge in network inference for Gaussian graphical models is to reduce the indirect (complete) to direct (partial) correlations.

Likelihood, given observations $\{x^{(i)}\}$:

$$p(D|M) = \sqrt{\frac{\det^K M}{(2\pi)^{NK}}} \exp\left(-\frac{K}{2} \text{Tr} DM\right)$$

with the observed correlation matrix:

$$D = D_{ij} = \frac{1}{K} \sum_{k=1}^K x_i^{(k)} x_j^{(k)}$$

Use Bayes' law to calculate the probability of a matrix M given the observations,

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}$$

this gives the probability distribution for matrices:

$$-\ln p(M|D) = \frac{K}{2} (-\ln \det M + \text{Tr} DM) + H(M)$$

$H(M)$ defines a prior on the matrices \Rightarrow sparsity, modularity.

Maximum-likelihood method:

$$-\frac{\partial}{\partial M_{ij}} \ln p(D|M) = \frac{K}{2} (-M_{ij}^{-1} + D_{ij}) = 0$$

Minimum is achieved at:

$$M_{ij} = D_{ij}^{-1} \quad \text{for } (i, j) \in E$$

$$M_{ij} = 0 \quad \text{for } (i, j) \notin E$$

if an edge set E is given.

Numerical method: gradient descent:

$$M_{ij}^{(n+1)} = M_{ij}^{(n)} - \eta \frac{K}{2} (-M_{ij}^{(n)-1} + D_{ij})$$

“Increase coupling if $D_{ij} > M_{ij}^{-1}$ ”.

\Rightarrow Maximizes likelihood under sparsity constraint

Use Jacobi iteration to calculate M^{-1} :

$$V^{(k+1)} = V^{(k)} + D^{-1} \underbrace{(I - MV^{(k)})}_{\text{Defect}}$$

$$\lim_{k \rightarrow \infty} V^{(k)} = M^{-1}$$

Small changes in $M \rightarrow$ few iterations to update V

Monte Carlo simulations:

Split model $M = (M', E)$

▷ E = edge set

▷ M' = restricted matrix such that $(i, j) \notin E \Rightarrow M'_{i,j} = 0$

Effective maximum-likelihood Hamiltonian:

$$H(E) = \min_{M'} H(M', E)$$

Perform Metropolis on E .

Monte Carlo strategies:

Greedy strategy:

▷ Choose as next edge the edge with highest defect

Metropolis/Simulated Annealing:

▷ Choose edges according to their defects with finite probabilities

▷ Accept edges according to the resulting $p(M)$

Hybrid Molecular Dynamics:

▷ Combines Metropolis + Molecular Dynamics method

▷ Basic step:

1. Propose (possibly discrete) Metropolis update
2. Evolve system continuously for some time
3. Accept according to final configuration

▷ Can combine discrete and continuous updates

▷ Increases acceptance rate

Network preferences:

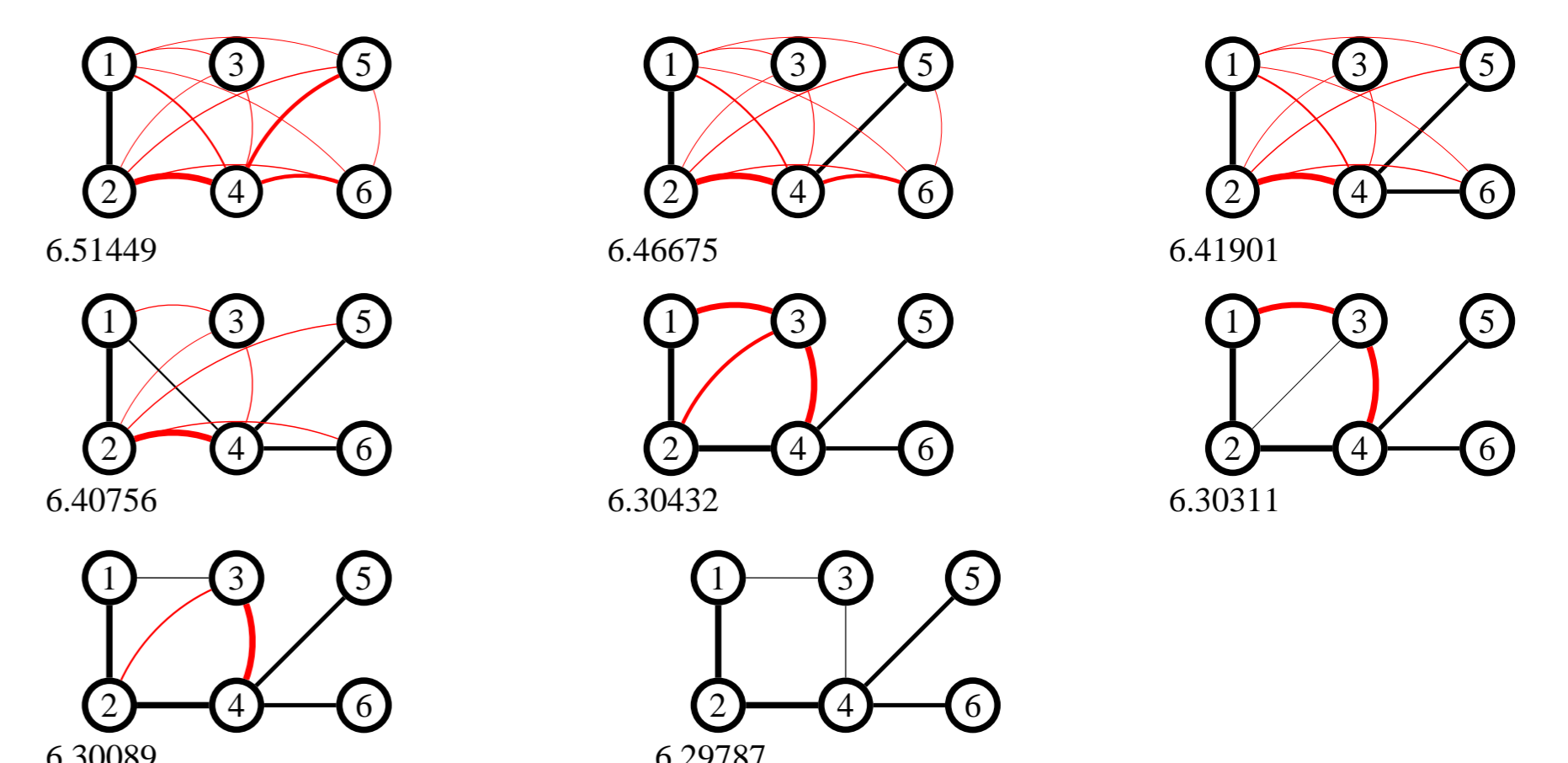
▷ Penalize high connectivity \rightarrow sparse networks

▷ Clustering: vs.

▷ Trees: vs.

▷ Hubs/Modules:

Few nodes have high connectivity
Nodes with high connectivity will preferentially receive more connections



Guided resolution of indirect correlations in an example network. The numbers give the log-probability of the network under the observed data.