

Energy/Stress Functions for Dimension Reduction and Graph Drawing: Power Laws and Their Clustering Properties*

Lisha Chen[†] and Andreas Buja

Yale University and University of Pennsylvania

This draft: August 5, 2009

Abstract

We introduce a parametrized family of energy/stress functions useful for proximity analysis, nonlinear dimension reduction, and graph drawing. The functions are inspired by the physics intuitions of attractive and repulsive forces common in graph drawing. Their minimization generates low-dimensional configurations (embeddings, graph drawings) whose interpoint distances match input distances as best as possible. The functions model attractive and repulsive forces with so-called Box-Cox transformations (essentially power transformations with the logarithmic transformation analytically embedded and negative powers made monotone increasing). Attractive and repulsive forces are balanced so as to attempt exact matching of input and output distances. To stabilize embeddings when only local distances are used, we impute infinite non-local distances that exert infinitesimal repulsive forces. The resulting energy/stress functions form a three-parameter family that contains many of the published energy functions in graph drawing and stress functions in multidimensional scaling. The problem of selecting an energy/stress function is therefore translated to a parameter selection problem which can be approached with a meta-criterion previously proposed by the authors (Chen and Buja 2009). Of particular interest is the tuning of a parameter associated with the notion of “clustering strength” proposed by Noack (2003). Such tuning greatly helps identifying “zero-dimensional manifolds”, i.e., clusters.

Key words and phrases: MDS, Force-Directed Layout, Cluster Analysis, Box-Cox Transformations, Unsupervised Learning

*Running title: Energy Functions for Dimension Reduction and Graph Drawing

[†]Corresponding author. Statistics Department, Yale University, 24 Hillhouse Ave, New Haven, CT 06511 (lisha.chen@yale.edu).

1 INTRODUCTION

In recent years an important line of work in machine learning has been non-linear dimension reduction and manifold learning. Among approaches used in this area many are based on inter-object distances and the faithful reproduction of such distances by so-called “embeddings,” that is, mappings of the objects of interest (images, signals, documents, genes, ...) to low dimensional spaces whereby the low-dimensional distances mimic the “true” inter-object distances as best as possible. Examples of distance-based methods include kernel PCA (KPCA; Schölkopf, Smola, and Müller 1998), “Isomap” (Tenenbaum, Silva, and Langford 2000), kernel-based semidefinite programming (SDP; Lu, Keles, Wright, and Wahba 2005; Weinberger, Sha, Zhu, and Saul 2006), and local multidimensional scaling (Chen and Buja 2009). These are all outgrowths of one of the two forms of multidimensional scaling, namely, Torgerson-Gower classical scaling which approximates inner products, and Kruskal-Shepard distance scaling which approximates distances directly.

The present article proposes a shift away from the established paradigms of non-linear dimension reduction and manifold learning. This shift is not meant to replace but to supplement the established paradigms by suggesting that in many practical situations the natural “manifold” is not an ideal smooth object but a collection of “zero-dimensional manifolds,” that is, clusters. In such situations it would be misleading to probe for smooth underlying structure; instead, a meaningful approach is to look for embeddings that reflect the underlying clustering. To this end, we introduce a parametrized family of surrogate loss functions (“energy functions”, “stress functions”) whose parameters can be used to tune the embeddings to reveal clusters at any scale as best as possible. The benefit of this approach is that it produces embeddings and visual clusterings in one process, whereas the usual approaches are ad hoc combinations of cluster analyses visualized in unrelated embeddings such as PCA.

Handing data analysts a dial to turn up and down the degree of clustering in a visualization creates a quandary: How real are the shown clusters? While this question may be difficult to answer, we provide nevertheless an independent tool for judging the faithfulness of embeddings in terms of the degree to which the underlying metric structure is reproduced, if not in magnitude, then in rank. That is, we use a family of criteria (Chen and Buja 2009; Chen 2006; Akkucuk and Carroll 2006) which allows us to diagnose the degree to which K -nearest neighborhoods are preserved in the mapping of objects to their images in an embedding. K -NN structure depends only on the ranks of sorted distances both in object space and in the embedding, and as such it is insensitive to non-linear but monotone transformations of the distances in both domains. The implication or, rather, benefit of this insensitivity is that it allows seemingly biased embeddings to be recognized as “correct” in a minimalist sense. For example, highly clustered embeddings may seem to exaggerate the clustering aspect but they may nevertheless

preserve essential topology by placing the clusters correctly and by preserving cluster-internal structure on a miniaturized scale internal to the clusters.

To solve the problem of tunable cluster-sensitivity in distance-based dimension reduction, we can look for inspiration in the graph-drawing literature, in particular in work by A. Noack (2003). This author introduced a “clustering postulate” which requires “energy functions” for graph drawing to satisfy a certain requirement of node placement when applied to a stylized two-clique situation. For two infinitely tight cliques, energy minimization can be performed analytically, and this exercise can be used to characterize the clustering properties of — and to impose desirable clustering properties on — energy functions. We make use of this possibility and extend Noack’s clustering postulate to a one-parameter family of postulates that characterize differing degrees of clustering strength. These postulates amount to very simple ordinary differential equations that put attractive and repulsive energies in a fixed relation to each other. The differential equations have some intuitive solutions in the form of power laws. By using the Box-Cox modification of power transformations (well-known in statistics), we extend the family of powers to include logarithmic and negative power laws as well. The resulting parametrized family of energy functions encompasses many of the better known energy functions from the graph drawing literature and characterizes them in terms of differing power laws and clustering strengths. In a next step we carry this approach over from graph data to distance data, thereby enriching distance-based multidimensional scaling with a large class of stress functions. Finally, we extend stress functions for complete distance data to the incomplete case, that is, to distance-weighted graphs, by extending a “trick” we previously used (Chen and Buja 2009) whereby missing non-local distances are imputed by “infinite” distances but with “infinitesimal” repulsive force.

While some readers may deplore the proliferation of energy/stress functions, it is a fact that many already exist without a unifying framework. By embedding a critical mass of previously proposed energy/stress functions in a single parametric family, we facilitate the understanding of their differences in terms of power laws and clustering strength, and we make them accessible in a single methodology that allows us to select good parameter combinations in terms of an independent meta-criterion (Section 4, based on Chen and Buja 2009).

We will illustrate the effects of tuning clustering strength on one artificial data example and two well-known image datasets, the Olivetti face data and the Frey face data. The artificial example shows how clustering can be revealed in multi-scale structure of embeddings. The Olivetti face data is naturally clustered into forty individuals with ten images each, and this is revealed by properly tuned clustering strengths. The Frey face data have mostly a cobweb-like structure whose cluster aspects get sharpened to various degrees by tuning the clustering strength.

This article proceeds as follows: Section 2 gives background on energy functions used

in graph drawing and describes Noack’s (2003) clustering postulate. Section 3 extends the clustering postulate to a power family of clustering postulates, and derives the Box-Cox or “B-C” family of energy/stress functions with tunable clustering strength. Section 4 introduces the meta-criterion for choosing tuning parameters. Section 5 shows the artificial and the two image data examples. Section 6 concludes with a summary and discussion. The Appendix ties up some loose ends.

2 Background: Graph Layout, Energy Functions and the Clustering Postulate

2.1 Graph Layout with Energy Functions

Let $G = (V, E)$ be a graph consisting of vertices V and edges E . The goal of graph layout is to draw a low-dimensional picture of points (vertices) and lines (edges) to best visualize the relational information in a graph. For now we consider unweighted graphs, but the application to nonlinear dimension reduction is immediate once the energy function framework is extended to distance-weighted graphs. In dimension reduction the graph is typically a symmetrized K -nearest neighbor graph, and finding low-dimensional embeddings is then a special case of the general distance-weighted graph layout problem.

A widely used approach to graph layout is force-directed placement. The idea can be best explained with Eades’ metaphor (Eades 1984):

To embed [lay out] a graph we replace the vertices by steel rings and replace each edge with a spring to form a mechanical system... The vertices are placed in some initial layout and let go so that the spring forces on the rings move the system to a minimal energy state.

Based on the heuristic of such physical systems, investigators in graph layout have proposed a number of energy-minimizing algorithms. In Table 1 we list a few well-known examples of energy functions. Perusing this table, it is not a priori clear what the advantage would be of one proposal over the others. The last example was most recently proposed by Noack based on his “clustering postulate,” the starting point for our own proposals.

We note that each of the energy examples is a function $U = U(G, X)$ of a graph $G = (V, E)$ and a configuration $X = (\mathbf{x}_i)_{i=1\dots N}$ through distances $d_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\|$. Minimization of an energy function $U(G, X)$ over all configurations X produces the best layout as measured by the particular energy function. A commonality of the examples listed in Table 1 is that attractive forces exist among the vertices connected by edges $(i, j) \in E$, while repulsive forces exist between all vertices $(i, j) \in V^2$ (= full graph). Hence such energy functions can be

Eades (1984)	$U = \sum_E d_{i,j} \cdot (\log(d_{i,j}) - 1) + \sum_{V^2} d_{i,j}^{-1}$
Fruchterman and Reingold (1991)	$U = \sum_E d_{i,j}^3/3 - \sum_{V^2} \log(d_{i,j})$
Davidson and Harel (1996)	$U = \sum_E d_{i,j}^2 + \sum_{V^2} d_{i,j}^{-2}$
Noack (2003)	$U = \sum_E d_{i,j} - \sum_{V^2} \log(d_{i,j})$

Table 1: *Some well-known energy functions U for graph layout. In each case the first term sums over edges $(i, j) \in E$ in the graph, and the second term sums over all possible edges $(i, j) \in V^2$ (= complete graph); the terms $d_{i,j}$ are the interpoint distances in the embedding.*

written in the form

$$U = \sum_{(i,j) \in E} f(d_{i,j}) - \sum_{(i,j) \in V^2} g(d_{i,j}). \quad (1)$$

Both $f(d)$ and $g(d)$ are monotone increasing to do justice to the terms “attraction” and “repulsion”. The idea is to balance graph structure according to the attractive force with a tendency to spread out according to the repulsive force.

2.2 Noack’s Clustering Postulate

Noack’s (2003) clustering postulate derives from the consideration of an idealized two-cluster situation that is so ingeniously simple that it lends itself to treatment with elementary calculus, and yet it gives insight into the response of energy functions to clusters in graphs. Noack laid down his postulate as a precise requirement on how strongly energy functions should attempt to separate clusters. We find his requirement too rigid for present purposes, which is why we will soften it up to describe an infinite range of responses to clustering.

Following Noack (2003), we consider an idealized graph consisting of two subgraphs, to be called “clusters.” To justify this term, we assume the vertices highly connected within each cluster and loosely connected between them. As depicted in Figure 1, we assume n_1 and n_2 vertices in the two clusters, respectively, and n_e edges between them. We assume that all the vertices of one cluster are drawn at the same position. As we are only interested in the relative placement of the clusters and not their internal structure, we ignore the energy within clusters and focus on the energy between them. The relevant energy, which we write as $U_{2cl}(d)$, has then the following form, where d is the distance at which the two clusters are placed in the embedding:

$$U_{2cl}(d) = n_e f(d) - n_1 n_2 g(d)$$

Noack defines the “coupling strength” C as a measure of how strongly the two clusters are

connected or “coupled”:

$$C = \frac{n_e}{n_1 n_2}.$$

This quantity is always between 0 and 1, with larger C implying stronger coupling strength. The energy function is proportional to

$$U_{2cl}(d) \sim f(d) - \frac{1}{C} g(d). \quad (2)$$

Noack postulates that energy functions should attempt to place strongly coupled clusters closer to each other than loosely coupled clusters. He goes even further by requiring that the energy-minimizing distance between the clusters in the embedding should be proportional to $1/C$:

Clustering Postulate: $d_{\min} = 1/C$

That is, maximally connected subgraphs ($C = 1$ or $n_e = n_1 n_2$) should be placed at a distance $d = 1$, whereas lesser connected subgraphs ($C < 1$ or $n_e < n_1 n_2$) should be placed at ever greater distance from each other, proportional to $1/C = n_1 n_2 / n_e > 1$. This postulate represents a fruitful idea, even though it is too rigid for our purposes. Its power is to take some of the arbitrariness out of energy functions by constraining the attractive and repulsive forces to each other, as we will show next: Assuming smoothness of $f(d)$ and $g(d)$, a necessary condition for achieving the minimal energy at a distance $d = d_{\min}$ is $U_{2cl}'(d_{\min}) \sim f'(d_{\min}) - (1/C) g'(d_{\min}) = 0$, that is,

$$f'(d_{\min}) = \frac{1}{C} g'(d_{\min}). \quad (3)$$

According to the Clustering Postulate above, $d_{\min} = 1/C$, for all possible values of the coupling strength $C \in (0, 1]$. Hence the minimum energy condition (3) combined with the Clustering Postulate results in the following constraint between attractive and repulsive forces:

Clustering Constraint: $f'(d) = d \cdot g'(d) \quad \forall d \geq 1$

This is a powerful condition that reduces the arbitrariness in the choice of energy functions: if either the attractive or the repulsive force is given, the other force is determined (up to an irrelevant additive constant). Noack introduces two new energy functions by applying the clustering constraint to two special cases: $g(d) = \log(d)$ and $g(d) = d$. The first results in what he calls the “LinLog” energy:

$$U_{LinLog} = \sum_{(i,j) \in E} d_{i,j} - \sum_{(i,j) \in V^2} \log d_{i,j},$$

the second in what he calls the “QuadLin” energy:

$$U_{QuadLin} = \sum_{(i,j) \in E} \frac{1}{2} d_{i,j}^2 - \sum_{(i,j) \in V^2} d_{i,j}.$$

Motivated by the LinLog energy, Noack goes one step further and considers a one-parameter family of “PolyLog” energy functions:

$$U_{PolyLog} = \sum_{(i,j) \in E} d_{i,j}^r - \sum_{(i,j) \in V^2} \log d_{i,j},$$

where $r \geq 1$. He notes that in the idealized two-cluster situation they attain their minimum energy distance at

$$d_{\min} = (1/C)^{1/r}.$$

We take the intuitions generated by this example as the starting point for systematically devising energy functions with variable clustering strengths.

