

Fast Euclidean Embedding of Ordinal Nearest Neighbor Graphs

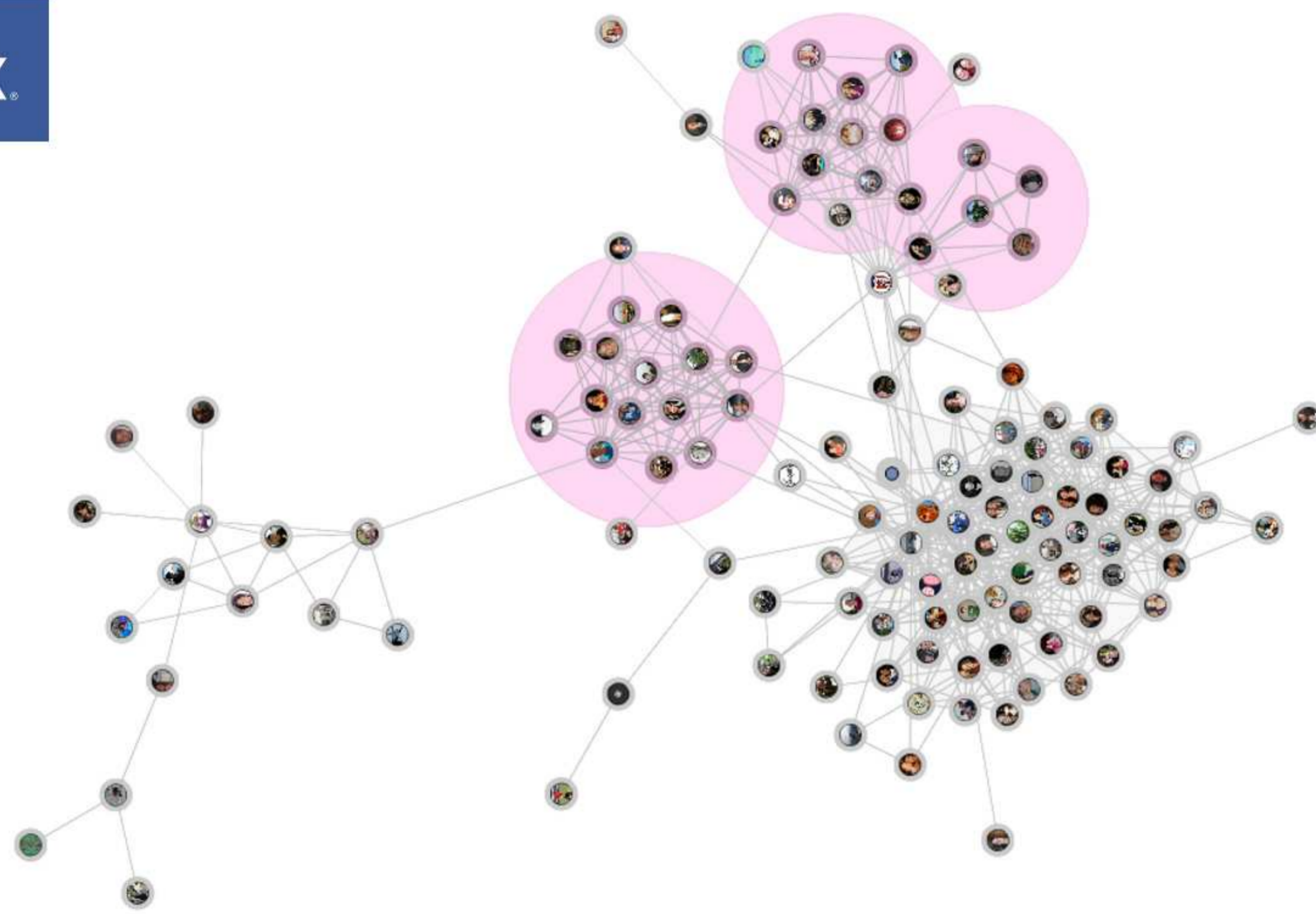
Brent Castle and Michael Trosset
Indiana University

Ordinal Nearest Neighbor Graphs

A number of datasets contain only ordinal nearest neighbor information about large proximity graphs.

- Social Graphs. E.g., Facebook.com prioritizes items in users' news feeds by "User Affinity" scores. These scores are pairwise and asymmetric.

