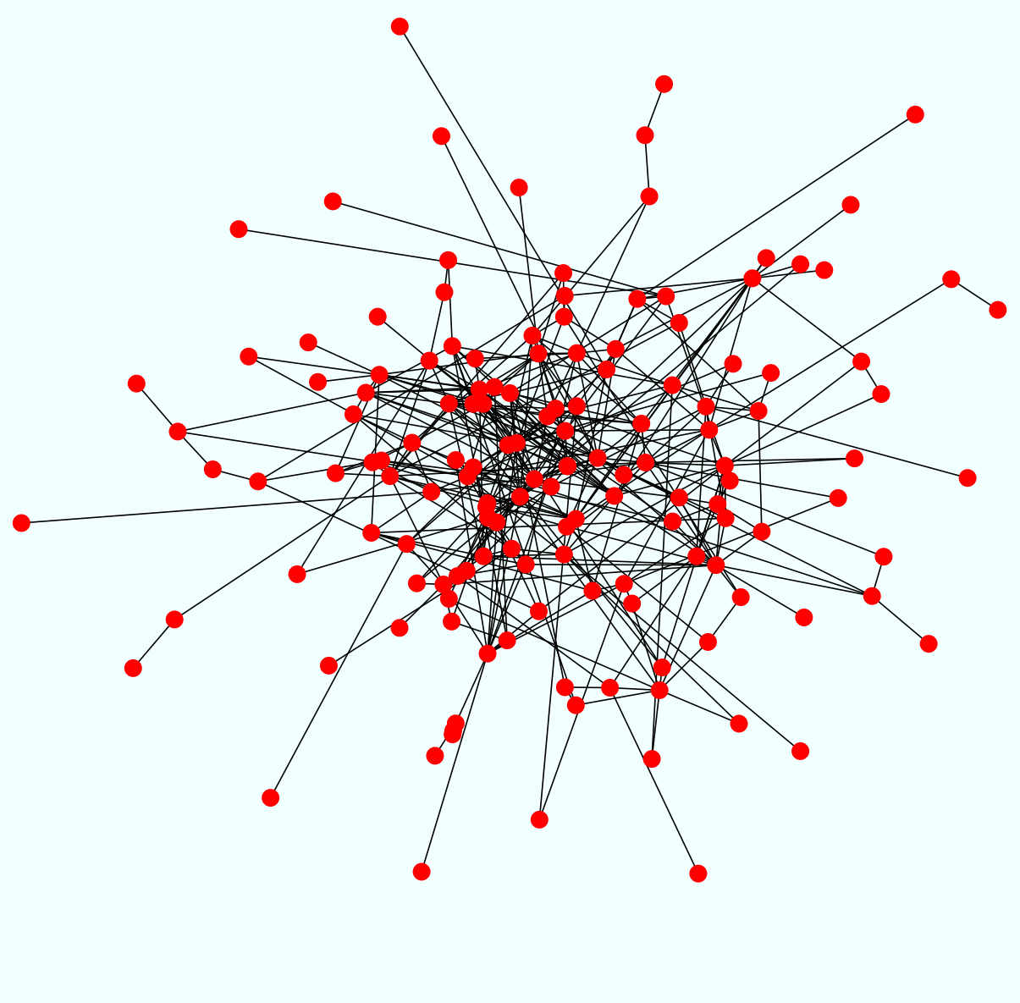
MA12: Information visualization

Jarkko Venna and Samuel Kaski

Adaptive Informatics Research Centre
Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK - Finland



Abstract — The goal of information visualization is to enlist the powerful processing capacity of human vision to help in finding patterns in the data. One of the basic tasks in exploratory visualization is to study the similarities or proximity relationships present in the data. For this the high dimensional data has to be converted to a two or three dimensional image. This can be achieved either by using specialized visualization techniques when the data is relatively low dimensional, or by using linear or nonlinear projection methods. It is not possible in general, however, to preserve all proximities present in the original data when reducing the dimensionality . Every projection method needs to make a compromise between trustworthiness and continuity. In a trustworthy projection the visualized similarities hold in the original data as well, whereas a continuous projection visualizes all proximities of the original data. We have developed methods for assessing the quality of different visualizations and a new nonlinear projection method that allows us to explicitely control the tradeoff done in the visualization process.