3 Box-Cox Energy Functions

3.1 B-C Energies for Unweighted Graphs

We make Noack’s clustering postulate more flexible by allowing the minimum distance in his two-cluster paradigm to be any power function of the inverse coupling strength:

Clustering Power Postulate: $d_{\min} = (1/C)^\lambda \quad (\lambda > 0)$

The parameter λ is a measure of the strength of the clustering effect: for larger λ , the postulate requires that loosely coupled clusters get separated more from each other. As λ approaches zero, the clustering effect vanishes. We therefore call λ the “**clustering power.**” Just as Noack’s original postulate does, the power postulate imposes a constraint between attractive and repulsive forces. Substituting $1/C = d_{\min}^{1/\lambda}$ in the minimum energy condition (3) we obtain the

Clustering Power Constraint: $f'(d) = d^{1/\lambda} \cdot g'(d) \quad \forall d \geq 1$

The energy functions by Fruchterman and Reingold (1991) and Davidson and Harel (1996) both satisfy the clustering power constraint with the clustering powers $\lambda = 1/3$ and $1/4$, respectively (see Table 1). This fact explains why these energy functions have a weaker clustering effect than the LinLog and QuadLin energy functions, which both have the clustering power $\lambda = 1$ (compare Noack 2003, Table 3)

In what follows we introduce a family of energy functions whose attractive and repulsive forces follow power laws. However, we would like this family to also include logarithmic laws such as LinLog. To accommodate such preferences, statisticians have long used the so-called Box-Cox family of transformations, defined for $d > 0$ by

$$BC_p(d) = \begin{cases} \frac{d^p - 1}{p} & (p \neq 0) \\ \log(d) & (p = 0) \end{cases}$$

This modification of the raw power transformations d^p not only affords analytical fill-in with the natural logarithm for $p = 0$, it also extends the family to $p < 0$ while preserving increasing monotonicity of the transformations (for $p < 0$ raw powers d^p are monotone decreasing while $BC_p(d)$ is not). The derivative is

$$BC_p'(d) = d^{p-1} > 0 \quad \forall d > 0, \quad \forall p \in \mathbb{R}.$$

See Figure 2 for an illustration of Box-Cox transformations.

In addition to imposing a Box-Cox family of power laws, we assume a power clustering constraint with a given clustering power λ . As a result, if either $f(d)$ or $g(d)$ is of the Box-Cox variety, the other will be, too. Thus, denoting the Box-Cox power of the repulsive energy $g(d)$ by μ , we obtain the following matched pairs of repulsive and attractive energies:

$$f(d) = BC_{\mu+\frac{1}{\lambda}}(d) = \begin{cases} \frac{d^{\mu+\frac{1}{\lambda}}-1}{\frac{1}{\lambda}+\mu} & (\mu + \frac{1}{\lambda} \neq 0) \\ \log(d) & (\mu + \frac{1}{\lambda} = 0) \end{cases} \quad (4)$$

$$g(d) = BC_{\mu}(d) = \begin{cases} \frac{d^{\mu}-1}{\mu} & (\mu \neq 0) \\ \log(d) & (\mu = 0) \end{cases} \quad (5)$$

Thus μ is the power law of repulsion, $\mu + \frac{1}{\lambda}$ the power law of attraction, and λ is the clustering power: $f'(d) = d^{\mu+\frac{1}{\lambda}} = d^{\frac{1}{\lambda}}g'(d)$. The full energy function is:

$$U = \sum_{(i,j) \in E} BC_{\mu+\frac{1}{\lambda}}(d) - \sum_{(i,j) \in V^2} BC_{\mu}(d). \quad (6)$$

This is what we call the **B-C family of energy functions** for unweighted graphs. It includes as special cases Noack's LinLog ($\mu = 0$, $\lambda = 1$), QuadLin ($\mu = 1$, $\lambda = 1$) and PolyLog ($\mu = 0$, $\lambda = 1/r$) energies, as well as the Fruchterman and Reingold ($\mu = 0$, $\lambda = 1/3$) and Davidson and Harel ($\mu = -2$, $\lambda = 1/4$) energies. These identifications are modulo irrelevant additive constants and (more surprisingly) modulo irrelevant relative reweighting of attractive and repulsive energies. Reweighting by constant weights only changes the scale, not the shape, of optimal embeddings, as will be seen in Section 3.4.

[One might worry that the clustering power constraint is only necessary but not sufficient for a local minimum of the 2-cluster energy function (2) at $d_{\min} = 1/C^{\lambda}$. However, for the B-C family of energy functions the second derivative is always positive at d_{\min} : $U_{2cl}''(d_{\min}) \sim d_{\min}^{\mu+1/\lambda-2}/\lambda > 0$ for $\lambda > 0$.]

3.2 B-C Energies for Complete Distance Data

We next address the situation in which complete distance data $D = (D_{i,j})_{i,j=1,\dots,n}$ is given instead of a graph $G = (V, E)$. The problem of embedding distance data in Euclidean spaces

is that of finding configurations $X = (\mathbf{x}_i)_{i=1\dots N}$ whose distances $d_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\|$ approximate the distance data $D_{i,j}$ as best as possible. The paradigmatic technique for solving this problem is multidimensional scaling (MDS). Graph layout can be reduced to the embedding problem for distance data by extracting a distance matrix from the graph, for example, by computing $D_{i,j}$ as the shortest-path length for all pairs of vertices (i, j) . This approach underlies Isomap (Tenenbaum et al. 2000) in the context of non-linear dimension reduction. The earliest reference that uses the distance approach for graph drawing is most likely Kruskal and Seery (1980), followed by Kamada and Kawai (1989) and Gansner, Koren and North (2004). In this subsection we carry the clustering power postulate and the notion of B-C energy functions over to the problem of embedding complete distance data, thereby enriching MDS with a large class of stress functions that contain Kruskal's (1964) original stress as a special case. In the next subsection we will arrive at the final form of energy/stress functions for distance-weighted graphs, that is, for incomplete distance data.

A natural way to extend the clustering postulate to distance data is to require that in Noack's two-cluster paradigm the minimum energy distance between clusters be achieved at a target distance D expanded by the clustering factor $(1/C)^\lambda$:

Clustering Power Postulate, Distance Weighted: $d_{\min} = D \cdot (1/C)^\lambda$

In the special case of two connected points (each cluster containing just one point), we have $n_e = n_1 = n_2 = 1$ and hence $C = 1$, hence the energy-minimizing distance is $d_{\min} = D$. In other words, the energy balance between connected vertices is such that two connected points i and j with target distance $D_{i,j}$ attempt to be placed at a distance $d_{i,j} = D_{i,j}$ from each other in the embedding.

Substituting $1/C = (d_{\min}/D)^{1/\lambda}$ from the new postulate in the minimum energy condition (3) results in the following constraint between attraction and repulsion:

Power Clustering Constraint, Distance Weighted: $f'(d) = (d/D)^{\frac{1}{\lambda}} g'(d)$

We can again prescribe $g(d)$ and obtain $f(d)$ from this constraint. The proposal for $g(d)$ is essentially the same as in Section 3.1, but multiplied with a power of the target D :

$$f(d) = BC_{\mu+\frac{1}{\lambda}}(d), \quad g(d) = D^{\frac{1}{\lambda}} BC_{\mu}(d). \quad (7)$$

There is, however, freedom to weight $f(d)$ and $g(d)$ by a common factor that can be an arbitrary function of the target distance D . Remaining true to the power theme, we let this common factor be an arbitrary power D^ν :

$$f(d) = D^\nu BC_{\mu+\frac{1}{\lambda}}(d), \quad g(d) = D^{\nu+\frac{1}{\lambda}} BC_{\mu}(d).$$

We arrive at the B-C family of stress functions for distance data:

$$U = \sum_{(i,j) \in V^2} D_{i,j}^\nu \left(BC_{\mu+\frac{1}{\lambda}}(d_{i,j}) - D_{i,j}^{\frac{1}{\lambda}} BC_\mu(d_{i,j}) \right). \quad (8)$$

The stress used in Kruskal-Shepard metric MDS is a special case obtained for $\mu = 1$, $\lambda = 1$ and $\nu = 0$. ‘‘SStress’’ used in the ALSCAL procedure of Takane, Young and de Leeuw (1977) is obtained for $\mu = 2$, $\lambda = 1/2$ and $\nu = 0$. Sammon’s mapping (1969) is covered by $\mu = 1$, $\lambda = 1$ and $\nu = -1$, whereas Kamada and Kawai (1989) use the same μ and λ but $\nu = -2$ for graph drawing (see Table 2).

Because down-weighting large distances destabilizes embeddings by increasing the number of local minima in the stress function, we dispense with the weight $D_{i,j}^\nu$ and let $\nu = 0$. If the goal is to exploit local structure, a better solution than down-weighting large distances is shown in Subsection 3.3. Our final form of stress function for complete distance data is therefore:

$$U = \sum_{(i,j) \in V^2} \left(BC_{\mu+\frac{1}{\lambda}}(d_{i,j}) - D_{i,j}^{\frac{1}{\lambda}} BC_\mu(d_{i,j}) \right). \quad (9)$$

3.3 B-C Energies for Distance-Weighted Graphs

In order to arrive at stress/energy functions for non-full graphs, we extend a device we used previously to transform Kruskal-Shepard MDS into a localized version called ‘‘local MDS’’ or ‘‘LMDS’’ (Chen and Buja 2009). We now assume target distances $D_{i,j}$ are given only for $(i, j) \in E$ in a graph. Starting with stress functions (9) for full graphs, we replace the dissimilarities $D_{i,j}$ for $(i, j) \notin E$ with a single large dissimilarity D_∞ which we let go to infinity. We down-weight these terms with a weight w in such a way that $wD_\infty^{\frac{1}{\lambda}} = t^{\frac{1}{\lambda}}$ is constant:

$$\begin{aligned} U &= \sum_{(i,j) \in E} \left(BC_{\mu+\frac{1}{\lambda}}(d_{i,j}) - D_{i,j}^{\frac{1}{\lambda}} BC_\mu(d_{i,j}) \right) \\ &\quad + w \sum_{(i,j) \notin E} \left(BC_{\mu+\frac{1}{\lambda}}(d_{i,j}) - D_\infty^{\frac{1}{\lambda}} BC_\mu(d_{i,j}) \right) \end{aligned}$$

As $D_\infty \rightarrow \infty$, we have $w = (t/D_\infty)^{1/\lambda} \rightarrow 0$, hence in the limit we obtain:

$$U = \sum_{(i,j) \in E} \left(BC_{\mu+\frac{1}{\lambda}}(d_{i,j}) - D_{i,j}^{\frac{1}{\lambda}} BC_\mu(d_{i,j}) \right) - t^{\frac{1}{\lambda}} \sum_{(i,j) \notin E} BC_\mu(d_{i,j}). \quad (10)$$

This we now call the *B-C family of stress or energy functions* for distance weighted graphs. The parameter t balances the relative strength of the combined attraction and repulsion inside the graph with the repulsion outside the graph. For completeness, we list the assumed ranges of the parameters:

$$t \geq 0, \quad \lambda > 0, \quad -\infty < \mu < \infty.$$

Choice of Parameters	Special Cases
$E = V^2, \lambda = \mu = 1, \nu = 0$	MDS (Kruskal 1964; Kruskal & Seery 1980)
$E = V^2, \mu = 2, \lambda = 1/2, \nu = 0$	ALSCAL (Takane, Young and de Leeuw 1977)
$E = V^2, \lambda = \mu = 1, \nu = -2$	Kamada & Kawai (1989)
$E = V^2, \lambda = \mu = 1, \nu = -1$	Sammon (1969)
$\lambda = \mu = 1, \nu = 0$	LMDS (Chen and Buja 2009)
$\lambda = 1/3, \mu = 0, D_{i,j} = 1, t = 1$	Fruchterman & Reingold (1991)
$\lambda = 1/4, \mu = -2, D_{i,j} = 1, t = 1$	Davidson & Harel (1996)
$\lambda = 1, \mu = 0, D_{i,j} = 1, t = 1$	Noack's LinLog (2003)
$\lambda = \mu = 1, D_{i,j} = 1, t = 1$	Noack's QuadLin (2003)
$\lambda > 0, \mu = 0, D_{i,j} = 1, t = 1$	Noack's PolyLog family (2003; his $r = 1/\lambda$)

Table 2: *Some special cases of energy functions. Where $\nu \neq 0$, the table refers to equation (8), else to equation (9). For unweighted graphs, $D_{i,j} = 1$, the power ν plays no role.*

3.4 An Irrelevant Constant: Weighting the Attraction

Noack (2003, Sec. 5.5) observed that for his LinLog energy function the relative weighting of the attractive energy relative to the repulsive energy is irrelevant in the sense that such weighting would only change the scale of the minimizing layout but not the shape. A similar statement can be made for all members of the B-C family of energy functions. To demonstrate this effect, we introduce B-C energies whose attraction is weighted by a factor $a^{\frac{1}{\lambda}}$ where $a > 0$:

$$U_a(d) = \sum_{(i,j) \in E} \left(a^{\frac{1}{\lambda}} BC_{\mu + \frac{1}{\lambda}}(d_{i,j}) - D_{i,j}^{\frac{1}{\lambda}} BC_{\mu}(d_{i,j}) \right) - t^{\frac{1}{\lambda}} \sum_{(i,j) \notin E} BC_{\mu}(d_{i,j}).$$

We denote by $d = (d_{i,j})$ the set of all configuration distances for all edges (i, j) , including those not in the graph E . Note that the repulsion terms are still differentially weighted depending on whether (i, j) is an edge of the graph E or not, which is in contrast to most energy functions proposed in the graph layout literature where invariably $t = 1$.

