# Visualizing Proximity Data

RICH DEJORDY
STEPHEN P. BORGATTI
CHRIS ROUSSIN
DANIEL S. HALGIN
*Boston College*

*In this article, the authors explore the use of graph layout algorithms for visualizing proximity matrices such as those obtained in cultural domain analysis. Traditionally, multidimensional scaling has been used for this purpose. The authors compare the two approaches to identify conditions when each approach is effective. As might be expected, they find that multidimensional scaling shines when the data are of low dimensionality and are compatible with the defining characteristics of Euclidean distances, such as symmetry and triangle inequality constraints. However, when one is working with data that do not fit these criteria, graph layout algorithms do a better job of communicating the structure of the data. In addition, graph layout algorithms lend themselves to interactive use, which can yield a deeper and more accurate understanding of the data.*

***Keywords:*** *multidimensional scaling; visualization; social network analysis; graph layout algorithms; cultural domain analysis; proximity matrices*

Visualization of proximity matrices is commonly used in cultural domain analysis (Weller and Romney 1988; Borgatti 1998). These visualizations facilitate interpretation when we collect item-by-item perceived similarity matrices. Since the 1960s, the most common method of visualizing such data has been multidimensional scaling (MDS) (Torgerson 1958), which represents similarities and differences among a set of items as Euclidean distances in a k-dimensional space (typically two dimensions for easy representation in printed form). However, we propose that the graph layout algorithms (GLAs), which underlie many popular social network analysis tools (e.g., Borgatti 2002; Batagelj and Mrvar 2003), offer an alternative approach to visualizing this type of data and can assist in analysis by producing visualizations that more effectively convey specific characteristics of the data.

*We thank Russ Bernard and two anonymous reviewers whose helpful comments on a previous draft of this article have improved its clarity and focus.*

*Field Methods*, Vol. 19, No. 3, August 2007 239–263
DOI: 10.1177/1525822X07302104
© 2007 Sage Publications

Our intent is not to suggest that GLAs should replace MDS visualizations, and we identify circumstances in which MDS produces superior visualizations. Our intent is also not to compare the methods from a mathematical perspective. Rather, our goals are to (1) introduce an additional tool that can assist in analysis and visualization of this type of data, (2) help identify conditions where each approach excels or falters, and (3) provide a visual comparison of each tool's representation when applied to the same data.

We organize this article as follows. First, we very briefly review both MDS and GLA techniques. Then we show how GLA methods can be applied in cultural domain analysis. Next, we systematically compare the two approaches, highlighting where each approach is more able to represent the underlying structure of the data. We conclude with a summary of our findings.

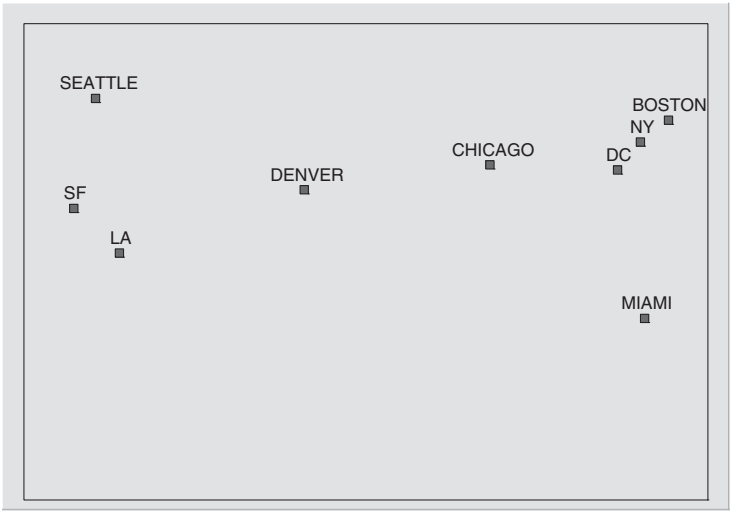
## MDS: MODELING PERCEIVED SIMILARITY AS EUCLIDEAN DISTANCE

The most common technique for visualizing perceived similarities or dissimilarities is MDS (Torgerson 1958; see Kruskal and Wish 1978 for an excellent introduction). MDS creates a graphical representation of a square item-by-item (or 1-mode) proximity matrix. The MDS algorithm determines coordinates for each item in a k-dimensional space such that the Euclidean distances among the points best approximates the input proximities. Input proximities may be either similarities or dissimilarities. In the case of dissimilarities (such as distances between cities), the relationship between input proximities and the Euclidean distances in the MDS map is positive: Larger input values correspond to larger map distances. In the case of similarities (such as correlations or perceived similarity data), the relationship is negative: Larger input values correspond to smaller map distances.

In cultural domain analysis, the input matrix is an aggregate similarity matrix, which represents the proportion of times that a given pair of items was seen as similar by respondents in elicitation tasks such as pile sorts or triads (Borgatti 1994). In other applications, the input matrix may be correlations among variables or distances between objects.

MDS has both metric and nonmetric variations. In metric MDS, coordinates in k-dimensional space are sought such that the Euclidean distance between any pair of items is linearly related (positively or negatively depending on whether the data are dissimilarities or similarities) to the input proximity of the same pair. In nonmetric MDS, the Euclidean distances are only required to match the rank order of the input proximities. In either case, a measure of fit between the Euclidean distances and the input

FIGURE 1
Metric Multidimensional Scaling Plot of U.S. Cities' Data (Stress = .014)




NOTE: SF = San Francisco; LA = Los Angeles; NY = New York; DC = Washington, D.C.

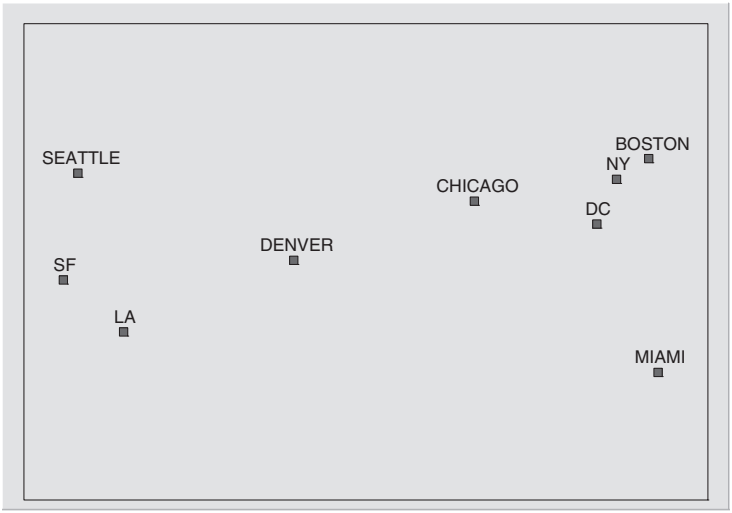
proximities is computed to allow assessment of the adequacy of the MDS representations. Most fit measures, such as the commonly used stress measures of Kruskal (1964), are simple normalizations of the sum of squared differences between the distances on the MDS representation and a function of the input proximities. High stress indicates a poor fit and that the MDS representation distorts the underlying data. Increasing the number of dimensions reduces the distortion; however, it also undermines the goals of both data reduction and useful visualization of the underlying data.