Background: Visualizing Similarities

The goal is to visualize the similarity relationships present in the data.

- Similarities are usually defined as distances; Short distance equals a high similarity. They can also be defined by the structure of the data as in graphs or other data with some form of network structure.
- Examples of methods: Scatterplots, dendrograms of hierarchical clustering, different non-linear projection methods and Self-Organizing maps.
- Dimensionality of the visualization is usually 2 or 3 and independent of the structure of the data.

Assessing the quality of visualizations: Is what we see really there?

The final arbiter on the quality of a visualization is the user who uses it to analyze a data set. Getting this kind of data on a quality of a visualization method is very hard to come by, however. That is why we have concentrated on studying what kinds of errors in the similarity structures are produced by the visualization process [1, 2].

- Two kinds of errors in the similarities can occur.

1. Objects might appear more similar in the visualization.(Trustworthiness)
2. Similar objects might appear dissimilar in the visualization. (Continuity)

Trustworthiness of a visualization

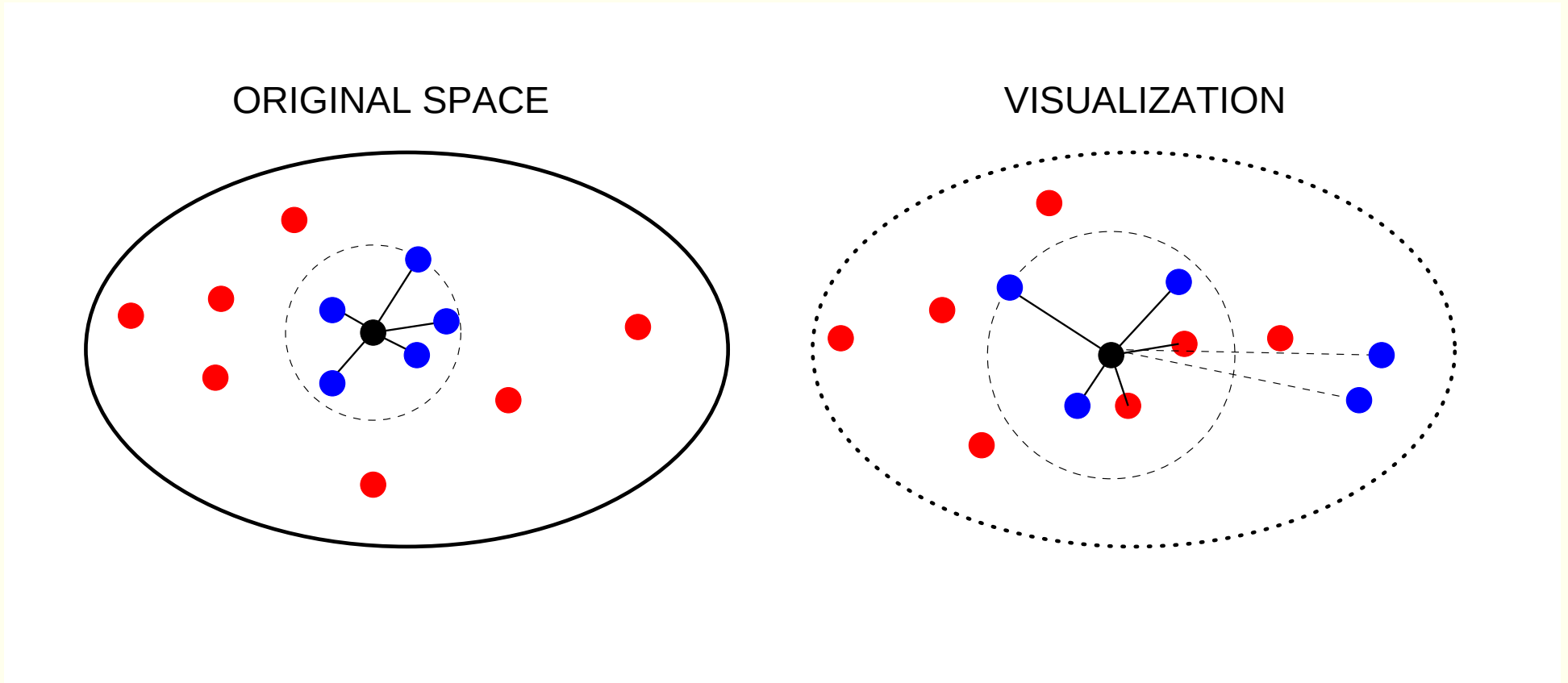
- We consider a projection onto a display *trustworthy* if the set of k closest neighbors of a point on the display are also close-by in the original space.