In analogy to Noack's argument, we observe the following form of scale equivariance:

$$U_1(ad) = a^{\mu} U_a(d) + \text{const} \quad (11)$$

As a consequence, if d is a minimizing set of configuration distances for $U_a(\cdot)$, then the distances ad of the configuration scaled by a factor a minimize the original unweighted B-C energy $U_1(\cdot)$. This makes sense as an example shows: for $\lambda = 1$ and $a = 2$, a set of configura-

tion distances d is minimal for $U_a(\cdot)$ iff $2d$ is minimal for $U_1(\cdot)$, that is, the increased attraction of $U_a(\cdot)$ shrinks the solution by a factor of $a = 2$ compared to $U_1(\cdot)$.

It is in this sense that Noack’s PolyLog family of energies can be considered as a special case of the B-C family: PolyLog energies agree with B-C energies for unweighted graphs ($D_{i,j} = 1$) for $\mu = 0$ and $t = 1$ up to a multiplicative factor of the attractive force.

3.5 Unit-Invariant Forms of the Repulsion Weight

In the B-C family of energy functions (10), the relative strength of attractive and repulsive forces is balanced by the parameter t . This parameter, however, has two deficiencies: (1) It suffers from a lack of invariance under a change of units in the target distances $D_{i,j}$; (2) it has stronger effects in sparse graphs than dense graphs because the number of terms in the summations over E and $V \setminus E$ vary with the size of the graph E . Both deficiencies can be corrected by reparametrizing t in terms of a new parameter τ as follows:

$$t^{\frac{1}{\lambda}} = \frac{|E|}{|V^2| - |E|} \cdot (\text{median}_{(i,j) \in E} D_{i,j})^{\frac{1}{\lambda}} \cdot \tau^{\frac{1}{\lambda}}. \quad (12)$$

This new parameter τ is unit free and adjusted for graph size. (Obviously the median can be replaced with any other statistic $S(D)$ that is positively homogeneous of first order: $S(cD) = cS(D)$ for $c > 0$.) These features enable us to formulate past experience in a problem-independent fashion as follows: in the examples we have tried, $\tau = 1$ has yielded satisfactory results. In light of this experience, there may arise few occasions in practice where there is a need to tune τ . As users work with different units in $D_{i,j}$ or different neighborhood sizes when defining NN-graphs, the recommendation $\tau = 1$ stands. Just the same, we will illustrate the effect of varying τ in an artificial example below.

3.6 Size-Invariant Forms of B-C Stress/Energy

It is a common experience that algorithms for minimizing energy spend much effort on getting the size of the embedding right. Size, however, is not of interest — shape is. We have therefore a desire to re-express energy in a manner that is independent of size. Fortunately, there exists a general method that achieves this goal: For any configuration, minimize energy with regard to size and replace the original energy with its size-minimized value. This works because the minimization with regard to size can be carried out explicitly with elementary calculus. The result is a new form of energy that is minimized by the same shapes as the original energy, but it is independent of size and hence purely driven by shape. The computational advantage of size-invariant energy is that gradient-based optimization descends along directions that change shape, not size. — We sketch the derivation.

It is convenient to collect the repulsion terms inside and outside the graph because they share the power law:

$$U = \sum_{(i,j) \in E} BC_{\mu+\frac{1}{\lambda}}(d_{i,j}) - \sum_{(i,j) \in V^2} \tilde{D}_{i,j}^{\frac{1}{\lambda}} BC_{\mu}(d_{i,j}).$$

where

$$\tilde{D}_{i,j} = \begin{cases} D_{i,j}, & (i,j) \in E \\ t, & (i,j) \notin E \end{cases}$$

Next, consider a configuration $X = (x_i)_{i=1\dots N}$ and resized versions $sX = (s x_i)_{i=1\dots N}$ thereof ($s > 0$). The configuration distances scale along with size: $d_{i,j}(sX) = s d_{i,j}(X)$. To find the stationary size factor s of the energy as a function of s , $U = U(s)$, we observe that

$$\frac{\partial}{\partial s} BC_{\mu}(s d_{i,j}) = s^{\mu-1} d_{i,j}^{\mu} \quad (\forall \mu \in \mathbb{R}).$$

In particular, this holds even for $\mu = 0$. Next we solve the stationary equation and check second derivatives:

$$U(s) = \sum_E BC_{\mu+\frac{1}{\lambda}}(s d_{i,j}) - \sum_{V^2} \tilde{D}_{i,j}^{\frac{1}{\lambda}} BC_{\mu}(s d_{i,j}), \quad (13)$$

$$U'(s) = s^{\mu+\frac{1}{\lambda}-1} S - s^{\mu-1} T, \quad (14)$$

$$U''(s) = (\mu + \frac{1}{\lambda} - 1) s^{\mu+\frac{1}{\lambda}-2} S - (\mu - 1) s^{\mu-2} T, \quad (15)$$

where $S = S(d)$ and $T = T(d)$ are defined for $d = (d_{i,j})$ by

$$S = \sum_E d_{i,j}^{\mu+\frac{1}{\lambda}}, \quad T = \sum_{V^2} \tilde{D}_{i,j}^{\frac{1}{\lambda}} d_{i,j}^{\mu},$$

and again (14) and (15) hold even for $\mu = 0$ and $\mu + \frac{1}{\lambda} = 0$. The stationary size factor s_* that satisfies $U'(s_*) = 0$ is

$$s_* = \left(\frac{T}{S} \right)^{\lambda} \quad (\forall \mu \in \mathbb{R}, \lambda > 0). \quad (16)$$

The factor s_* is a strict minimum:

$$U''(s_*) = \frac{1}{\lambda} \frac{T^{\lambda\mu+1-2\lambda}}{S^{\lambda\mu-2\lambda}} > 0 \quad (\forall \mu \in \mathbb{R}, \lambda > 0).$$

Evaluating $U(s_*)$ we arrive at a **size-invariant** yet **shape-equivalent** form of the energy function. For the evaluation we need to separate power laws from the two logarithmic cases:

$$U(s) \approx \begin{cases} \frac{1}{\mu+\frac{1}{\lambda}} s^{\mu+\frac{1}{\lambda}} S - \frac{1}{\mu} s^{\mu} T & (\mu + \frac{1}{\lambda} \neq 0, \mu \neq 0) \\ |E| \log(s) + \sum_E \log(d_{i,j}) - \frac{1}{\mu} s^{\mu} T & (\mu + \frac{1}{\lambda} = 0) \\ \lambda s^{\frac{1}{\lambda}} S - \sum_{V^2} \tilde{D}_{i,j}^{\frac{1}{\lambda}} (\log(s) + \log(d_{i,j})) & (\mu = 0) \end{cases}$$

where “ \approx ” means “equal up to additive constants that are irrelevant for optimization.” We calculate $\tilde{U} = U(s_*)$ separately in the three cases with s_* from (16):

- $\mu + \frac{1}{\lambda} \neq 0, \mu \neq 0$: Several algebraic simplifications produce the following.

$$\tilde{U} = \left(\frac{1}{\mu + \frac{1}{\lambda}} - \frac{1}{\mu} \right) \frac{\left(\sum_{V^2} \tilde{D}_{i,j}^{\frac{1}{\lambda}} d_{i,j}^\mu \right)^{\lambda\mu+1}}{\left(\sum_E d_{i,j}^{\mu+\frac{1}{\lambda}} \right)^{\lambda\mu}}.$$

[Note that \tilde{U} gets minimized, hence the ratio on the right gets minimized when the left factor is positive (i.e., $\mu < 0 < \mu + 1/\lambda$), and it gets maximized when the left factor is negative (i.e., $\mu > 0$ or $\mu + 1/\lambda < 0$.)]

- $\mu + \frac{1}{\lambda} = 0$: We take advantage of the fact that $S = |E|$ and $\mu = -\frac{1}{\lambda}$.

$$\tilde{U} \approx |E| \lambda \log \sum_{V^2} \left(\frac{\tilde{D}_{i,j}}{d_{i,j}} \right)^{\frac{1}{\lambda}} + \sum_E \log(d_{i,j})$$

- $\mu = 0$: We take advantage of the fact that $S = \sum_E d_{i,j}^{\frac{1}{\lambda}}$ and also that $T = \sum_{V^2} \tilde{D}_{i,j}^{\frac{1}{\lambda}}$ is a constant for optimization with regard to $d = (d_{i,j})$, and any additive term that is just a function of $\tilde{D}_{i,j}$ but not $d_{i,j}$ can be neglected.

$$\tilde{U} \approx \lambda \left(\sum_{V^2} \tilde{D}_{i,j}^{\frac{1}{\lambda}} \right) \log \left(\sum_E d_{i,j}^{\frac{1}{\lambda}} \right) - \sum_{V^2} \tilde{D}_{i,j}^{\frac{1}{\lambda}} \log(d_{i,j})$$

Even though size-invariance holds by construction, one checks it easily in all three cases: $\tilde{U}(sd) = \tilde{U}(d)$.

4 Meta-Criteria For Parameter Selection

In previous work (Chen and Buja 2009) the authors introduced “meta-criteria” to measure the quality of configurations independently of the primary energy/stress functions. The main purpose of these meta-criteria is to guide the selection of parameters such as those in the B-C family, λ , μ and τ . The idea is to compare “input neighborhoods” defined in terms of $D_{i,j}$ with “output neighborhoods” defined in terms of $d_{i,j}$ by measuring the size of their overlaps. Such neighborhoods are typically constructed as K -NN sets or, less frequently, in metric terms as ϵ -neighborhoods. In a dimension reduction setting one may define for the i 'th point the input neighborhood $\mathcal{N}_D(i)$ as the set of K -NNs with regard to $D_{i,j}$ and similarly the output neighborhood $\mathcal{N}_d(i)$ as the set of K -NNs with regard to $d_{i,j}$. In an unweighted graph setting, one may define $\mathcal{N}_D(i)$ as the metric $\epsilon = 1$ neighborhood, that is, the set of points connected with the i 'th point in the graph E , and hence the neighborhood size $K(i) = |\mathcal{N}_D(i)|$ is the degree of the i 'th point in the graph E and will vary from point to point. The corresponding output neighborhood $\mathcal{N}_d(i)$ can then be defined as the $K(i)$ -NN set with regard to $d_{i,j}$. The

pointwise meta-criterion at the i 'th point is defined as size of the overlap between $\mathcal{N}_d(i)$ and $\mathcal{N}_D(i)$, hence it is in frequency form

$$N_d(i) = |\mathcal{N}_d(i) \cap \mathcal{N}_D(i)|,$$

and in proportion form, using $|\mathcal{N}_D(i)|$ as the baseline,

$$M_d(i) = \frac{|\mathcal{N}_d(i) \cap \mathcal{N}_D(i)|}{|\mathcal{N}_D(i)|}.$$

The global meta-criteria are simply the averages over all points:

$$N_d = \frac{1}{|V|} \sum_i N_d(i) \quad \text{and} \quad M_d = \frac{1}{|V|} \sum_i M_d(i).$$

Only when all input neighborhood sizes are equal, $|\mathcal{N}_D(i)| = K$, is there a simple relationship between N_d and M_d : $M_d = \frac{1}{K} N_d$. We subscript these quantities with d because they serve to compare different outputs $(\mathbf{x}_i)_{i=1\dots N}$ (configurations, embeddings, graph drawings), but all that is used are the interpoint distances $d_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\|$. The proportion form M_d is obviously advantageous because it allows comparisons across different K (or ϵ).

Whether the meta-criterion values are small or large should be judged not against their possible ranges ($[0, 1]$ for M_d) but against the possibility that $d_{i,j}$ (hence the embedding) and $D_{i,j}$ are entirely unrelated and generate only random overlap in their respective neighborhoods $\mathcal{N}_d(i)$ and $\mathcal{N}_D(i)$. The expected value of random overlap is not zero, however; rather, it is $E[|\mathcal{N}_d(i) \cap \mathcal{N}_D(i)|] = |\mathcal{N}_d(i)| \cdot |\mathcal{N}_D(i)| / (|V| - 1)$ because random overlap should be modeled by a hypergeometric distribution with $|\mathcal{N}_D(i)|$ “defectives” and $|\mathcal{N}_d(i)|$ “draws” from a total of $|V| - 1$ “items.” The final adjusted forms of the meta-criteria are therefore:

$$\begin{aligned} N_d^{adj}(i) &= |\mathcal{N}_d(i) \cap \mathcal{N}_D(i)| - \frac{1}{|V| - 1} |\mathcal{N}_d(i)| \cdot |\mathcal{N}_D(i)|, \\ M_d^{adj}(i) &= \frac{|\mathcal{N}_d(i) \cap \mathcal{N}_D(i)|}{|\mathcal{N}_D(i)|} - \frac{1}{|V| - 1} |\mathcal{N}_d(i)|, \\ N_d^{adj} &= \frac{1}{|V|} \sum_i N_d^{adj}(i), \quad M_d^{adj} = \frac{1}{|V|} \sum_i M_d^{adj}(i). \end{aligned}$$

When the neighborhoods are all K -NN sets, $|\mathcal{N}_d(i)| = |\mathcal{N}_D(i)| = K$, these expressions simplify:

$$\begin{aligned} N_d^{adj}(i) &= |\mathcal{N}_d(i) \cap \mathcal{N}_D(i)| - \frac{K^2}{|V| - 1}, \\ M_d^{adj}(i) &= \frac{|\mathcal{N}_d(i) \cap \mathcal{N}_D(i)|}{K} - \frac{K}{|V| - 1} = \frac{N_d^{adj}(i)}{K}, \\ N_d^{adj} &= N_d - \frac{K^2}{|V| - 1}, \quad M_d^{adj} = M_d - \frac{K}{|V| - 1} = \frac{N_d^{adj}}{K}. \end{aligned}$$

An important general observation is that if the neighborhoods are defined as K -NN sets, the meta-criteria are invariant under monotone transformations of both inputs $D_{i,j}$ and outputs $d_{i,j}$. Methods that have this invariance are called “non-metric” in proximity analysis/multidimensional scaling because they depend only on the ranks and not the actual values of the distances.