Figure 1 shows a metric MDS representation of the CITIES data set that is included in the UCINET software package (Borgatti, Everett, and Freeman 2002). These data, presented in Table 1, comprise travel distances, in miles, among nine U.S. cities. MDS plots can be rotated around the origin or reflected through either axis to facilitate interpretation without affecting the representation of the data. In fact, Figure 1 was reflected through the horizontal axis to better approximate most maps of the United States. This close visual approximation is consistent with a low stress value of .014, indicating that the very little distortion was introduced when representing the data in a two-dimensional plane. If, however, we had included cities from around the globe, we would have had to choose between a

TABLE 1
U.S. Cities' Travel Distances (miles)

| | *1* *Boston* | *2* *New York* | *3* *Washington, D.C.* | *4* *Miami* | *5* *Chicago* | *6* *Seattle* | *7* *San Franscisco* | *8* *Los Angeles* | *9* *Denver* |
|---|---|---|---|---|---|---|---|---|---|
| Boston | 0 | 206 | 429 | 1,504 | 963 | 2,976 | 3,095 | 2,979 | 1,949 |
| New York | 206 | 0 | 233 | 1,308 | 802 | 2,815 | 2,934 | 2,786 | 1,771 |
| Washington, D.C. | 429 | 233 | 0 | 1,075 | 671 | 2,684 | 2,799 | 2,631 | 1,616 |
| Miami | 1,504 | 1,308 | 1,075 | 0 | 1,329 | 3,273 | 3,053 | 2,687 | 2,037 |
| Chicago | 963 | 802 | 671 | 1,329 | 0 | 2,013 | 2,142 | 2,054 | 996 |
| Seattle | 2,976 | 2,815 | 2,684 | 3,273 | 2,013 | 0 | 808 | 1,131 | 1,307 |
| San Francisco | 3,095 | 2,934 | 2,799 | 3,053 | 2,142 | 808 | 0 | 379 | 1,235 |
| Los Angeles | 2,979 | 2,786 | 2,631 | 2,687 | 2,054 | 1,131 | 379 | 0 | 1,059 |
| Denver | 1,949 | 1,771 | 1,616 | 2,037 | 996 | 1,307 | 1,235 | 1,059 | 0 |

FIGURE 2
Nonmetric Multidimensional Scaling Plot of U.S. Cities (Stress = .000)

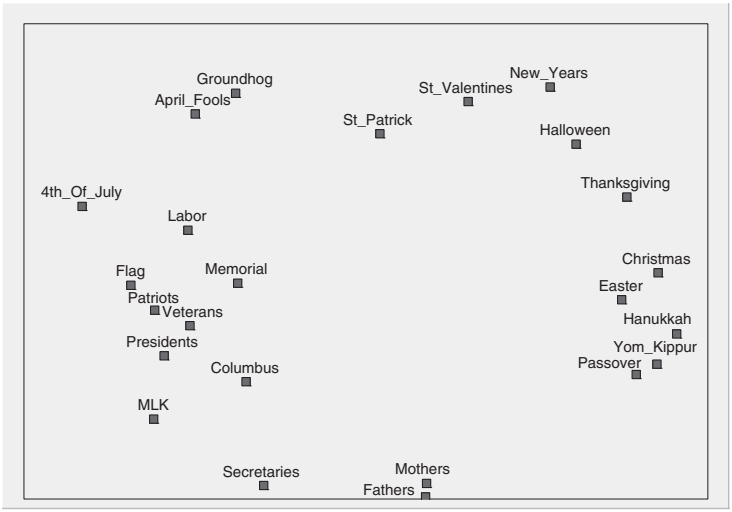



NOTE: SF = San Francisco; LA = Los Angeles; NY = New York; DC = Washington, D.C.

highly distorted representation in two dimensions and an undistorted but difficult to print representation in three dimensions.

Figure 2 shows a nonmetric scaling of the same data. Although the two MDS pictures are generally similar, the relative positioning of the cities in the metric version matches their physical locations slightly better than the nonmetric one. This is because the nonmetric MDS algorithm considers only the rank order of the input proximities (i.e., distances), stripping the data of their interval/ratio properties. In cases where the data are inherently interval or ratio, some information is lost in a nonmetric representation. Even so, in many cases, nonmetric MDS is the more appropriate choice. This is especially true in cultural domain analysis, where the data consist of the proportion of respondents who consider each pair of items similar. Although these data can be seen as ratio level (since they are frequencies), it is not commonly believed that there is a linear relationship between the proportions and the degree of similarity of items. That is, items indicated as similar 50% of the time are not necessarily "twice as similar" as items indicated as similar 25% of the time. However, we do expect the rank ordering is right: The first pair is more similar than the latter. Thus, the nonmetric

FIGURE 3
Metric Multidimensional Scaling Plot of Holiday Data (Stress = .269)



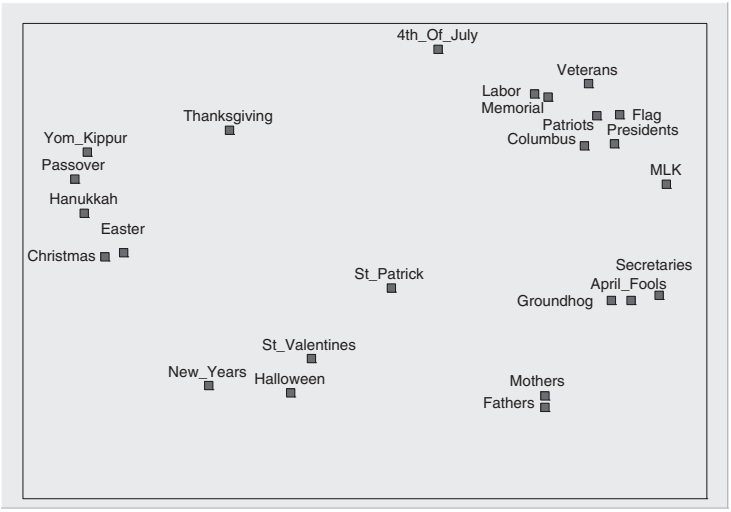

NOTE: MLK = Martin Luther King Day.