$$M_1(k) = 1 - A\sum_{i=1}^{N}\sum_{j\in U_k(i)}(r(i,j)-k)\ . \qquad (1)$$

Continuity of the neighborhoods

- We consider a projection onto a display to preserve the *continuity* if the set of k closest neighbors of a point in original space are also close-by on the display.

$$M_2(k) = 1 - A\sum_{i=1}^{N}\sum_{j\in V_k(i)}(\hat{r}(i,j)-k)\ . \qquad (2)$$



Comparing different projection methods

Data Sets

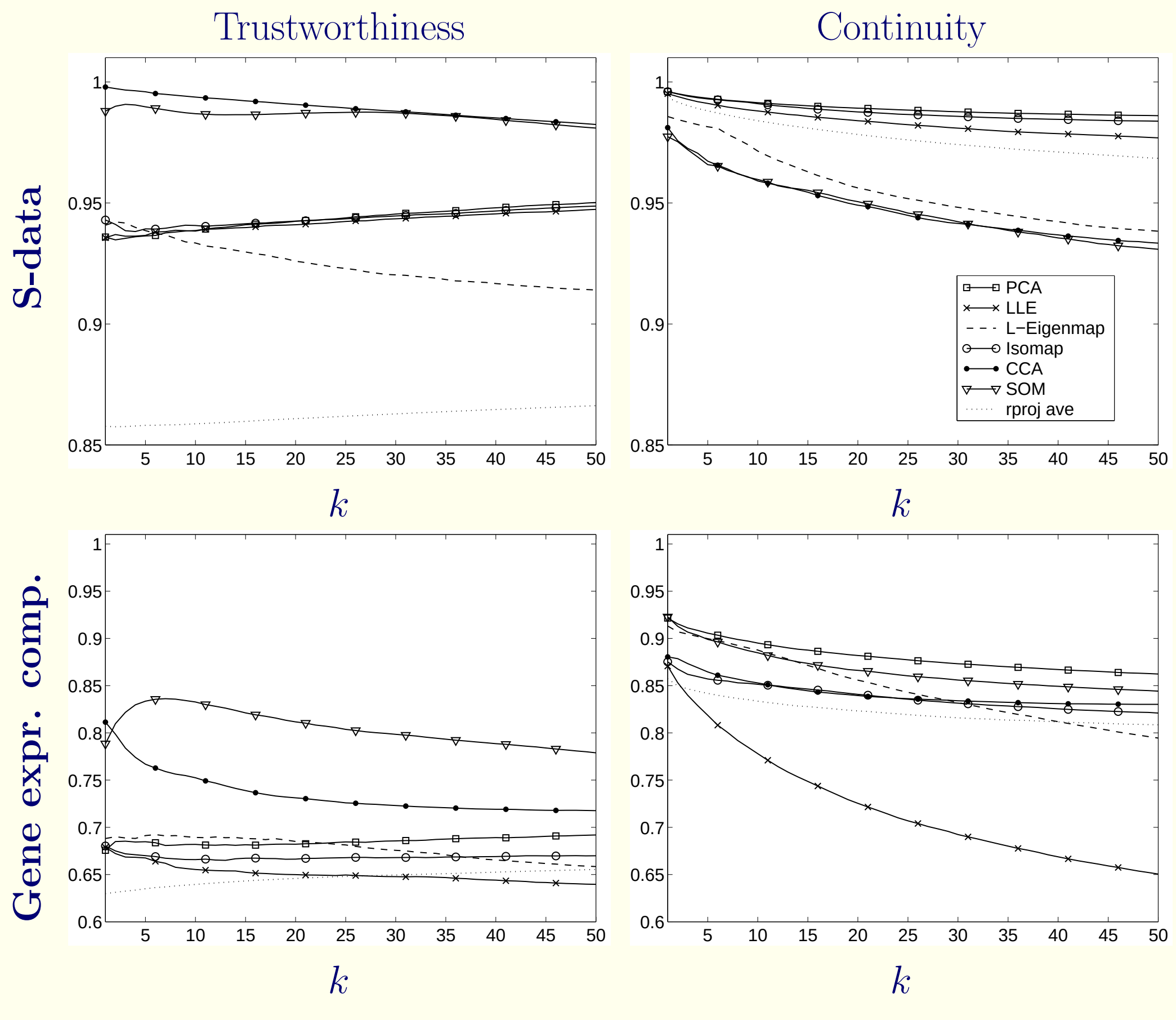
Thick S-curve. A simple data set lying on a thick S-shaped manifold in a three-dimensional space.