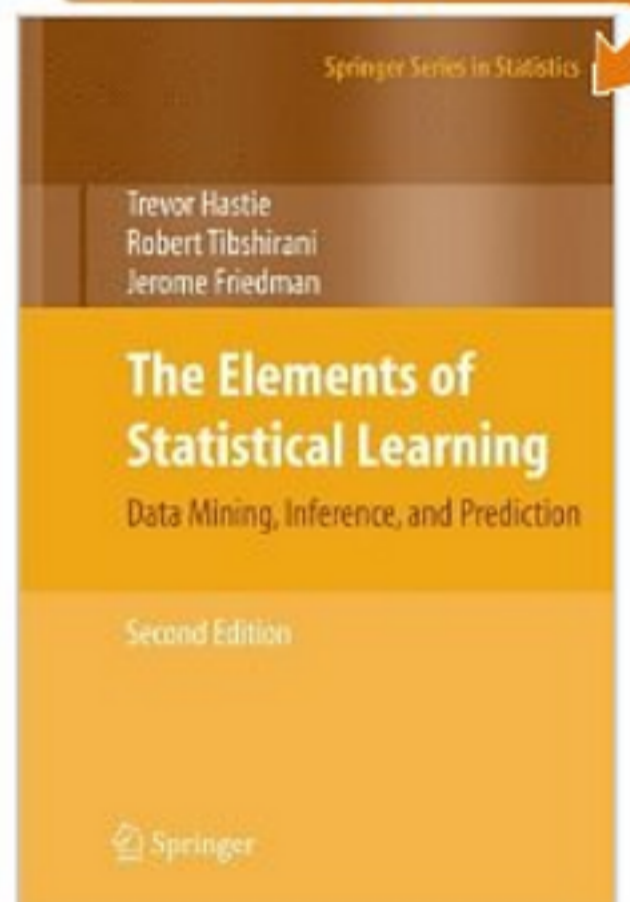
facebook



- "Co-occurrence" Graphs. E.g., Amazon.com lists the books most often purchased when a target book was purchased.

amazon.com

Click to LOOK INSIDE!



Customers Who Bought This Item Also Bought

1. Pattern Recognition and Machine Learning
2. Handbook of Statistical Analysis and Data Mining Applications
3. Probabilistic Graphical Models: Principles and Techniques
4. All of Statistics: A Concise Course in Statistical Inference
5. Machine Learning: An Algorithmic Perspective
6. Introduction to Machine Learning, 2nd Ed.
7. Data Mining: Practical Machine Learning Tools and Techniques, 2nd Ed.

We consider applications in which the information of interest is ordinal. Sometimes, only ranks of the nearest neighbors are provided without a corresponding quantitative measure of proximity. In other instances, quantitative measures are available, but are not immediately useful.

We seek methods for embedding nearest neighbor graphs into low-dimensional Euclidean spaces for subsequent analysis, e.g., visualization or classification.

Embedding Issues

Existence. There may be no configuration that satisfies the partial ordering of the distances. This requires us to minimize an error criterion.

Asymmetry. Nearest neighbor graphs are naturally asymmetric, i.e., it is possible that $r_{ij} \neq r_{ji}$. In fact, this asymmetry can be quite extreme, e.g., $r_{ij} = 1$ and $r_{ji} = \infty$. However, Euclidean distances are symmetric.

Sparsity. Pairwise information in a nearest neighbor graph is sparse, viz., we only have information about the pairwise relationship to a small number of neighbors of any target object.

Scale. Nearest neighbor graphs are especially useful representations when n is very large. Naive implementations of nonmetric multidimensional scaling with partial orderings are typically $O(n^4)$ [Burdakov, Sysoev, Grimvall and Hussian, 2006].

An Optimization Problem

We have a directed graph $G = (V, E)$ with $n = |V|$ vertices and ranks $r : E \mapsto \mathbb{N}$. The ranks are the rank order of neighbors from a source vertex, e.g., $r_{ij} = 2$ means that vertex j is the 2nd nearest neighbor of vertex i . For convenience we define $r_{ij} = \infty$ if $(i, j) \notin E$.

To illustrate, suppose that v_i 's nearest neighbor, second nearest neighbor, and third nearest neighbor are respectively v_j , v_k , and v_ℓ . Then $r_{ij} = 1$, $r_{ik} = 2$, and $r_{i\ell} = 3$. We are interested in configurations for which the interpoint distances have the same rank ordering, i.e.,

$$d(x_i, x_j) \leq d(x_i, x_k) \leq d(x_i, x_\ell).$$

Ideally, if $r_{ij} < r_{ik}$, then X should satisfy $d_{ij}(X) \leq d_{ik}(X)$. Often, the only configurations that satisfy this condition are degenerate (collapse to one point). We might search for an X that violates this condition as infrequently as possible, but the resulting optimization is intractable.

Instead, to obtain a tractable formulation we borrow from the nonmetric multidimensional scaling literature. We seek a configuration X and a set of disparities $\Delta = [\delta_{ij}]$ that minimize the *asymmetric stress*[†] criterion subject to ordinal constraints on the disparities:

$$\begin{aligned} &\text{minimize} && \sigma_a(\Delta, X) = \frac{1}{2} \sum_i \sum_j w_{ij} (\delta_{ij} - d_{ij}(X))^2 \\ &\text{subject to} && \delta_{ij} \leq \delta_{ik} \text{ if } r_{ij} < r_{ik} \text{ for } i \in 1, \dots, n \text{ and } (i, j), (i, k) \in E; \\ &&& \sum_i \sum_j \delta_{ij}^2 \geq c \end{aligned}$$

† - The asymmetric stress criterion allows w_{ij} not equal to w_{ji} and δ_{ij} to vary independently of δ_{ji} .