MDS technique that relies only on the rank ordering of similarities is typically more appropriate, even though it is "throwing away" some of the richness of the data.

For example, Figures 3 and 4 show metric and nonmetric MDS plots for perceived similarities among twenty-four holidays, respectively, collected by Boston College student Heidi Stokes from undergraduate students as part of a research methods class. The stress for the metric MDS plot is .269, whereas the nonmetric stress is only .171. As the underlying algorithms are different, a direct comparison of the two stress values is not meaningful. However, the nonmetric stress is more clearly within the accepted rules of thumb (Kruskal and Wish 1978). More important, the nonmetric visualization produced is more meaningful, better identifying the clustering of holidays that represents the students' understanding of the domain.

In both metric and nonmetric forms, MDS attempts to minimize the "error" (i.e., distortion) between the n-dimensional solution and the ideal solution through a least-squared mechanism. One consequence of this is that the (mis)placement of distant items has a greater impact on the error calculation than does the placement of near items. Kruskal and Wish (1978)

FIGURE 4
Nonmetric Multidimensional Scaling Plot of Holiday Data (Stress = .171)




NOTE: MLK = Martin Luther King Day.

describe this as better representing the data's "global structure" than their "local structure." As such, whenever there is stress, the interpretation of smaller distances in MDS plots is less reliable than that of larger distances. For example, in Figure 4, which shows the nonmetric MDS representation of the holiday data, it is perfectly reasonable to draw an inference that the Fourth of July is perceived to be very different from the group of religious holidays to the left (Christmas, Easter, Hanukkah, Yom Kippur, and Passover) and distinct from but more similar to the nationalistic or patriotic holidays to the right (Veterans Day, Patriots Day, Columbus Day, Labor Day, Memorial Day, Presidents' Day, and Flag Day). It would be less reasonable to make any inferences about the differences in distance between the Christmas/Easter and the Hanukkah/Passover pairs. In fact, we will see later that, despite appearances, Hanukkah and Passover are perceived to be more similar than Christmas and Easter. Although the general clustering of points into distinct clumps provides useful information, the closer distances are less meaningful, and MDS is not typically useful in identifying the relative ranking of similarities within a group of items positioned closely together (except when stress is 0).

TABLE 2
Game Playing among Wiring Personnel

| | *1* *I1* | *2* *I3* | *3* *W1* | *4* *W2* | *5* *W3* | *6* *W4* | *7* *W5* | *8* *W6* | *9* *W7* | *10* *W8* | *11* *W9* | *12* *S1* | *13* *S2* | *14* *S4* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 I1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 I3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 W1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 4 W2 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 5 W3 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 6 W4 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 7 W5 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 8 W6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| 9 W7 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| 10 W8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |
| 11 W9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| 12 S1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 S2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 S4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |

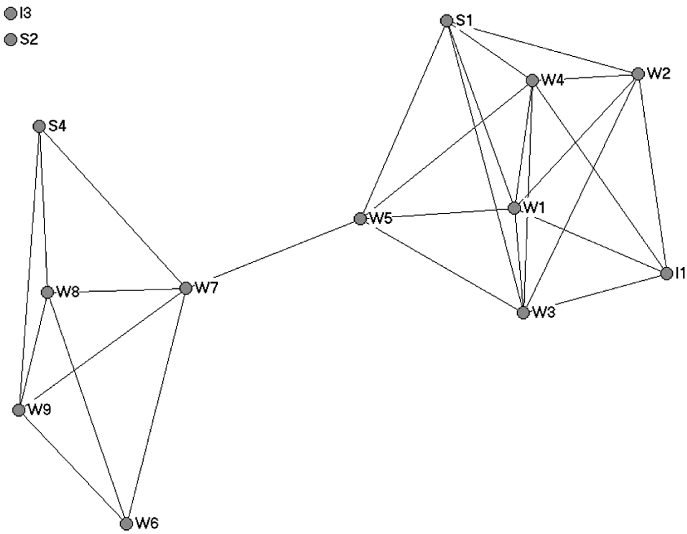
NOTE: I = inspectors; W = workers; S = supervisors. A "1" in any cell indicates that the associated pair of workers was observed playing games together, whereas a "0" indicates they did not.

## GLAs: MODELING PROXIMITIES AS NETWORKS

Outside computer science and electrical engineering, GLAs have mostly been used to represent social network data. With the ever-increasing popularity of social network analysis research (Borgatti and Foster 2003), considerable advances have been made in the application of GLAs to visualize social networks of all sorts: from HIV transmission (Klovdahl 1985) to communication networks (Freeman 1978) and from the bank wiring room interactions in the famous Hawthorne studies (Roethlisberger and Dickson 1939) to attendance at society events (Davis, Gardner, and Gardner 1941). Virtually all social network visualization tools use a GLA to depict the network graphically.

As an example, consider the well-known wiring room data set. These data are presented in Table 2. In the matrix, a "1" in any cell indicates that the associated pair of workers was observed playing games together, whereas a "0" indicates they did not. The data are depicted graphically in Figure 5 using NetDraw (Borgatti 2002), although similar results would have been obtained with other tools (e.g., Batagelj and Mrvar 2003). It is striking how clearly the visual representation conveys the underlying structure of the relationships. In particular, the graph makes it immediately apparent that there are two main groups of persons and that W5 and W7 represent the only connection between them.

FIGURE 5
Graph Layout Algorithm Representation of Game-Playing Data




NOTE: I = inspectors; W = workers; S = supervisors.

Most GLAs are based on drawing an analogy between networks and physical systems. One of the oldest and best-known GLAs is the spring-embedding algorithm of Eades (1984). As its name suggests, the algorithm works by modeling a network of social ties as a system of springs stretched between posts. If a pair of posts with a spring between them is placed too closely together, the spring is compressed and tries to push the posts apart (a property called node repulsion). If the posts are too far apart, the spring is stretched and tries to pull the posts together (a property called node attraction). The algorithm is essentially a method of locating the posts in such a way as to balance the repulsive and attractive forces throughout the entire system. A number of variations on this basic idea have been proposed that seek to improve the ability to find the equilibrium point of the system, including the well-known FR algorithm of Fruchterman and Reingold (1991).