Gene expression compendium. A large collection of human gene expression arrays collected by Segal et al. (2004). This is a very hard data set to visualize.

Methods

- Principal component analysis (PCA)
- Locally linear embedding (LLE)
- Laplacian Eigenmap
- Isomap
- Curvilinear component analysis (CCA)
- Self-Organizing Map (SOM)

Results



Summary

Each method makes an intrinsic tradeoff between trustworthiness and continuity. Some methods like CCA and SOM produce results that have a high trustworthiness while other like PCA are usually good at producing displays that have a high continuity. Understanding the behavior of the methods is necessary for selecting the method that is the most suitable for the task at hand.

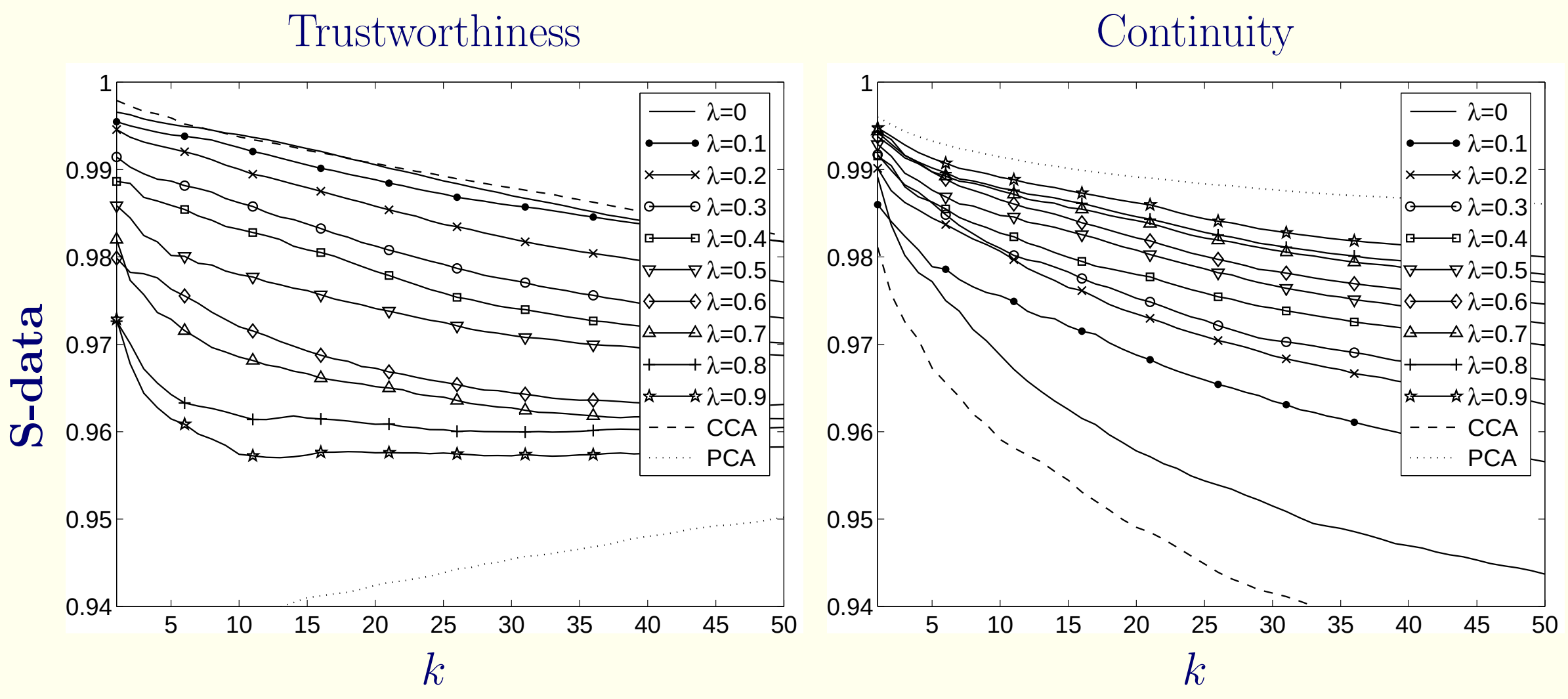
Local Multidimensional Scaling (MDS)