In what follows, we will report M_d^{adj} for each configuration shown in the figures, and we will also use the pointwise values $M_d(i)$ as a diagnostic by highlighting points with $M_d(i) < 1/2$ as problematic in some of the figures.

5 B-C ENERGIES APPLIED

5.1 Simulated Data

We introduced three parameters in the B-C energy functions, namely, λ , μ and τ , and we interpreted them according to intuitions borrowed from physics. In this subsection, we will examine how these parameters affect configurations in terms of their local and global structure by experimenting with an artificial example.

The artificial data set consists of 83 points which form a geometric shape represented in Figure 3 (top). The design of this example was inspired by a simulation example used by Trosset (2006). The distance between any pair of adjacent points is set to 1. To define an initial local graph, as input to the energy functions, we connected each point with its adjacent neighbors with distance 1 (Figure 3, bottom). That is, we use metric nearest neighborhoods with radius 1. Thus, interior points have node degree 4, corner points and connecting points have node degree 2, and the remaining peripheral points have node degree 3. The geometry of this input graph is intended to represent three connected clusters with an internal structure that is relatively tight compared to the connecting structure.

We produced 2-D configurations using B-C energy functions, and to explore the effect of the parameters on the configurations, we varied each of the three parameters one at a time. For each combination of parameters, we use two different starting configurations: the inputs from Figure 3, and a random start, respectively. By starting from the inputs, the optimization has no need to find a globally correct structure and can focus on figuring out the local structure, whereas starting from a random configuration indicates how stable the solutions will be for given parameter combinations. For starts from the inputs, the results are shown in Figures 4, 6, and 8, and for starts from random configurations they are shown in Figures 5, 7, and 9, along with their M^{adj} values that measure the local faithfulness of the configuration. We also colored red the points whose neighborhood structure is not well preserved in terms of a proportion $< 1/2$ of shared neighbors between input and output configurations ($M(i) < 1/2$).

Parameter λ : We set $\mu = 0$ and $\tau = 1$ and let λ vary as follows: $\lambda = 0.2, 0.5, 1, 1.5$. The resulting configurations are shown in Figures 4 and 5. The overall observation from both figures is that the clustering effect is stronger in the configurations with larger λ values. This confirms the role of λ as the parameter that controls the clustering effect. For example, the bottom-right panels with the largest λ show the strongest clustering effect. We also notice in both figures that M^{adj} increases with increasing λ , which indicates local structure within clusters is better recovered when the clusters are well separated. To confirm this, we zoom in on the nearly collapsed points in the bottom right configurations and observe the square structure in the input configuration is almost perfectly recovered. This indicates that by tuning λ properly the resulting configurations can reveal both macro structure in terms of relative cluster placement as well as micro structure within each cluster.

Parameter μ : In Figures 6 and 7, we examine the effect of μ . We fix $\lambda = 0.2$ and $\tau = 1$ and let μ vary as follows: $\mu = -1, 0, 1, 2$. An overall observation is that the larger μ , the more spread out are the points. This is consistent with the interpretation of μ as the power law of repulsion. With smaller μ (such as $\mu = -1$) the energy function flattens out as the distance increases (Figure 2) and the points are subject to very weak repulsion. The top left plots ($\mu = -1$) in both figures show that weak repulsion is suitable for generating locally faithful structure. However, the repulsion can be too weak to generate globally meaningful structure, as illustrated by the configurations obtained from random starts (Figure 7). In the top left plot of Figure 7, the three clusters are not aligned properly due to the weak repulsion. With stronger repulsion the points are placed in a globally more correct position, as shown in bottom two plots in Figure 7, with a sacrifice of local faithfulness (as reflected by the lower value of M^{adj}). The distortion of local structure is not surprising considering the fact that the repulsion is stronger in the direction in which points line up, which in this case is the horizontal direction. By comparison, points are squeezed flat in the vertical direction because repulsion has no traction vertically.

Parameter τ : We fix $\lambda = 0.2$ and $\mu = 0$ and vary τ as follows: $\tau = 0.01, 1, 10^3, 10^5$. The configurations from original design and a random start are shown in Figures 8 and 9. Figure 8 shows that the configuration closest to the input configuration is achieved with relatively small τ ($\tau = 0.01$). This indicates that the B-C energy functions for small τ are quite successful in recreating local distances as they are supposed to. From a random start, however, the configuration can be easily trapped in a local minimum with small λ (Figure 9, top two plots), which indicates that the relative weight of repulsion to attraction controlled by τ plays an essential role in achieving a stable configuration. With relatively larger τ , the points are more spread-out and configurations reveal the underlying structure more faithfully, both locally and globally (bottom two plots, Figure 9).

5.2 Olivetti Faces Data

This dataset, published on Sam Roweis' website (<http://www.cs.toronto.edu/~roweis/data.html>), contains 400 facial images of 40 people, 10 images for each person. All images are of size 64 by 64. Mathematically, each image can be represented by a long vector, each element of which records the light intensity of one pixel in the image. Given this representation, we treat each image as a data point lying in a 4096-dimensional space ($64 \times 64 = 4096$). For visualization purposes we reduce the dimension from 4096 to 2. As the 40 faces form natural groups, we would expect effective dimension reduction methods to show clusters in their configurations. If this expectation is correct, then this dataset provides an excellent test bed for the effect of the clustering power λ .

We first centered the data, a 400×4096 matrix, at their row means to adjust the brightness of the images to the same level. We constructed a pairwise distance matrix using Euclidean distances in the original high dimensional space R^{4096} . We then defined a local graph using 4-NN, that is, we connected each point to its four nearest neighbors. In the resulting graph five small components were disconnected from the main component of the graph. Each of them contained images from a single person, with 5 images for one person and 5 images for another four persons. Since the disconnected components are trivially pushed away from the main component in any embedding due to the complete absence of attraction, we discarded them and kept the 355 images representing 36 people for further analysis. We created for each person a unique combination of color and symbol to code the points representing it.

Figure 10 shows 2D configurations generated by different energy functions (10) with different clustering powers, $\lambda = 0.5, 1, 1.5, 2$, while the other parameters are fixed at $\mu = 0$ and $\tau = 1$. With relatively small λ ($\lambda = 0.5, 1$), we either do not see any clusters or see fussy clusters. As λ increases the clusters become clearer. The coloring and symboling show that these clusters are not artificial but real, mostly representing the images of the same person. The configurations do not produce exactly 36 clusters: some images of different people are not quite distinguishable in the configurations and some images of the same person are torn apart. The former could really present similar images. The latter could be due to the placement of images from the random start. However, the overall impression indeed confirms the clustering effect of λ . An interesting observation from this experiment is that the M_k^{adj} increases as clustering effect becomes stronger, which assures us with the faithfulness of local topology.

Figure 11 shows 2D configurations generated from four popular dimension reduction methods: PCA, MDS, Isomap and LLE. PCA and MDS did not find any clusters. Isomap and LLE did find a couple of clusters, but not as many nor as clearly as our method, even though we tuned the neighborhood size for Isomap and LLE to achieve the best visualization. For example, we chose a different neighborhood size $K = 8$ for LLE and Isomap $K = 4$; LLE

configurations degenerated to lines or lines and a big cluster when $K = 4$ and 6 , respectively.

5.3 Frey Face Data

In the Olivetti data, we were able to see the clustering effect of λ . In the present example we study the effect of μ and its interaction with λ . The Frey face data was originally published with the LLE article (Roweis and Saul, 2000). We studied its low dimensional configurations from various dimension reduction methods in Chen and Buja (2009). The data contains 1965 facial images of “Branden Frey,” which are stills of a short video clip recorded when Frey was making different facial expressions. Each image is of size 20×28 which can be thought of as a data point in 560-dimensional space. In our experiments we use a subset of 500 images in order to save on computations and in order to obtain less cluttered low dimensional embeddings. The fact is that the intrinsic structure of the full dataset is well preserved in this subset, partly due to the inherent redundancies in video sequences: the images close in order are very similar because the stills are taken more frequently than Frey’s facial expression changes.

In Figure 12 we show the first two principal components of the 3D configurations for varying μ as λ and τ are fixed at one. The neighborhood size K is six. The coloring scheme is adapted from the LMDS configurations of Chen and Buja (2009) where the points were colored to highlight the clustering structure found in those configurations. The bottom left configuration with parameters $\lambda = \mu = 1$ is an LMDS configuration. We found that subsampling changes little of the overall structure in the configurations. We observe that the clustering effect of a fixed λ value 1 varies as μ varies. With smaller μ such as -1 and 0 , the clustering effect is most pronounced: bigger clusters in the configurations with larger μ are split into smaller ones. On the other hand, larger μ such as 1 and 2 clearly provide connectivity between clusters and therefore better capture the global structure in the data. The local neighborhood structure, though, is better reflected in the configurations with smaller values of μ , as suggested by the values of meta-criteria.

6 Summary and Discussion

Our work contributes to the literature on proximity analysis, non-linear dimension reduction and graph drawing by systematizing the class of distance-based approaches whose commonality is that they generate maps (embeddings, configurations, graph drawings) of objects in such a way that given input distances between the objects are well-approximated by output distances in the map. The systematization consists of devising a three-parameter family of loss (stress, energy) functions that comprises many published proposals from the literatures on proximity analysis (multidimensional scaling) and graph drawing. A benefit of this endeavor is that the

seemingly arbitrary selection of a loss function is turned into a parameter selection problem, for which we propose a systematic approach based on “meta-criteria” that measure the quality of maps independently of the loss functions. The parameters of the proposed family have the following interpretations:

- λ : The clustering power which, for increasing λ , increases the clustering strength, i.e., the tendency to group points in the embedding. Range: $\lambda > 0$.
- μ : The power law of repulsion, which is balanced against the power law of attraction, $\mu + \frac{1}{\lambda}$. The greater μ , the greater the sensitivity to large discrepancies between $d_{i,j}$ and $D_{i,j}$. Range: $-\infty < \mu < +\infty$.
- τ : A regularization parameter that stabilizes configurations when only local input distances are used, at the cost of some bias (stretching of configurations), achieved by imputing infinite input distances with infinitesimal repulsion. Range: $\tau > 0$.

Because the power laws are implemented using Box-Cox transformations, they encompass logarithmic laws as well ($\mu = 0$ and $\mu + \frac{1}{\lambda} = 0$, respectively). The regularization parameter τ plays a role only when the input distance matrix is incomplete, as in the case of a graph or in the case of localization by restricting the loss function to small distances (as in LMDS; Chen and Buja 2009). If the input distances are complete, as assumed in proximity analysis with multidimensional scaling, the loss functions form a two-parameter family of stress functions.

The problem of incomplete distance information is often solved by completing it with additive imputations provided by the shortest-path algorithm, so that MDS-style stress functions can be used — the route taken by Isomap (Tenenbaum et al. 2000). The argument against such completion is that stress functions tend to be driven by the largest distances, which are imputed and hence noisy. Conversely the argument against not completing is that the use of pervasive repulsion to stabilize configurations amounts to imputation also, albeit of an uninformative kind. A full understanding of the trade-offs between completion and repulsion is currently lacking, but practitioners can meanwhile experiment with both approaches and compare them on their data. In both cases the family of loss functions proposed here offers control over the clustering power λ and the repulsion power μ .

Another issue with distance-based approaches is that there is often much freedom in choosing the distances, in particular when applied to dimension reduction. There is therefore a need to systematize the choices and provide guidance for “distance selection.”