The system of springs can also be seen in terms of minimizing potential energy. This is the approach taken by Kamada and Kawai (1991). A key difference between their algorithm and that of Eades (1984) and Fruchterman and Reingold (1991) is that Kamada and Kawai propose that the physical

distance between points (in the GLA representation) should be proportional to the geodesic distance among the corresponding nodes in the network. Geodesic distance, known as *degrees of separation* in the popular press, refers to the number of links in the shortest path between a pair of nodes. Thus, the Kamada-Kawai algorithm is essentially an MDS of the associated geodesic distance matrix. The figures presented in this article were drawn using a variation of the Kamada-Kawai algorithm.

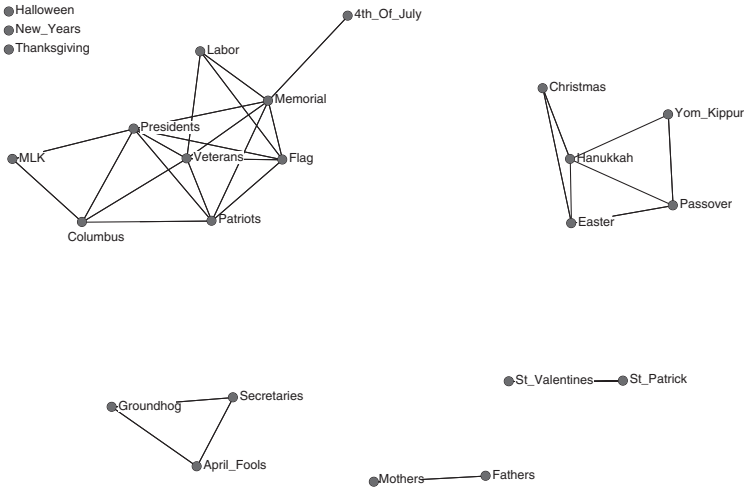
Finally, a third well-known approach to graph layout involves direct optimization of desirable layout qualities. This is the approach taken by Blythe, McGrath, and Krackhardt (1994) as well as Davidson and Harel (1996). Here, simulated annealing is used to maximize a complex function that contains a term for each desirable layout quality. For example, Davidson and Harel propose five principles of good layouts: (1) All nodes are visible at one time, (2) available space is utilized as fully as possible, (3) line lengths are of approximately uniform length, (4) line crossings are minimized, and (5) nodes maintain a margin of separation from nearby lines. Each of these qualities is quantified, and a generic optimization algorithm seeks to maximize all five properties simultaneously.

## APPLYING GLAs TO VISUALIZE PERCEIVED SIMILARITY DATA

GLAs are designed to work with binary data representing the presence or absence of relationships. To use them with valued proximity data, such as the proportions obtained in pile sort tasks, we must dichotomize the proximities. In effect, we must decide how similar two items must be in order to be linked by a line in the visualization. In practice, we are typically interested in dichotomizing at various different levels in order to get a complete understanding of the structure of the data. GLA-based software tools such as NetDraw (Borgatti 2002) allow you to work interactively with continuous data by specifying filtering criteria for when lines are drawn between nodes on the graph.

Figure 6 is the GLA representation of the holiday data from Figure 4. A line is shown between those holidays deemed similar by at least 50% of the respondents. The groupings visible in Figure 4 and Figure 6 are quite similar. For example, we see a group of patriotic holidays and a group of religious holidays in both. Figure 7 shows four GLA representations of the same holiday data, but with lines present based on different proportions of the respondents indicating the holidays were similar. In particular, panel d has lines only when at least 75% of the respondents indicated the holidays were similar.

FIGURE 6
Graph Layout Algorithm Representation of Holiday Data (filtered at .50)

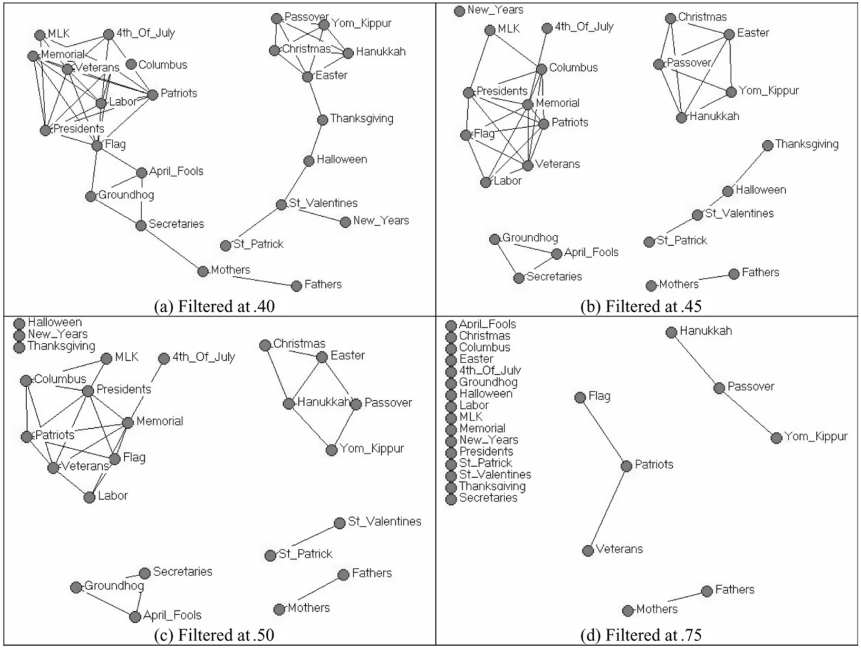



NOTE: MLK = Martin Luther King Day.

Naturally, there are fewer lines on this graph because fewer relationships meet this criterion. However, as mentioned earlier, the presence of a line between Hanukkah and Passover at this level, and the lack of a line between Christmas and Easter, highlights the stronger perceived similarity between Hanukkah and Passover compared to Christmas and Easter within the population sampled, in this case undergraduates at a Catholic university.