We have recently introduced a new visualization method for nonlinear projection of data sets [3]. It minimizes a cost function which is a tunable compromise between two types of errors: errors in preserving distances for data points that are neighbors on the *visualization*, and for points that are proximate in the *original space*. The cost function of local MDS is

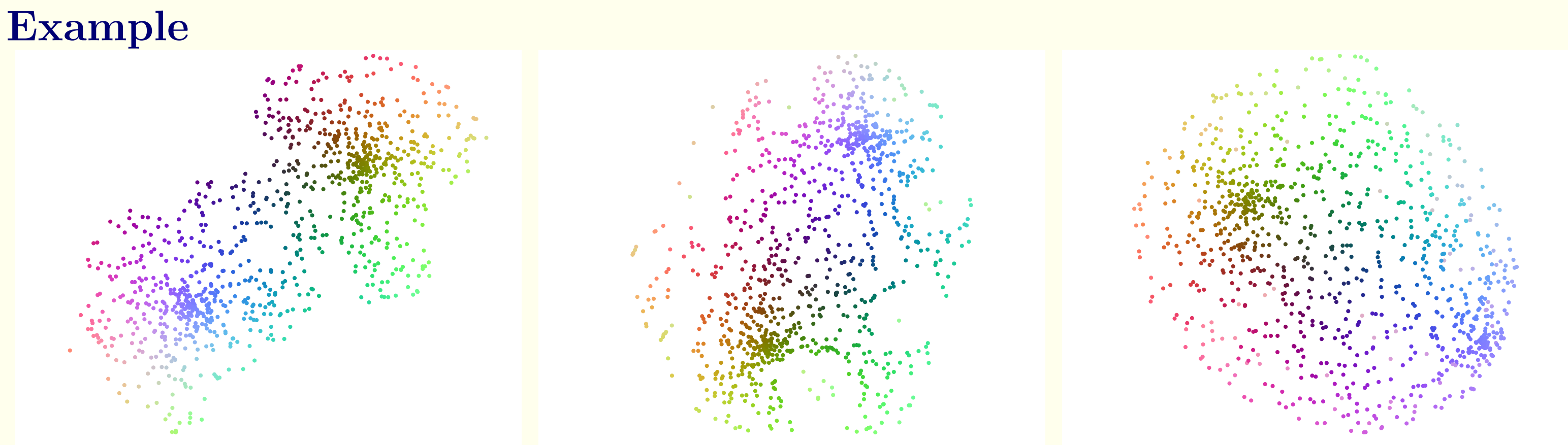
$$E = \frac{1}{2}\sum_{i}\sum_{j\neq i}(d(\mathbf{x}_i,\mathbf{x}_j)-d_{ij})^2[(1-\lambda)F(d(\mathbf{x}_i,\mathbf{x}_j),\sigma_i)+\lambda F(d_{ij},\sigma_i)]\ ,$$

$$F(d,\sigma)) = \begin{cases} 1 \text{ if } d\leq\sigma \\ 0 \text{ if } d>\sigma\ . \end{cases}$$

Here $d(\mathbf{x}_i,\mathbf{x}_j)$ is the distance between the points i and j in the visualization and d_{ij} is the distance between the points i and j in the original space.



Example



Three projections of a three-dimensional spherical cell with local MDS. On the left, trustworthiness of the projection is maximized by selecting $\lambda = 0$. In the middle and right, discontinuity of the projection is penalized as well, by setting $\lambda = 0.1$ and $\lambda = 0.9$, respectively.