The final constraint discourages degenerate solutions.

Weights

We use 0-1 weights, i.e., $w_{ij} = 1$ or $w_{ij} = 0$. We consider three possibilities:

1. $w_{ij} = 1$ iff $(i, j) \in E$
2. $w_{ij} = 1$ for all i, j
3. $w_{ij} = 1$ for all $(i, j) \in E$ and some additional $(i, j) \notin E$

A natural choice is #1, but there is rarely enough pairwise information to obtain useful configurations. Choice #2 is the other extreme, in which a large amount of "missing" information is emphasized and computation is expensive. As a compromise, we add a set of $5q$ pairs $(i, j) \notin E$ (where q is the maximum out-degree in G) and denote the enlarged edge set by E^+ .

The KNNSCAL Algorithm

1. Set $q = \max(\{r_{ij}\}) + 1$.
2. Set $\Delta_{\text{sym}} = [\min(r_{ij}, r_{ji}, q)]$ and $\Delta = [\min(r_{ij}, q)]$.
3. Construct an initial X using Δ_{sym} .
4. Repeat until convergence:
 - a. Fix Δ and minimize σ_a wrt X (Asymmetric Stress Majorization).
 - b. Fix X and minimize σ_a wrt Δ (Isotonic Regression).
 - c. Rescale Δ to satisfy the nondegeneracy constraint.

Minimizing Stress

► Asymmetric Stress Majorization

Symmetric formulations often use the raw stress criterion [Kruskal, 1964]. Let $A_{ij} = [a_{ij}]$ with $a_{ii} = a_{ij} = 1$, $a_{ij} = a_{ji} = -1$, and 0 otherwise. Then,

$$\begin{aligned} \sigma_r(\Delta, X) &= \sum_{i < j} w_{ij} (\delta_{ij} - d_{ij}(X))^2 \\ &= \eta_\delta^2 + \text{tr}(X^T V X) - 2 \text{tr}(X^T B(X) X), \end{aligned}$$

where $V = \sum_{i < j} w_{ij} A_{ij}$ and $B(X) = \sum_{i < j} \frac{w_{ij} \delta_{ij}}{d_{ij}} A_{ij}$.

We seek to minimize the asymmetric stress criterion:

$$\begin{aligned} \sigma_a(\Delta, X) &= \frac{1}{2} \sum_i \sum_j w_{ij} (\delta_{ij} - d_{ij}(X))^2 \\ &= \frac{1}{2} \left[\eta_\delta^2 + \text{tr}(X^T V X) - 2 \text{tr}(X^T B(X) X) \right], \end{aligned}$$

where $V = \sum_{i < j} (w_{ij} + w_{ji}) A_{ij}$ and $B(X) = \sum_{i < j} \frac{w_{ij} \delta_{ij} + w_{ji} \delta_{ji}}{d_{ij}} A_{ij}$.

For fixed Δ , both criteria can be majorized by functions of the same form. Hence, given the appropriate V and $B(X)$, methods for majorizing raw stress can be used to majorize asymmetric stress.

To majorize stress we use the Diagonal Majorization Algorithm [Trosset and Groenen, 2005], which updates the configuration as follows

$$X \leftarrow X + \frac{1}{2} \text{diag}(V)^{-1} [B(X) - V] X$$

When V is sparse, we need only update as follows

$$X_i \leftarrow X_i + \frac{1}{2\gamma(i)} \left[\sum_{(i,j) \in E^+} \left(\left(\frac{\delta_{ij}}{d_{ij}} - 1 \right) (X_i - X_j) \right) + \sum_{(j,i) \in E^+} \left(\left(\frac{\delta_{ij}}{d_{ij}} - 1 \right) (X_j - X_i) \right) \right]$$

where $\gamma(i)$ is the in-degree plus the out-degree of vertex v_i .

► Isotonic Regression

We perform n completely ordered isotonic regressions ($O(k)$) [Grotzinger and Witzgall, 1984], one for each vertex. Each isotonic regression requires an $O(k \log k)$ sort to determine the ordinal constraints, resulting in an average case computation time of $O(nk \log k)$.

Numerical Experiment: Amazon Books

A graph containing $|V| = 1845$ was extracted from Amazon's "Customer's who bought this book ..." graph. We iterated step 4a 5x per iteration and step 4b 3x in total. The plot shows the $\log(\sigma_a)$ after each substep.