In our view, the addition of lines to represent relationships (filtered at a certain level) and the use of an optimization algorithm to locate nodes so as to maximize readability generates a display that is highly aesthetic and exceptionally easy to grasp. It should be noted that this kind of plot is different from the practice described by Kruskal and Wish (1978) of adding lines to an otherwise standard MDS plot. In the latter approach, the positions of the items in space are based on the raw proximity matrix, whereas in the GLA approach, the data matrix is dichotomized at a given cutoff level to create a network, and then path distances are computed and used as the basis for positioning points. Thus, the position of points changes as one chooses different cutoff levels. In addition, unlike MDS, the GLA algorithm takes into account aesthetic criteria such as avoiding placing points too close to

FIGURE 7
Graph Layout Algorithm Representation of Holiday Data with Various Filterings




NOTE: MLK = Martin Luther King Day.

each other. Thus, a GLA representation of proximity matrices produces a powerfully comprehensible and aesthetic representation, but one that is more abstract than the corresponding MDS picture.

## COMPARISON OF METHODS

In this section, we systematically compare the MDS and GLA visualization techniques. In particular, we examine how the methods deal with four data issues (high dimensionality, outliers, violations of triangle inequality, and asymmetry) and three tool properties (interactivity, precision in visualizations, and node repositioning). We discuss each of these concepts below, providing examples to show how each visualization technique handles such variance.

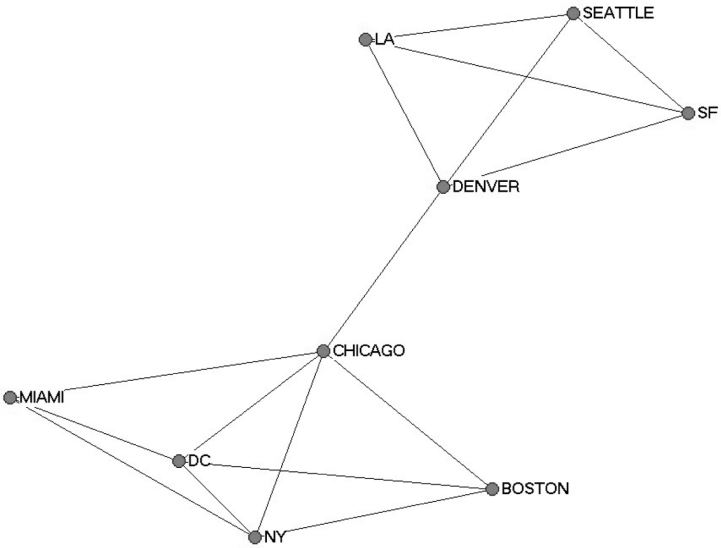
### Dimensionality

In this context, dimensionality refers to the number of dimensions (or map axes) needed to accurately represent the data. For example, the CITIES data used in Figure 1 are well represented by two dimensions. This is not surprising because they consist of distances among locations on a roughly flat surface. Of course, there are altitude differences, and there is curvature of the Earth's surface, but for the set of U.S. cities, these factors are minor compared to the variation along the dimensions of latitude and longitude. For data that require more than two dimensions to adequately represent their variability (i.e., data that have higher dimensionality), MDS visualizations become much more difficult to interpret as well as impracticable to represent in print.

Figure 8 shows a GLA representation of the same data with lines drawn between cities that are within 1,500 miles of each other. The picture tells the essential story—a West Coast group of cities and an East Coast group of cities bridged from the west by Denver and bridged from the east by Chicago—very clearly. However, the MDS visualization in Figure 1 is richer in the sense that it provides a far more nuanced representation than the GLA. For example, one can tell that San Francisco and Los Angeles are much closer to each other than they are to Seattle: The MDS representation retains degrees of proximity whereas the GLA has a more cartoonish or schematic feel.

Of course, close interpretation of an MDS picture is only possible when stress is low. When stress is high, perhaps because of inherently high-dimensional data, MDS plots cannot be interpreted so closely, and the GLA representation can be interpreted more reliably. For example, consider again the nonmetric MDS plot in Figure 4, which plots the relative similarity of holidays according to undergraduate students. The stress value (.171) is low enough that most researchers would not feel compelled to go to a difficult to

FIGURE 8
Graph Layout Algorithm Representation of U.S. Cities (filtered at 1,500 miles)




NOTE: SF = San Francisco; LA = Los Angeles; NY = New York; DC = Washington, D.C.

publish three-dimensional plot. Yet the stress is high enough to indicate the plot is not a completely faithful representation of the data, potentially misleading the viewer and introducing potentially significant ambiguity into the interpretation.

Consider the location of Martin Luther King Day (MLK) on the right side of the plot. Its positioning between the cluster of patriotic holidays and the cluster containing Secretary's Day, Groundhog Day, and April Fool's Day suggests not only separation from those patriotic holidays but also some association with the other group, as if some respondents saw MLK as patriotic and others saw it as more whimsical. One might be tempted to conjecture that the MDS plot is cleverly uncovering an implicit attitude (Greenwald, McGhee, and Schwartz 1998) about race among some informants. Or one might hypothesize that respondents are placing MLK near the Secretary's Day cluster because those are fairly recently adopted holidays and have less lore, ritual, and ceremony than most of the other holidays.

However, a GLA representation of the same data suggests that neither interpretation is supported by the data, because MLK is not connected to the new or whimsical holidays at all. The MDS plot, in this case, is misleading,

and MLK belongs firmly in the cluster of patriotic holidays. Returning to Figure 6, we see that the GLA filtered at .50 (i.e., a line is drawn between two holidays if at least 50% of respondents saw them as similar) shows a clear separation between the patriotic holidays (including MLK) and four other clusters of holidays, with no line linking MLK to the whimsical holidays. Lowering the cutoff to 40% (see Figure 7a) shows that MLK is still not connected to those holidays but that Flag Day does show a tie, which was also not obvious from the MDS representation. Thus, the apparent intermediacy of MLK in the MDS plot (see Figure 4) was actually spurious—a distortion consistent with moderately high stress.

Extending the MDS to three dimensions reduces the stress to .109 and produces a more accurate picture (not shown). Visualizing the three-dimensional MDS using the MAGE software tool (Richardson and Richardson 1992), which allows the user to view a three-dimensional picture on a computer screen, we were able to see that MLK was positioned firmly with the patriotic holidays and showed no tendency to drift toward the whimsical cluster.

Although tools like MAGE (Richardson and Richardson 1992) are available for visualizing and interpreting three-dimensional data interactively, they do not always translate well to static, two-dimensional media, such as journals. Furthermore, beyond three dimensions it becomes impossible to visualize the data at all. In these cases, GLA representations are often more "economical," representing high-dimensional information nicely in two dimensions. Even if multiple diagrams using different cutoff levels are needed, this is still easier than schemes for printing three-dimensional data, such as adding perspective cues or printing scatterplots of each dimension against every other.