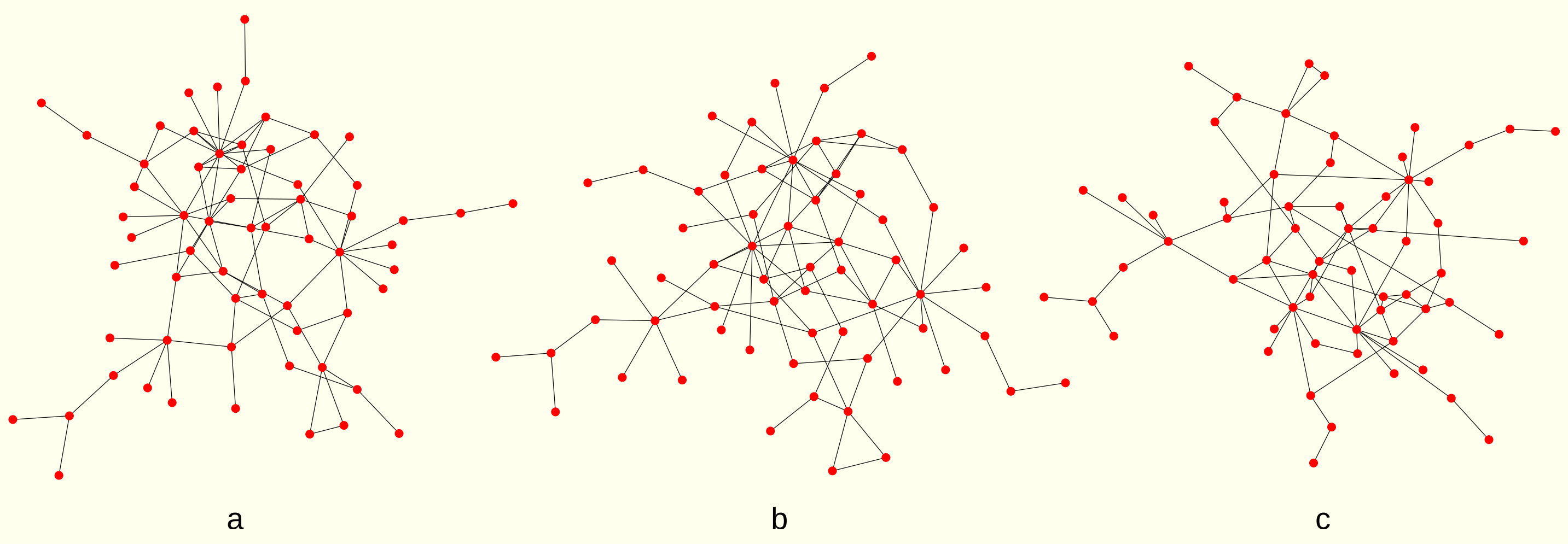
Application: Producing graph layouts for undirected graphs

Results [4]

Lee data, unweighted graph, 106 nodes 182 edges

Method	Trustworthiness (M_1^u)	Continuity (M_2^u)	Edge crossings
Graphviz	0.93	**0.96**	68
LGL	0.92	0.95	71
lMDS $\lambda = 0.2$	**0.99**	**0.96**	**33**

Trustworthiness (M_1), continuity of the mapping (M_2) and number of edge crossings produced by different methods.



Graph layouts for the Lee data: **a)** Graphviz **b)** LGL **c)** local MDS ($\lambda = 0.2$).

References

[1] Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Proceedings of ICANN 2001, International Conference on Artificial Neural Networks*, pages 485–491, Berlin, 2001. Springer.

[2] Samuel Kaski, Janne Nikkilä, Merja Oja, Jarkko Venna, Petri Törönen, and Eero Castrén. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics*, 4:48, 2003.

[3] J. Venna and S. Kaski. Local multidimensional scaling with controlled tradeoff between trustworthiness and continuity. In *Proceedings of 5th Workshop on Self-Organizing Maps*, pages 695–702, Paris, France, 2005.

[4] J. Venna and S. Kaski. Visualizing gene interaction graphs with local multidimensional scaling. In *Proceedings of 14th European Symposium of Artificial Neural Networks*, pages 557–562, Bruges, Belgium, 2006.