APPENDIX: ENERGY MINIMIZATION

6.1 Gradients for Energy Functions

Let the $n \times d$ matrix $X = (x_1, \dots, x_n)^T$ represent the set of points in d dimensions. As always let D_{ij} be the input distances and $d_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\|$ the output distances. The B-C energy function for $\mu \neq 0$ and $\mu + \frac{1}{\lambda} \neq 0$ is

$$U(\mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{(i,j) \in E} \left(\frac{d_{i,j}^{\mu + \frac{1}{\lambda}} - 1}{(\mu + \frac{1}{\lambda})} - D_{ij}^{\frac{1}{\lambda}} \frac{d_{i,j}^{\mu} - 1}{\mu} \right) - t^{\frac{1}{\lambda}} \sum_{(i,j) \in E^c} \frac{d_{i,j}^{\mu} - 1}{\mu} \quad (17)$$

Let $\nabla U = (\nabla_1, \dots, \nabla_n)^T$ be the gradient of the energy function with respect to X :

$$\nabla_i = \frac{\partial U}{\partial \mathbf{x}_i} = \sum_{j \in E(i)} \left(d_{i,j}^{\mu + \frac{1}{\lambda} - 2} - D_{ij}^{\frac{1}{\lambda}} d_{i,j}^{\mu - 2} \right) (\mathbf{x}_i - \mathbf{x}_j) - t^{\frac{1}{\lambda}} \sum_{j \in N^c(i)} d_{i,j}^{\mu - 2} (\mathbf{x}_i - \mathbf{x}_j)$$

Define a $N \times N$ matrix M as follows:

$$M_{ji} = \begin{cases} d_{i,j}^{\mu + \frac{1}{\lambda} - 2} - D_{ij}^{\frac{1}{\lambda}} d_{i,j}^{\mu - 2} & \text{if } j \in E(i) \\ -t^{\frac{1}{\lambda}} d_{i,j}^{\mu - 2} & \text{if } j \notin E(i) \end{cases}$$

Note that M is symmetric. The gradient can be simplified to

$$\begin{aligned} \nabla_i &= \frac{\partial U}{\partial \mathbf{x}_i} = \sum_j M_{ji} (\mathbf{x}_i - \mathbf{x}_j) \\ &= \left(\sum_j M_{ji} \right) \mathbf{x}_i - \sum_j M_{ji} \mathbf{x}_j, \end{aligned}$$

and

$$\nabla U = X * (M \cdot E) - M \cdot X,$$

where E is a $N \times d$ matrix with all elements being 1. The symbol '*' represents elementwise multiplication of the two matrices of the same size, and the symbol '.' stands for regular matrix multiplication

6.2 Gradients for Size-Invariant Energy Functions

We write the energy function as follows:

$$\begin{aligned}
U(d) &= \sum_{(i,j) \in E} \left(\frac{d_{i,j}^{\mu+\frac{1}{\lambda}} - 1}{\mu + \frac{1}{\lambda}} - D_{i,j}^{\frac{1}{\lambda}} \frac{d_{i,j}^{\mu} - 1}{\mu} \right) - t^{\frac{1}{\lambda}} \sum_{(i,j) \notin E} \frac{d_{i,j}^{\mu} - 1}{\mu} \\
&= \sum_{(i,j) \in E} \frac{d_{i,j}^{\mu+\frac{1}{\lambda}} - 1}{\mu + \frac{1}{\lambda}} - \sum_{(i,j) \in V^2} \tilde{D}_{i,j} \frac{d_{i,j}^{\mu} - 1}{\mu} \\
&\sim \sum_{(i,j) \in E} \frac{d_{i,j}^{\mu+\frac{1}{\lambda}}}{\mu + \frac{1}{\lambda}} - \sum_{(i,j) \in V^2} \tilde{D}_{i,j} \frac{d_{i,j}^{\mu}}{\mu}
\end{aligned}$$

where $\tilde{D}_{ij} = \begin{cases} D_{ij}^{\frac{1}{\lambda}}, & (i,j) \in E \\ t^{\frac{1}{\lambda}}, & (i,j) \notin E \end{cases}$ and \sim denotes equal up to a constant.

Consider a scaling factor γ

$$U(\gamma d) = \gamma^{\mu+\frac{1}{\lambda}} \sum_{(i,j) \in E} \frac{d_{i,j}^{\mu+\frac{1}{\lambda}}}{\mu + \frac{1}{\lambda}} - \gamma^{\mu} \sum_{(i,j) \in V^2} \tilde{D}_{i,j} \frac{d_{i,j}^{\mu}}{\mu}$$

The derivative with respect to γ

$$U'_{\gamma}(\gamma d) = \gamma^{\mu+\frac{1}{\lambda}-1} \sum_{(i,j) \in E} d_{i,j}^{\mu+\frac{1}{\lambda}} - \gamma^{\mu-1} \sum_{(i,j) \in V^2} \tilde{D}_{i,j} d_{i,j}^{\mu}$$

Set $U'_{\gamma}(\gamma d) = 0$, the optimal scaling γ^* has the following representation

$$\gamma^* = \left(\frac{\sum_{(i,j) \in V^2} \tilde{D}_{i,j} d_{i,j}^{\mu}}{\sum_{(i,j) \in E} d_{i,j}^{\mu+\frac{1}{\lambda}}} \right)^{\lambda}$$

The optimal scaling γ^* is indeed the minimizer of the energy function $U(\gamma d)$ as $U''_{\gamma}(\gamma^* d) > 0$

$$U''_{\gamma}(\gamma^* d) = \frac{1}{\lambda} (\gamma^*)^{\mu+\frac{1}{\lambda}-2} \left(\sum_{(i,j) \in E} d_{i,j}^{\mu+\frac{1}{\lambda}} \right)$$

Plug γ^* into $U(\gamma d)$ we get the scale invariant stress function

$$\begin{aligned}
\tilde{U}(d) &= U(\gamma^* d) = \left(\frac{1}{\mu + \frac{1}{\lambda}} - \frac{1}{\mu} \right) (\gamma^*)^{\mu} \left(\sum_{(i,j) \in V^2} \tilde{D}_{i,j} d_{i,j}^{\mu} \right) \\
&= \left(\frac{1}{\mu + \frac{1}{\lambda}} - \frac{1}{\mu} \right) \frac{\left(\sum_{(i,j) \in V^2} \tilde{D}_{i,j} d_{i,j}^{\mu} \right)^{\lambda\mu+1}}{\left(\sum_{(i,j) \in E} d_{i,j}^{\mu+\frac{1}{\lambda}} \right)^{\lambda\mu}} \\
&\equiv \left(\frac{1}{\mu + \frac{1}{\lambda}} - \frac{1}{\mu} \right) \frac{T(d)^{\lambda\mu+1}}{S(d)^{\lambda\mu}}
\end{aligned}$$

where

$$T(d) = \sum_{(i,j) \in V^2} \tilde{D}_{i,j} d_{i,j}^\mu; \text{ and } S(d) = \sum_{(i,j) \in E} d_{i,j}^{\mu + \frac{1}{\lambda}}$$

To optimize $\tilde{U}(d)$ we use the gradient decent method. Let $\nabla U = ((\nabla U)_1, \dots, (\nabla U)_n)^T$ be the gradient the $\tilde{U}(d)$ with respect to configuration $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$

$$\begin{aligned} (\nabla U)_i &= \left(\frac{1}{\mu + \frac{1}{\lambda}} - \frac{1}{\mu} \right) \left[(\lambda\mu + 1) \left(\frac{T}{S} \right)^{\lambda\mu} (\nabla T)_i - (\lambda\mu) \left(\frac{T}{S} \right)^{\lambda\mu+1} (\nabla S)_i \right] \\ (\nabla T)_i &= \frac{\partial T(d)}{\partial \mathbf{x}_i} = \mu \sum_j \tilde{D}_{i,j} d_{i,j}^{\mu-2} (\mathbf{x}_i - \mathbf{x}_j) \\ (\nabla S)_i &= \frac{\partial S(d)}{\partial \mathbf{x}_i} = \left(\mu + \frac{1}{\lambda} \right) \sum_{j \in E(i)} d_{i,j}^{\mu + \frac{1}{\lambda} - 2} (\mathbf{x}_i - \mathbf{x}_j) \end{aligned}$$

Plug $(\nabla T)_i$ and $(\nabla S)_i$ into the $(\nabla U)_i$

$$\begin{aligned} (\nabla U)_i &= \left(\frac{T}{S} \right)^{\lambda\mu} \left(\frac{T}{S} \sum_{j \in E(i)} d_{i,j}^{\mu + \frac{1}{\lambda} - 2} (\mathbf{x}_i - \mathbf{x}_j) - \sum_j \tilde{D}_{i,j} d_{i,j}^{\mu-2} (\mathbf{x}_i - \mathbf{x}_j) \right) \\ &= \left(\frac{T}{S} \right)^{\lambda\mu} \left(\sum_{j \in E(i)} \left(\frac{T}{S} d_{i,j}^{\mu + \frac{1}{\lambda} - 2} - D_{j,i}^{\frac{1}{\lambda}} d_{i,j}^{\mu-2} \right) (\mathbf{x}_i - \mathbf{x}_j) - \sum_{j \in E^c(i)} t d_{i,j}^{\mu-2} (\mathbf{x}_i - \mathbf{x}_j) \right) \end{aligned}$$

Define a $N \times N$ matrix M by

$$M_{ij} = \begin{cases} \left(\frac{T}{S} d_{i,j}^{\mu + \frac{1}{\lambda} - 2} - D_{j,i}^{\frac{1}{\lambda}} d_{i,j}^{\mu-2} \right) & \text{for } j \in E(i) \\ t d_{i,j}^{\mu-2} & \text{for } j \notin E(i) \end{cases}$$

The gradient can be simplified to

$$\begin{aligned} (\nabla U)_i &= \left(\frac{T}{S} \right)^{\lambda\mu} \sum_j M_{ij} (\mathbf{x}_i - \mathbf{x}_j) \\ &= \left(\frac{T}{S} \right)^{\lambda\mu} \left(\sum_j M_{ij} \mathbf{x}_i - \sum_j M_{ij} \mathbf{x}_j \right) \end{aligned}$$

and

$$\nabla U = \left(\frac{T}{S} \right)^{\lambda\mu} (X * (M \cdot E) - M \cdot X)$$

We did the calculation separately for $\mu = 0$ and $\frac{1}{\lambda} + \mu = 0$ which showed that the representations of γ^* and M include these two cases by noting $z^0 = 1$ for $z \neq 0$. The scale invariant energy function $\tilde{U}(d)$ for the general case does not include these special cases.

$$\begin{aligned} \mu = 0 : \quad \tilde{U}(d) &\sim \lambda \left(\sum_{(i,j) \in V^2} \tilde{D}_{ij} \right) \log \left(\sum_{(i,j) \in E} d_{i,j}^{\frac{1}{\lambda}} \right) - \sum_{(i,j) \in V^2} \tilde{D}_{ij} \log d_{ij} \\ \mu + \frac{1}{\lambda} = 0 : \quad \tilde{U}(d) &\sim \lambda \left(\sum_{(i,j) \in E} \right) \log \left(\sum_{(i,j) \in V^2} \tilde{D}_{ij} d_{ij}^\mu \right) + \sum_{(i,j) \in E} \log d_{ij} \end{aligned}$$

REFERENCES

- Akkucuk, U. and Carroll, J.D., 2006, PARAMAP vs. Isomap: A Comparison of Two Non-linear Mapping Algorithms, *Journal of Classification*, **23** (2), 221-254.
- Brandes, U., 2001, Drawing on Physical Analogies, in: *Drawing Graphs*, Kaufmann, M. and Wagner, D., eds., Springer: Berlin, 71-86.
- Chen, L., 2006, Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Layout and Proximity Analysis, Ph.d. Thesis, University of Pennsylvania.
- Chen, L. and Buja, A., 2009, Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Drawing and Proximity Analysis, *Journal of American Statistical Association*, **104**, No. 485, 209-219.
- Davidson, R. and Harel, D., 1996, Drawing graphs nicely using simulated annealing, *ACM Transactions on Graphics*, **15**(4), 301-331.
- Di Battista, G., Eades, P., Tamassia, R., Tollis, I. G., 1999, Graph Drawing: Algorithms for the Visualization of Graphs. Prentice Hall.
- Eades P., 1984, A heuristic for graph drawing, In *Congressus Numerantium*, **42**, 149-160.
- Fruchterman, T. M. J. and Reingold, E. M., 1991, Graph drawing by force-directed placement, *Software-Practice and Experience*, **21**(11), 1129-1164.
- Gansner, E., Koren, Y., and North, S., 2004, Graph Drawing by Stress Majorization, *Graph Drawing*, 239-250.
- Graef J., and Spence, I., 1979, Using Distance Information in the Design of Large Multidimensional Scaling Experiments, *Psychological Bulletin*, **86**, 60-66.
- Kaufmann, M. and Wagner, D. (eds.), 2001, Drawing Graphs. Springer: Berlin.
- Kamada, T., and Kawai, S., 1989, An algorithm for drawing general undirected graphs. *Information Processing Letters*, **31**, 7-15.
- Kruskal, J. B., 1964a, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, **29**, 1-27.
- Kruskal, J. B., 1964b, Nonmetric multidimensional scaling: a numerical method, *Psychometrika*, **29**, 115-129.
- Kruskal, J. B. and Seery, J. B., 1980, Designing Network Diagrams, Technical Memorandum, Bell Laboratories, Murray Hill, NJ.
- Lu, F., Keles, S., Wright, S. J., and Wahba, G., 2005, Framework for kernel regularization with application to protein clustering, *PNAS* **102**, (35), 12332-12337 (<http://www.pnas.org/cgi/content/full/102/3>)

- Michailidis, G. and de Leeuw, J., 2001, Data visualization through graph drawing, *Computation. Stat.*, **16**, 435-50.
- Noack, A., 2003, Energy models for drawing clustered small-world graphs, Computer Science Report 07/2003, Brandenburg Technical University at Cottbus, Germany.
- Roweis S. T. and Saul, L. K., 2000, Nonlinear dimensionality reduction by local linear embedding, *Science*, **290**, 2323-2326.
- Sammon, J., 1969, A Non-Linear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*, **C-18**(5), 401-409.
- Schölkopf, B., Smola, A. J., and Müller, K.-R., 1998, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, **10**, 1299-1319.
- Saul, L. K. and Roweis, S. T., 2003, Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds, *J. of Machine Learning*, **4**, 119-155.
- Takane, Y., Young, F. W. and de Leew, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, **42**, 7-67.
- Tenenbaum, J. B., Silva, V. d., and Langford, J. C., 2000, A global geometric framework for nonlinear dimensionality reduction, *Science*, **290**, 2319-2323.
- Torgerson, W.S., 1952, *Psychometrika*, **17**, 401-419.
- Trosset, M.W., 2006, Classical Multidimensional Scaling and Laplacian Eigenmaps, a talk given at the 2006 Joint Statistical Meeting, Seattle, WA.
- Weinberger, K. Q., Sha, F., Zhu, Q., and Saul, L. K., 2006, Graph Laplacian Regularization for Large-Scale Semidefinite Programming. *NIPS 2006*, 1489-1496.