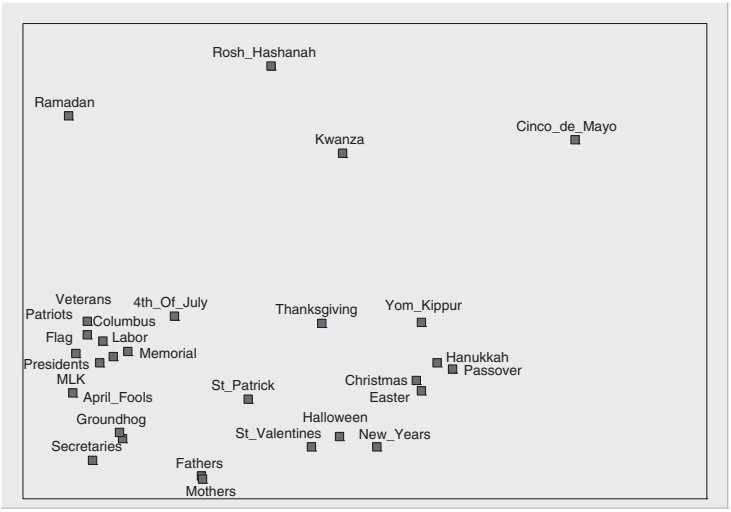
In addition, the GLA diagram has a certain clarity in its relationship to the data: A line is drawn between two items only if their similarity is greater than a user-specified level. Therefore, no matter where the points are placed, it is always clear whether or not any given pair is closer than some cutoff. On the other hand, information regarding how much closer is not represented.

In summary, with respect to dimensionality, our view is that if the data can be represented well in two dimensions, an MDS plot in dimensions with low stress provides a rich visualization of the data that encodes all of the subtlety of the raw data in the pattern of distances. As the underlying dimensionality of the data increases, however, GLAs, which present a kind of stylized or simplified view of the data, provide a way of representing the essential features of complex data on a two-dimensional page without distortion.

### Outliers

Because MDS attempts to represent all dyadic relationships present in the data at once, any items that have unusual relationships to the rest of the

FIGURE 9
Multidimensional Scaling Plot of Holiday Data
with Outliers (Stress = .207)

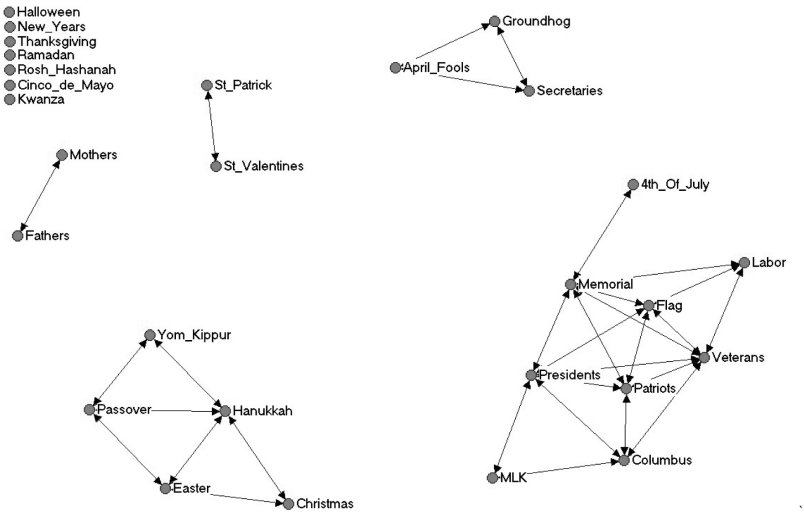



NOTE: MLK = Martin Luther King Day.

data affect the overall picture. In the holiday data we have been presenting, there were actually four additional holidays included in the pile sort task: Ramadan, Rosh Hashanah, Kwanza, and Cinco de Mayo. However, none of these holidays was ever grouped with other holidays or with each other. They were indicated as either unique or unknown by the respondent. The nonmetric MDS representation of these data, when these holidays are included, is presented in Figure 9. As you can see, the four outliers are scattered across the top, whereas the bottom approximates the same structure as the representation in Figure 4, but compressed to use less of the plot area. Because there are four holidays that are more dissimilar to the rest of the data (and each other) than most of the other relationships, almost half of the MDS plot is devoted to setting them apart from the other data. The stress of this MDS representation is also noticeably higher at .207 instead of .171.

Compare the difference between Figure 4 and Figure 9 with the difference between Figure 6 and Figure 10. Here, because relationships are filtered at a certain level, holidays without any relationship at that strength are simply set aside along the left side of the graph. In graph terms, these are termed "isolates" and do not affect the positioning of the other nodes in the

FIGURE 10
Graph Layout Algorithm Representation of Holiday Data with Outliers (filtered at .50)




NOTE: MLK = Martin Luther King Day.

graph. Thus, the representation of relationships in the data is automatically unaffected by the outliers in the data in a GLA, whereas we had to remove the items from the similarity matrix input to MDS in order to not have the outliers distort the overall representation of the other relationships.

## Transitivity

MDS uses distance in Euclidean space to represent relationships. Euclidean distances have certain properties. Data that do not conform to these properties cannot be accurately represented in a Euclidean space. One of these properties is triangle inequality (also referred to as transitivity), which states that $d(i,j) \leq d(i,k) + d(k,j)$. This means that if an object *k* is close to both *i* and *j,* then in a Euclidean (e.g., physical) space, there is a limit to how far away *i* and *j* can be from each other. Data in which the relationships among the items are not so constrained simply cannot be represented in a Euclidean space without distortion.