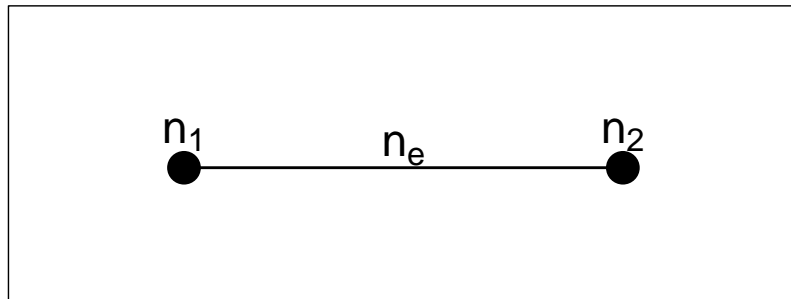


Figure 1: *Noack's simple graph with two subgraphs that have n_1 and n_2 internal vertices and n_e connecting edges.*

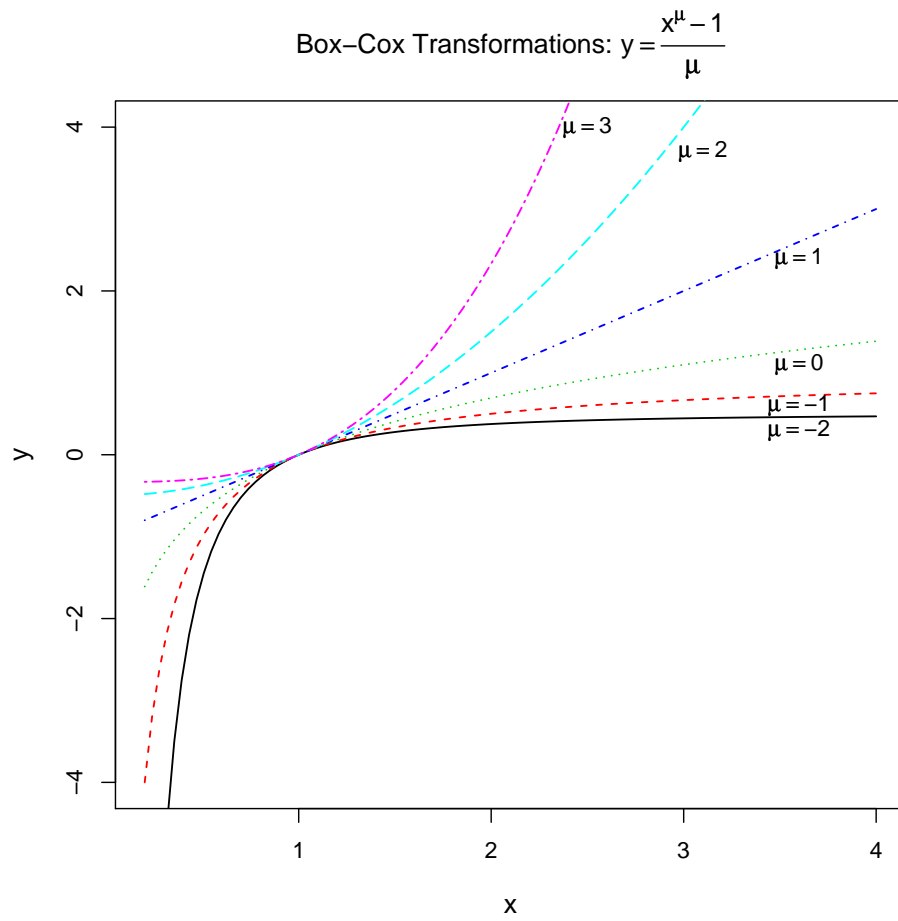


Figure 2: *Box-Cox Transformations: $y = \frac{x^\mu - 1}{\mu}$*

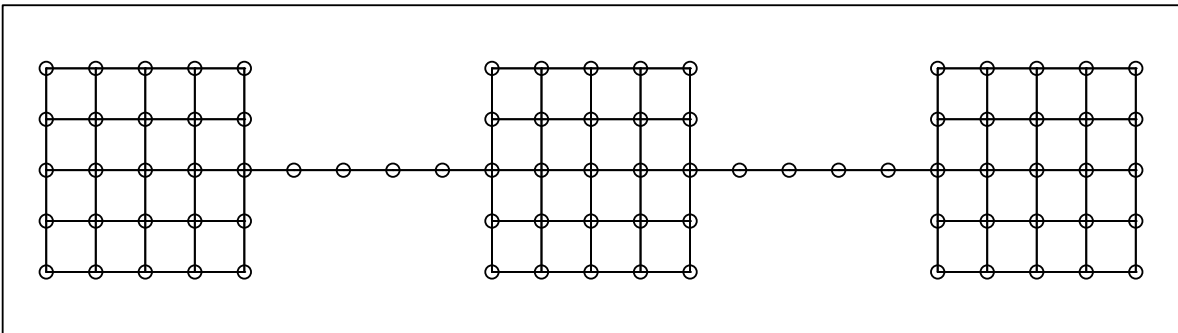
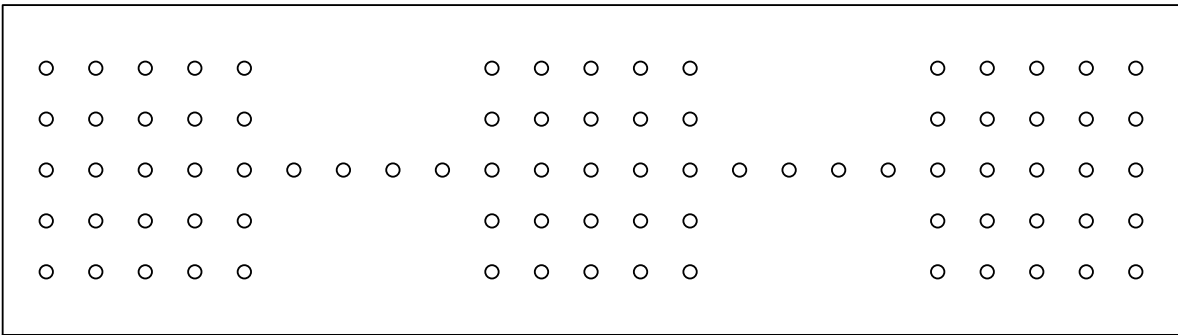


Figure 3: *The original configuration (top) and initial local graph (bottom)*

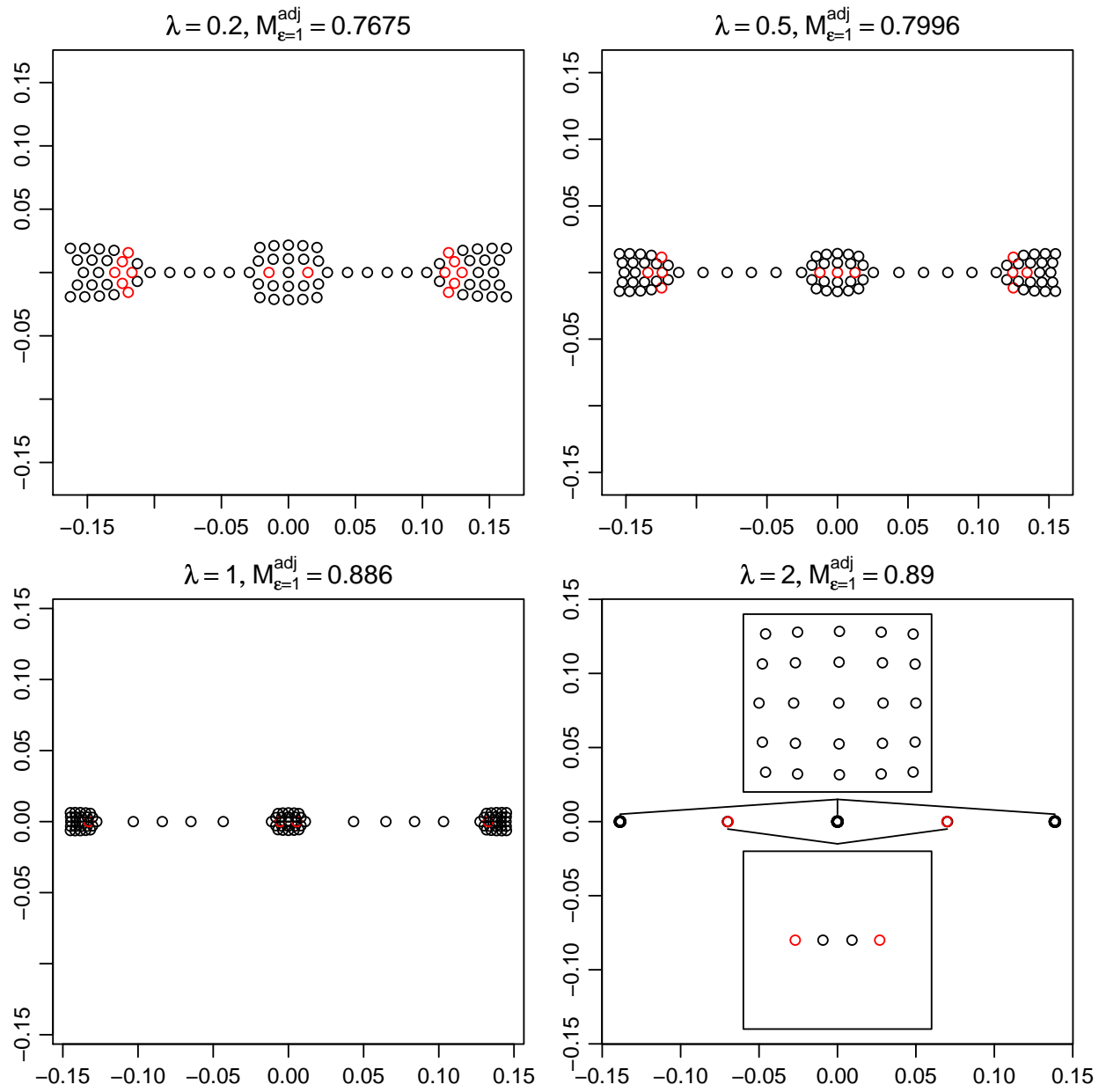


Figure 4: *The configurations with varying λ with other parameters fixed at $\mu = 0$, $\tau = 1$, starting from the input configuration (Figure 3).*

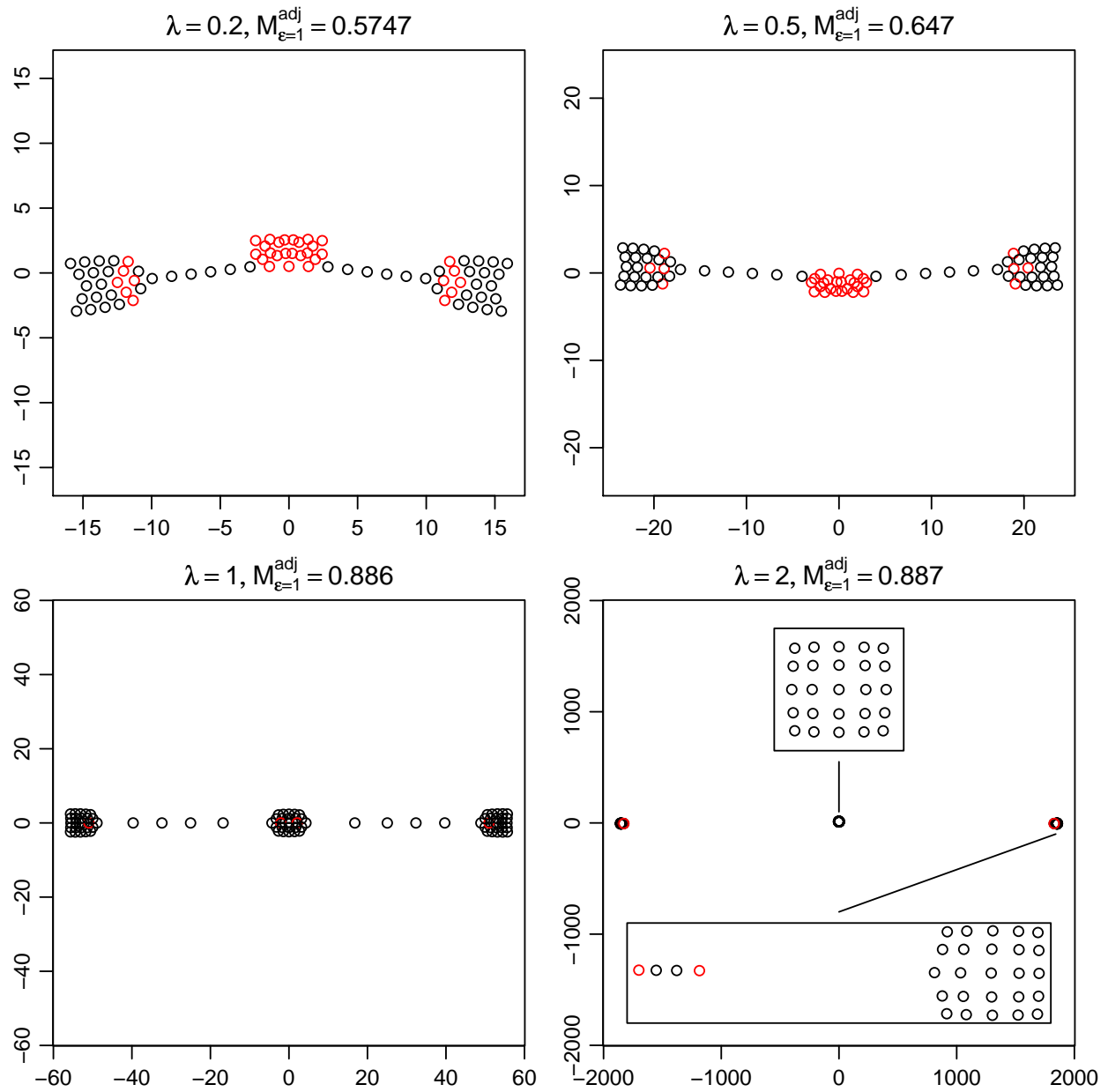


Figure 5: *The configurations with varying λ with other parameters fixed at $\mu = 0$, $\tau = 1$, starting from a random configuration.*

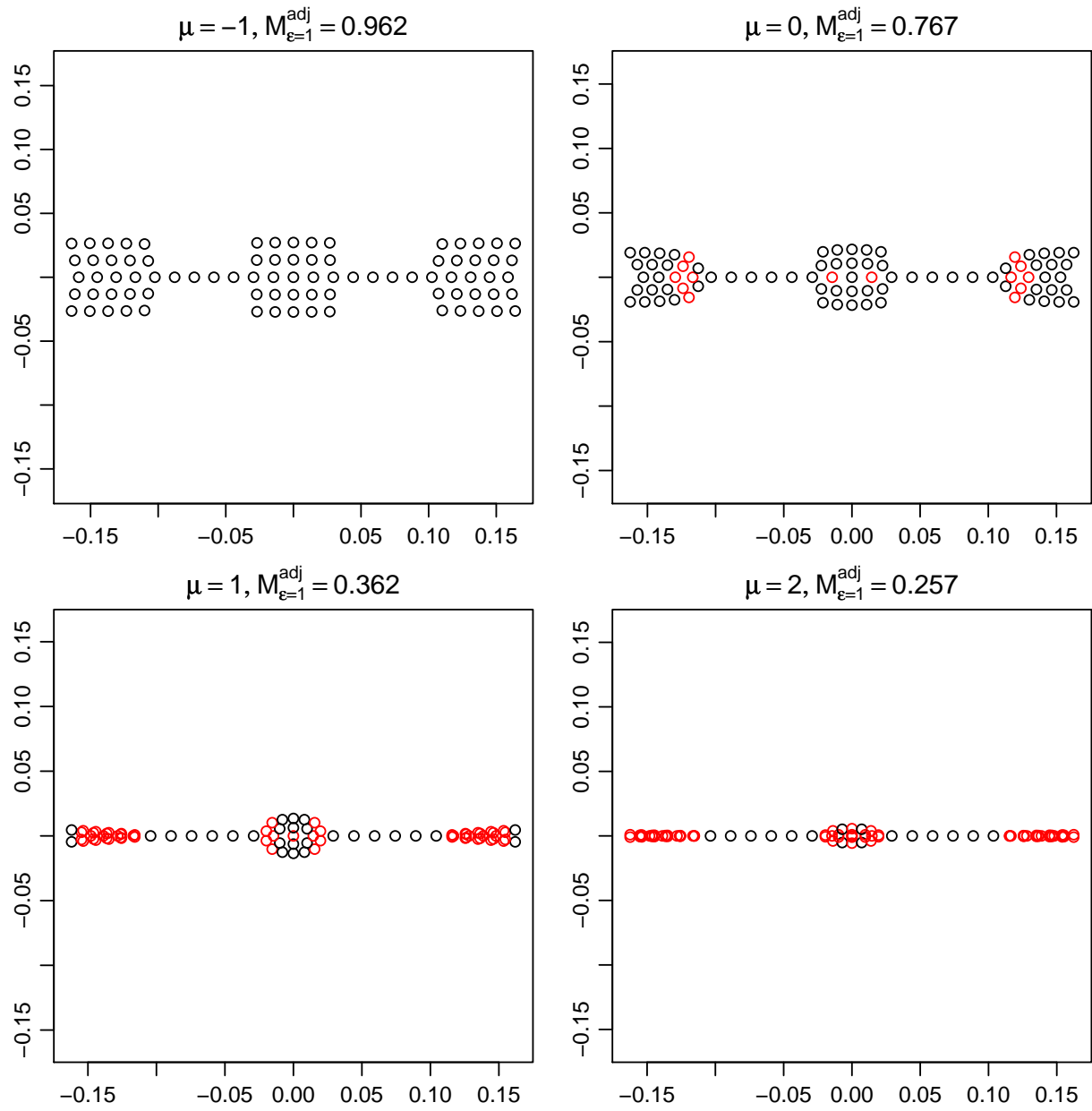


Figure 6: *The configurations with varying μ with other parameters fixed at $\lambda = .2, \tau = 1$, starting from the input configuration (Figure 3).*

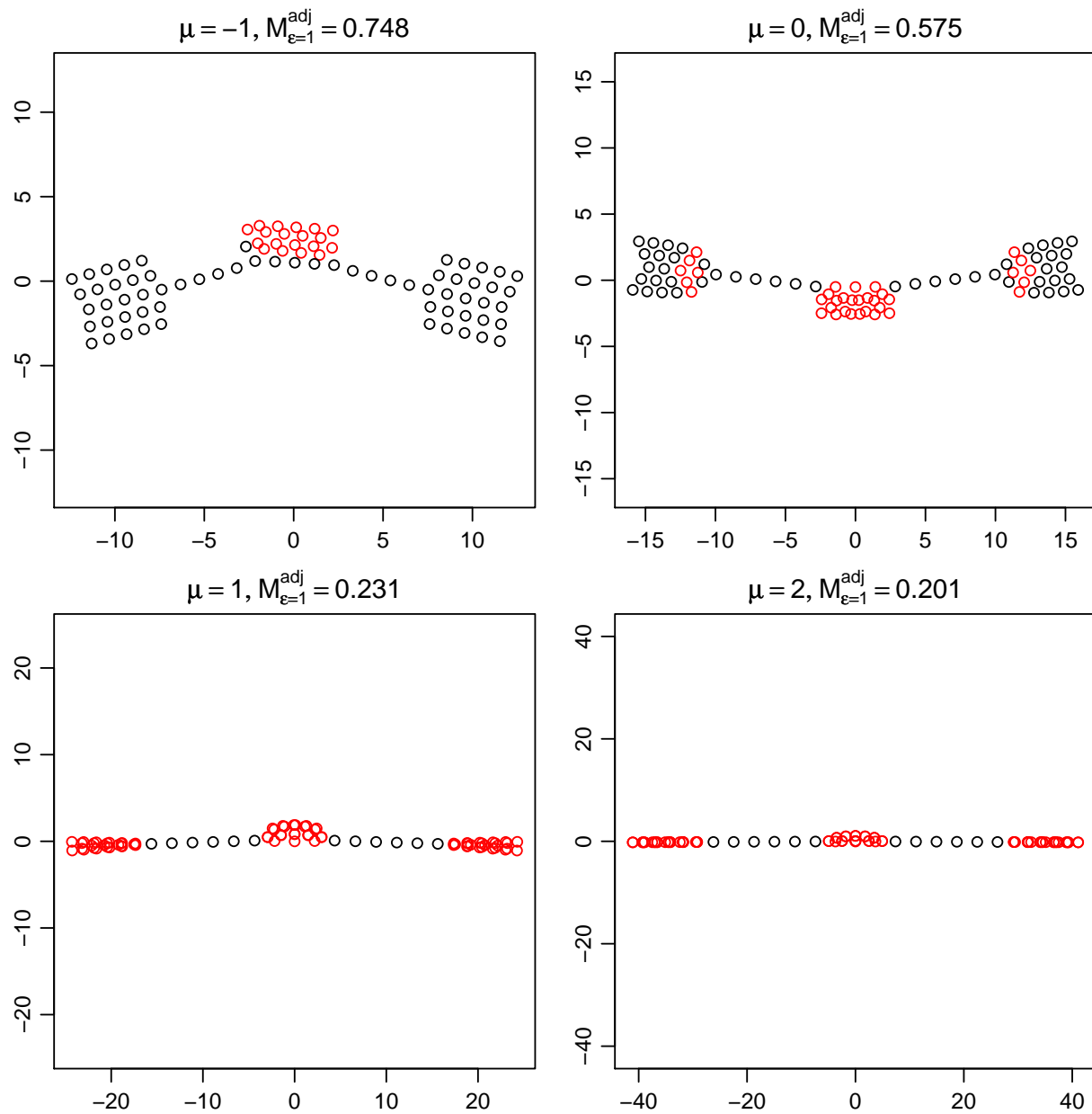


Figure 7: *The configurations with varying μ with other parameters fixed at $\lambda = .2, \tau = 1$, starting from a random configuration.*

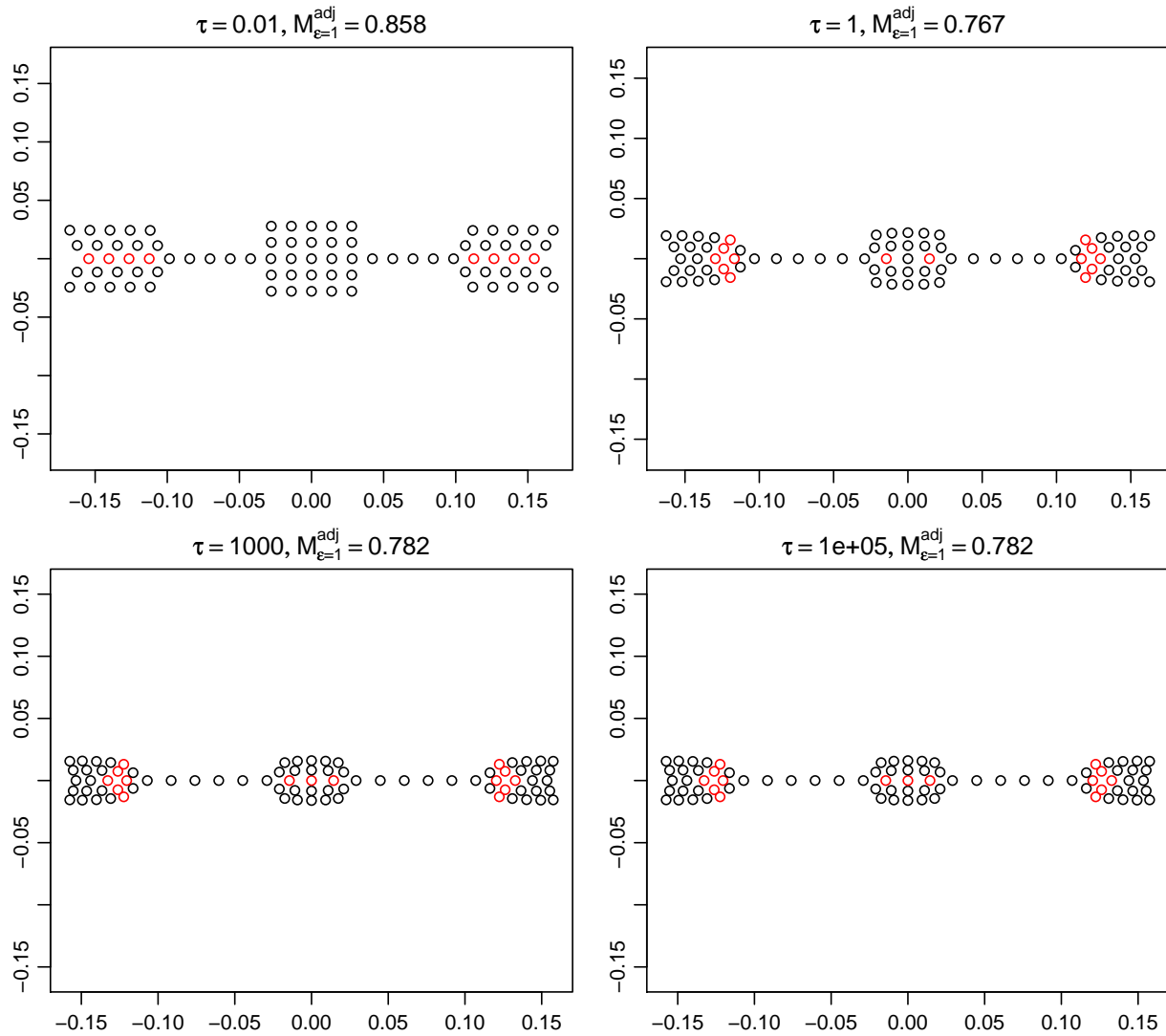


Figure 8: *The configurations with varying τ with other parameters fixed at $\lambda = .2, \mu = 0$, starting from the input configuration (Figure 3).*