In Figure 4 (the nonmetric MDS plot of the holiday data), Thanksgiving is somewhat isolated toward the top of the graph just left of center. A look at the raw data in Table 3 indicates that Thanksgiving was perceived as most

# TABLE 3
## Input Similarity Matrix for Holiday Data

| | *1* *April Fool's Day* | *2* *Christmas* | *3* *Columbus Day* | *4* *Easter* | *5* *Father's Day* | *6* *Flag Day* | *7* *Fourth of July* | *8* *Groundhog Day* | *9* *Halloween* | *10* *Hanukkah* | *11* *Labour Day* | *12* *MLK* | *13* *Memorial Day* | *14* *Mother's Day* | *15* *New Year* | *16* *Passover* | *17* *President's Day* | *18* *St. Patrick's Day* | *19* *St. Valentine's Day* | *20* *Thanksgiving* | *21* *Veteran's Day* | *22* *Yom Kippur* | *23* *Patriot's Day* | *24* *Secretary's Day* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| April Fool's Day | .00 | .00 | .19 | .15 | .22 | .41 | .11 | .67 | .26 | .00 | .26 | .22 | .30 | .30 | .11 | .11 | .19 | .33 | .19 | .00 | .19 | .04 | .33 | .56 |
| Christmas | .00 | .00 | .00 | .74 | .11 | .04 | .11 | .11 | .30 | .70 | .04 | .11 | .04 | .11 | .37 | .48 | .11 | .19 | .33 | .37 | .00 | .44 | .00 | .04 |
| Columbus Day | .19 | .00 | .00 | .00 | .22 | .44 | .30 | .19 | .15 | .04 | .41 | .59 | .48 | .22 | .04 | .04 | .63 | .19 | .07 | .26 | .70 | .19 | .56 | .26 |
| Easter | .15 | .74 | .00 | .00 | .15 | .04 | .15 | .00 | .26 | .52 | .04 | .00 | .15 | .22 | .26 | .63 | .00 | .22 | .22 | .41 | .00 | .48 | .07 | .15 |
| Father's Day | .22 | .11 | .22 | .15 | .00 | .15 | .19 | .19 | .19 | .04 | .19 | .22 | .15 | .93 | .11 | .07 | .19 | .15 | .19 | .11 | .15 | .00 | .15 | .33 |
| Flag Day | .41 | .04 | .44 | .04 | .15 | .00 | .37 | .41 | .07 | .04 | .52 | .33 | .56 | .15 | .04 | .04 | .59 | .19 | .07 | .07 | .67 | .04 | .81 | .33 |
| Fourth of July | .11 | .11 | .30 | .15 | .19 | .37 | .00 | .04 | .15 | .07 | .44 | .22 | .56 | .11 | .15 | .07 | .37 | .26 | .19 | .26 | .41 | .07 | .41 | .04 |
| Groundhog Day | .67 | .11 | .19 | .00 | .19 | .41 | .04 | .00 | .19 | .11 | .26 | .33 | .19 | .19 | .22 | .00 | .33 | .26 | .30 | .00 | .19 | .04 | .26 | .52 |
| Halloween | .26 | .30 | .15 | .26 | .19 | .07 | .15 | .19 | .00 | .15 | .07 | .04 | .00 | .19 | .37 | .11 | .04 | .26 | .48 | .48 | .15 | .19 | .07 | .15 |
| Hanukkah | .00 | .70 | .04 | .52 | .04 | .04 | .07 | .11 | .15 | .00 | .04 | .11 | .04 | .04 | .22 | .78 | .11 | .19 | .26 | .22 | .00 | .74 | .00 | .04 |
| Labor Day | .26 | .04 | .41 | .04 | .19 | .52 | .44 | .26 | .07 | .04 | .00 | .26 | .67 | .11 | .11 | .04 | .41 | .26 | .11 | .19 | .52 | .11 | .48 | .30 |
| MLK | .22 | .11 | .59 | .00 | .22 | .33 | .22 | .33 | .04 | .11 | .26 | .00 | .37 | .22 | .15 | .00 | .70 | .15 | .22 | .04 | .44 | .04 | .37 | .30 |
| Memorial Day | .30 | .04 | .48 | .15 | .15 | .56 | .56 | .19 | .00 | .04 | .67 | .37 | .00 | .22 | .07 | .15 | .56 | .30 | .07 | .22 | .70 | .07 | .67 | .30 |
| Mother's Day | .30 | .11 | .22 | .22 | .93 | .15 | .11 | .19 | .19 | .04 | .11 | .22 | .22 | .00 | .11 | .15 | .19 | .19 | .19 | .11 | .15 | .00 | .19 | .41 |
| New Year | .11 | .37 | .04 | .26 | .11 | .04 | .15 | .22 | .37 | .22 | .11 | .15 | .07 | .11 | .00 | .11 | .11 | .33 | .44 | .30 | .00 | .11 | .00 | .07 |
| Passover | .11 | .48 | .04 | .63 | .07 | .04 | .07 | .00 | .11 | .78 | .04 | .00 | .15 | .15 | .11 | .00 | .00 | .19 | .11 | .22 | .00 | .78 | .07 | .15 |
| President's Day | .19 | .11 | .63 | .00 | .19 | .59 | .37 | .33 | .04 | .11 | .41 | .70 | .56 | .19 | .11 | .00 | .00 | .22 | .22 | .11 | .70 | .04 | .63 | .26 |
| St. Patrick's Day | .33 | .19 | .19 | .22 | .15 | .19 | .26 | .26 | .26 | .19 | .26 | .15 | .30 | .19 | .33 | .19 | .22 | .00 | .52 | .07 | .19 | .15 | .19 | .26 |
| St. Valentine's Day | .19 | .33 | .07 | .22 | .19 | .07 | .19 | .30 | .48 | .26 | .11 | .22 | .07 | .19 | .44 | .11 | .22 | .52 | .00 | .22 | .04 | .07 | .00 | .11 |
| Thanksgiving | .00 | .37 | .26 | .41 | .11 | .07 | .26 | .00 | .48 | .22 | .19 | .04 | .22 | .11 | .30 | .22 | .11 | .07 | .22 | .00 | .26 | .26 | .19 | .04 |
| Veteran's Day | .19 | .00 | .70 | .00 | .15 | .67 | .41 | .19 | .15 | .00 | .52 | .44 | .70 | .15 | .00 | .00 | .70 | .19 | .04 | .26 | .00 | .15 | .81 | .22 |
| Yom Kippur | .04 | .44 | .19 | .48 | .00 | .04 | .07 | .04 | .19 | .74 | .11 | .04 | .07 | .00 | .11 | .78 | .04 | .15 | .07 | .26 | .15 | .00 | .07 | .00 |
| Patriot's Day | .33 | .00 | .56 | .07 | .15 | .81 | .41 | .26 | .07 | .00 | .48 | .37 | .67 | .19 | .00 | .07 | .63 | .19 | .00 | .19 | .81 | .07 | .00 | .37 |
| Secretary's Day | .56 | .04 | .26 | .15 | .33 | .33 | .04 | .52 | .15 | .04 | .30 | .30 | .30 | .41 | .07 | .15 | .26 | .26 | .11 | .04 | .22 | .00 | .37 | .00 |

NOTE: MLK = Martin Luther King Day.

similar to Halloween, which is on the bottom side of the plot. Although we would typically expect Thanksgiving to be positioned close to Halloween, looking more closely at the data in Table 3, we see that Thanksgiving was also perceived as dissimilar from St. Valentine's Day, but that Halloween and St. Valentines Day were perceived as similar. Consequently, Thanksgiving was positioned far away from St. Valentine's Day, and therefore from Halloween as well. Because MDS uses Euclidean distances, which are transitive, the relationship between Thanksgiving and Halloween is constrained by each of their relationships with St. Valentine's Day. Thanksgiving cannot be simultaneously close to Halloween and far from St. Valentine's Day because those two are close to each other.