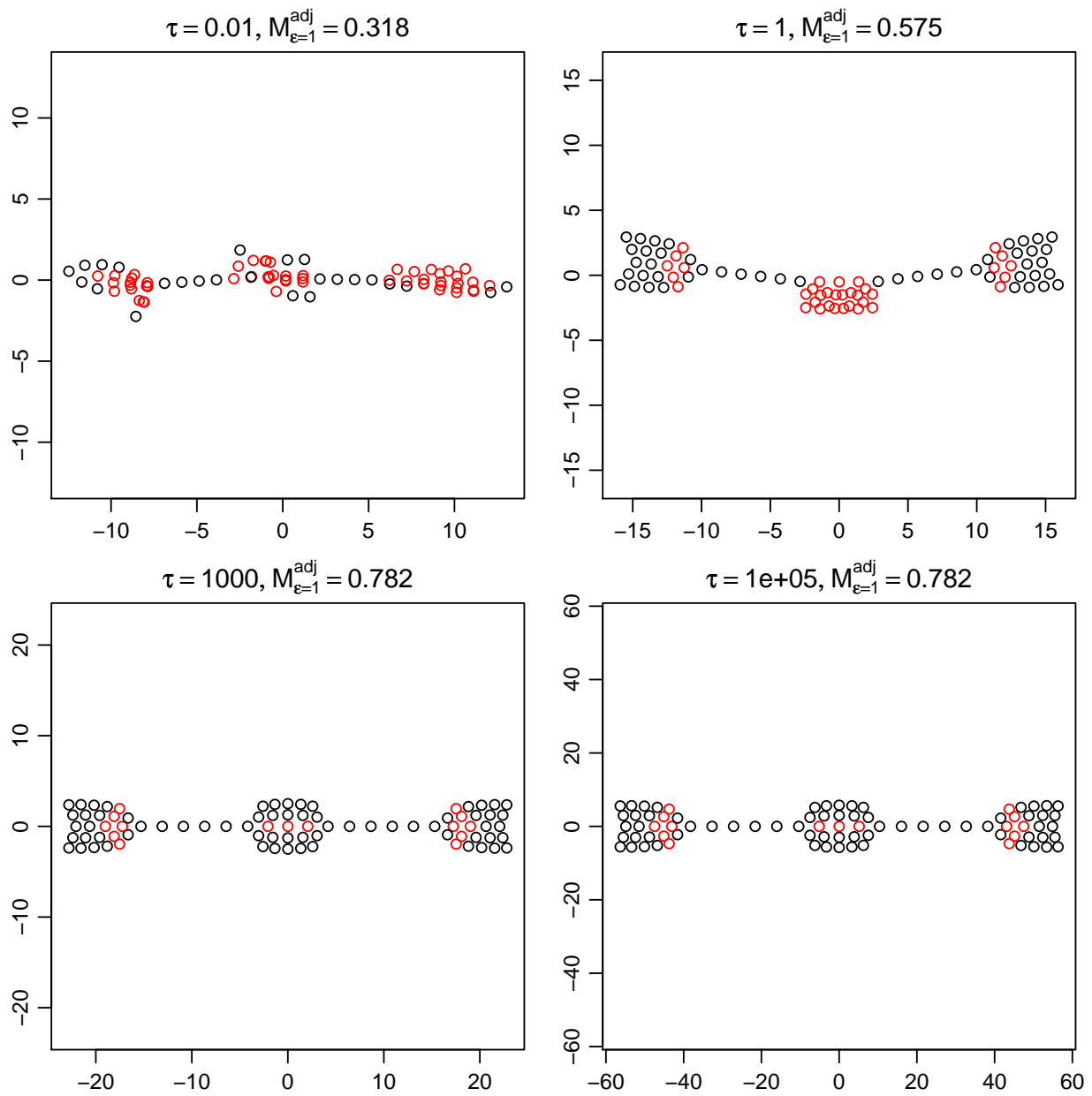


Figure 9: *The configurations with varying τ with other parameters fixed at $\lambda = .2, \mu = 0$, starting from a random configuration.*

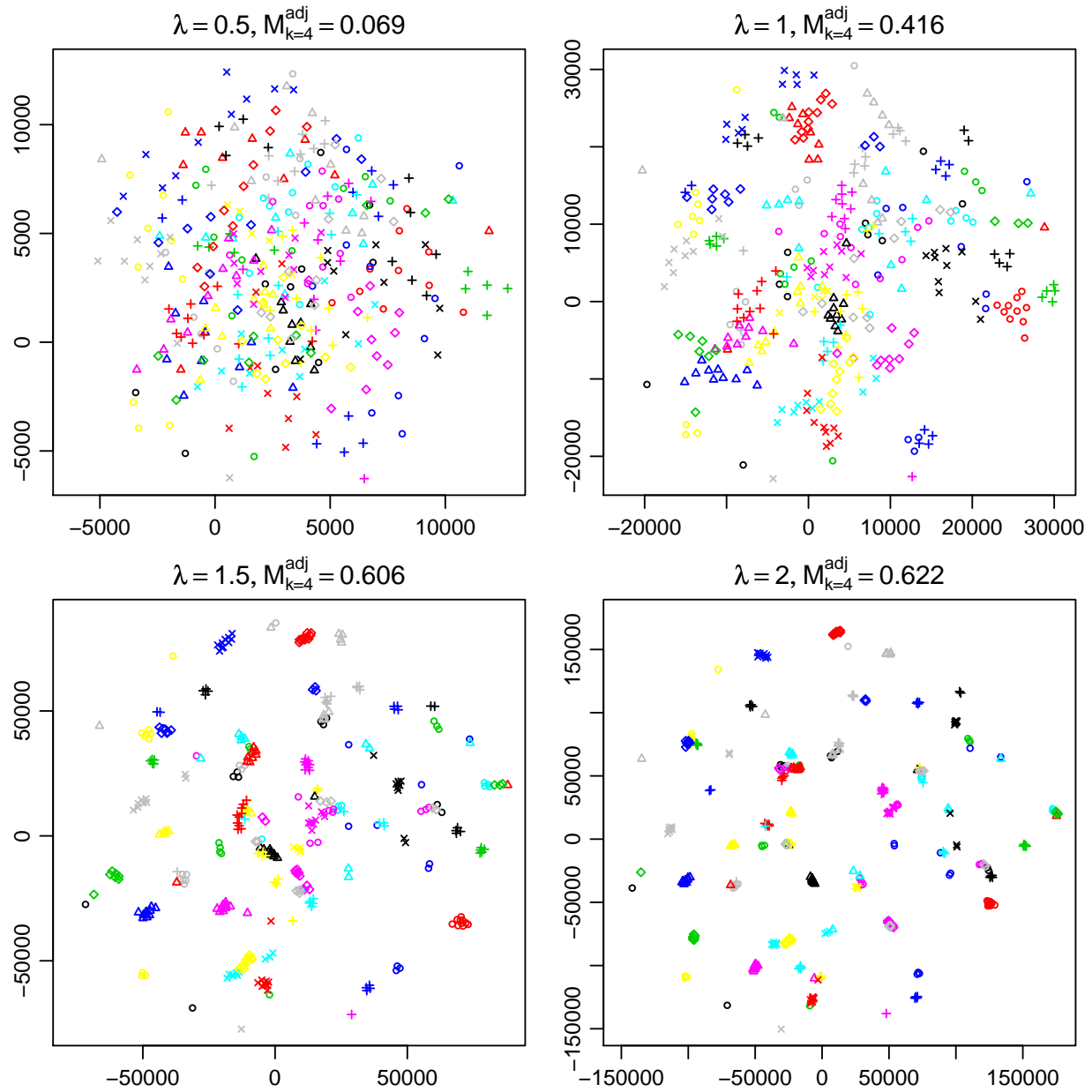


Figure 10: Olivetti Faces: Configurations for four choices from the B-C family of energy functions.

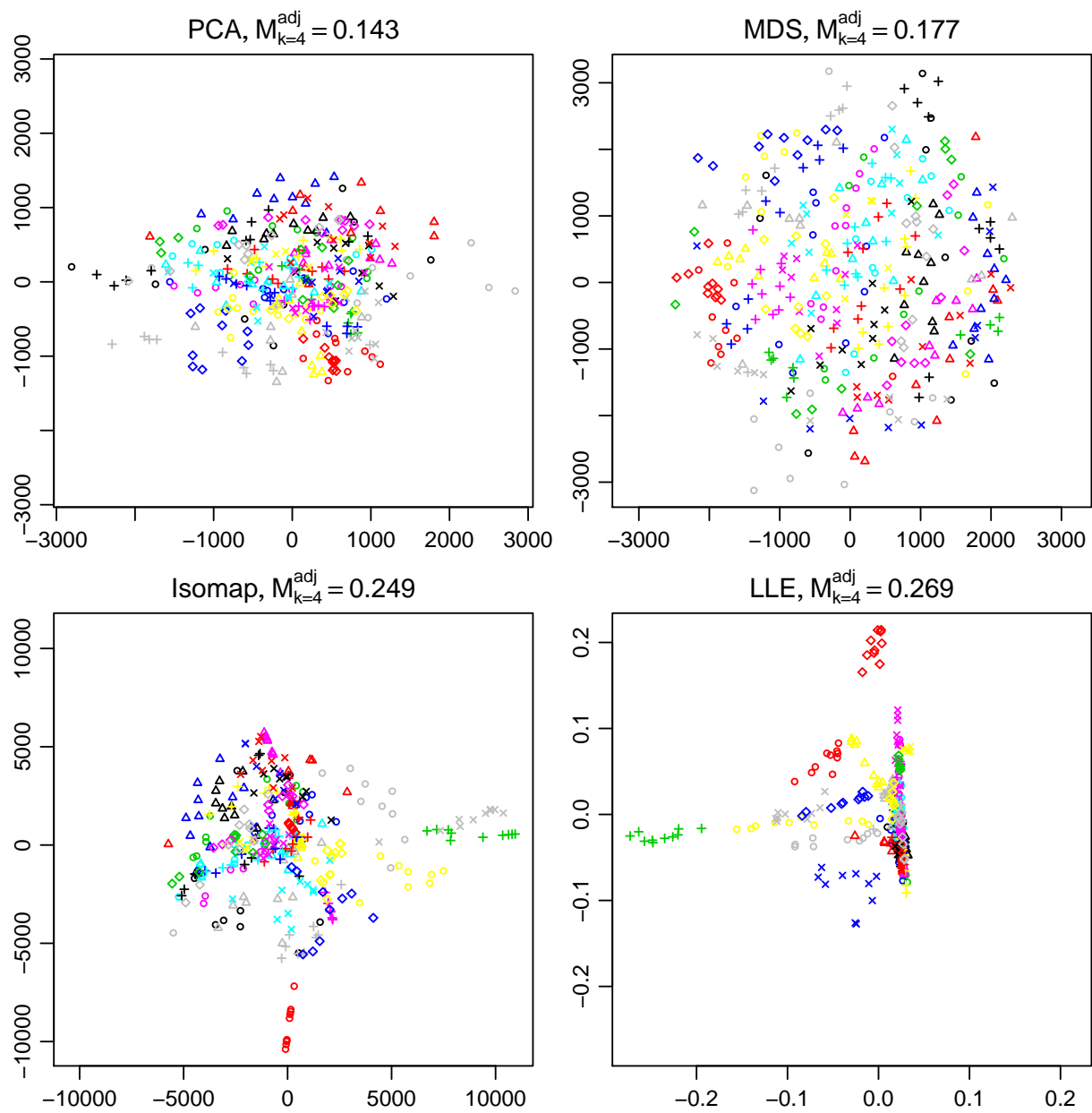


Figure 11: Olivetti Faces: Configurations from PCA, MDS, Isomap and LLE.

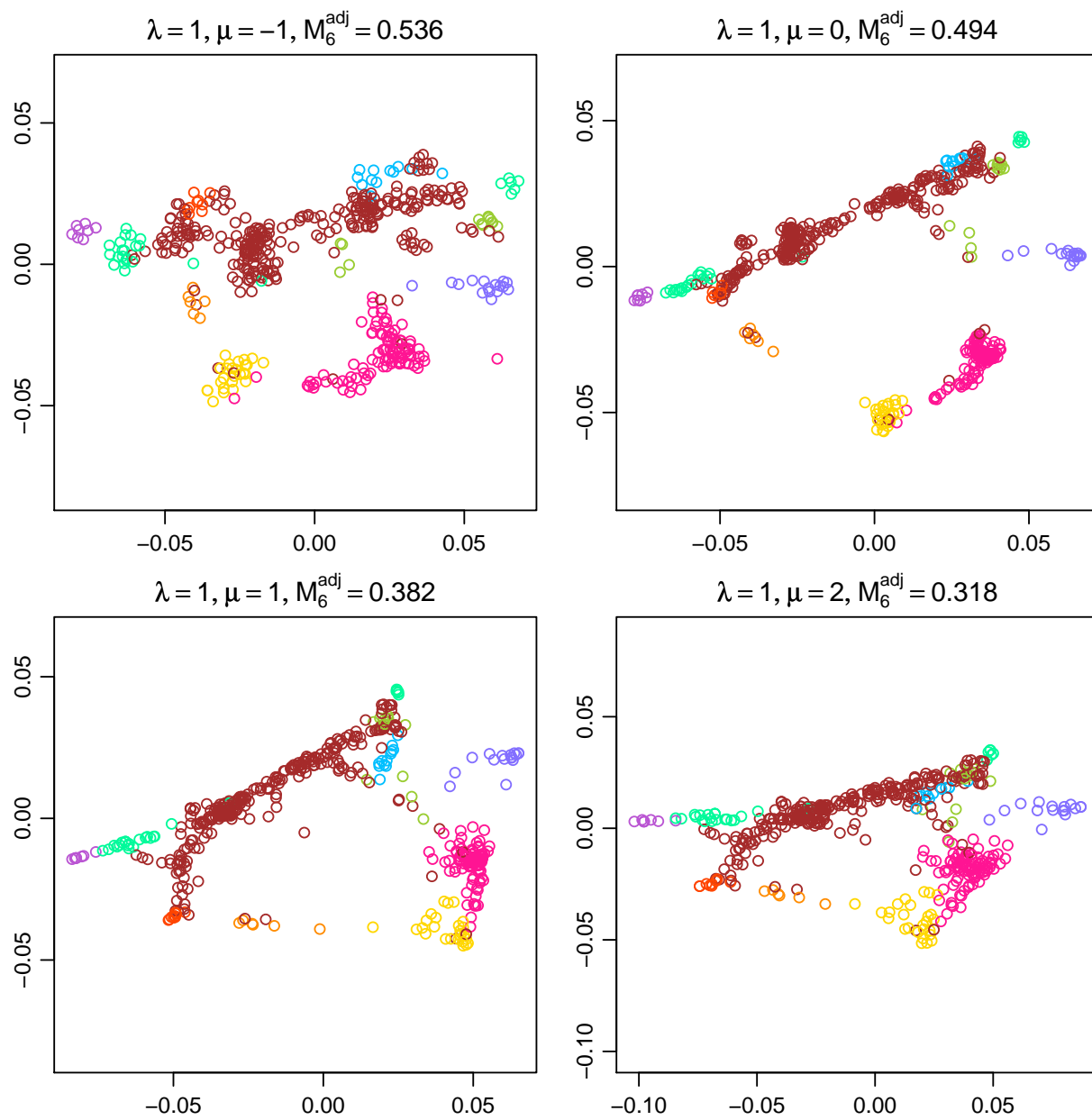


Figure 12: Frey Face Data: Configurations with varying μ when $\lambda = 1$.