GLAs are not troubled by violations of the triangle inequality law. To be more specific, force-directed methods such as Fruchterman-Reingold are not based on Euclidean spaces and so are not bothered by intransitivity. Other methods, like Kamada-Kawai, explicitly construct their own geodesic distance metric from the dichotomized data, guaranteeing that the transformed data, based on "degrees of separation" or "shortest path length" at any given level of similarity, conform to distance properties. As a result, GLA representations do a nice job with intransitive data. Figure 7a, for example, easily conveys the intransitive relationships between these three items because the edges are drawn to show that Halloween has a relationship with both Thanksgiving and St. Valentine's Day, but the lack of a line between Thanksgiving and St. Valentine's Day indicates the lack of a similarity between the latter pair.

The problems of intransitivity are often exacerbated when the number of items being analyzed or compared increases. The greater the number of items, the more opportunities there are for intransitive triads in the data. However, some data frequently visualized with MDS or GLAs are inherently transitive. For example, correlations among a set of variables never violate the triangle inequality property, independent of the number of variables correlated. When data are not intransitive, both MDS and GLAs can represent the data visually without distortion, but when they are intransitive, GLAs will provide a more accurate representation.

### Symmetry

Another property of Euclidean distance is symmetry, which states that the distance from *a* to *b* is the same as the distance from *b* to *a.* For many forms of cultural domain analysis, this does not present a problem, as the data are intrinsically symmetric. In fact, all of the data sets presented in this article are symmetric. However, there are proximity matrices that do exhibit asymmetry. One example is the set of asymmetric measures of variable association, such as lambda ($\lambda$) or ordinary regression coefficients. Another example is the set

of directional semantic relations among items in a domain, such as "cause of," "precedes," or "may substitute for."

As noted by Kruskal and Wish (1978), MDS is not well suited to asymmetric data because in the Euclidean space it constructs the distance d(x,y) equals the distance d(y,x), and so the only way to get low stress is if both the proximities p(x,y) and p(y,x) are the same. When the data are asymmetric, the best MDS can do is spread out the error and locate x and y so that the distance between them is a compromise of the two input proximities.

In contrast, GLA representations effectively ignore directionality because if either p(x,y) or p(y,x) are greater than the user-specified threshold, there will be a line between *x* and *y.* Directionality can then be indicated by adding arrowheads, but these play no part in the calculations of node coordinates. Consistent with Kruskal and Wish (1978), we find that MDS is ill suited to asymmetric data and GLAs provide a more appropriate representation.

### Interactivity

In the previous sections, we compared MDS and GLA techniques on how they handle data with respect to different properties (dimensionality, outliers, transitivity, and symmetry). In these sections, we focus on the ways that tools based on these techniques compare when interrogating a data set.

MDS is not a particularly interactive technique. There are few options that generate different maps: One can choose between metric and nonmetric, and one can choose the number of dimensions, although practical considerations typically constrain the choice to just two or maybe three dimensions. We can preprocess the data in various ways, such as taking logs or normalizing away marginal effects, but these kinds of transformations are normally dictated by the needs of the data, rather than providing ways to see the data from different perspectives. In this sense, for any set of data, MDS techniques typically only generate one publishable visualization.

GLAs, on the other hand, lend themselves to highly interactive implementations that yield alternative visualizations. The algorithms require dichotomous data as input. As a result, tools implementing GLAs, such as NetDraw (Borgatti 2002), make it easy to cut the data at different levels, affording multiple views of the data. In fact, we commonly investigate the structure of the data by systematically visualizing the data at increasing levels of relational strength, to understand when the global structure breaks down into a set of individual clusters, and then when those, in turn, break down to smaller clusters.

As an example, revisit Figure 7, which shows relationships between holidays at levels of .40 (panel a), .45 (panel b), .50 (panel c), and .75 (panel d). These figures show how interactivity can help uncover the structure in the data and also how the GLA-enabled tool separates drawing lines from locating

points in space. Filtering at different levels, the researcher can determine the strength of relationships between nodes indicated by the existence of lines. Looking at panel a, the two components in the graph both seem somewhat eclectic. Clearly, the component on the left has some patriotic component but also has family-/role-oriented holidays connected to it. The component on the right clearly has most of the holidays based in religious tradition, but New Years and Thanksgiving are not specifically religious. By increasing the strength of the relationship represented by from .40 edges to .50 in panel c, we see Halloween, New Years, and Thanksgiving as all unique holidays, and the remaining clusters are easily identified (e.g., patriotic, religious, etc.). The ability to slice the data representation at different levels helps the researcher uncover the underlying structure in the data.

### Specificity and Precision in Visualization

Similarly, because MDS attempts to represent all dyadic relationships in one visual representation, it sometimes has to compromise based on some of the data properties described above (e.g., transitivity). In contrast, GLAs do not compromise and always represent the presence or absence of relationships at a certain level faithfully. In this sense, they are better suited to answering questions about the data at specific levels of precision. For example, returning again to the MDS plot of U.S. cities in Figure 1, it is difficult to determine whether Miami is within 1,500 miles of either Boston or Chicago, or to which it is closer. However, using a GLA and filtering at 1,500 miles (see Figure 8), the presence of a line between Miami and Chicago and the absence of the same between Miami and Boston makes the answer to both questions easily determined visually. As such, GLAs are also very well suited to analyze data when the researcher is particularly interested in the existence or absence of relationships at precise levels, and when accurate answers to precise questions are more important than a single representation of the range and relative strength of relationships considered along a continuum.

### Manual Repositioning of Nodes

It is also important to reiterate that in a GLA representation, the physical distances on the map do not bear an exact relationship to the input data. The only information conveyed in the graph is dichotomous, which items are tied to each other at a specified level of similarity (or dissimilarity). Although GLA-based tools use positions in an attempt to maximize the clarity of the graph, the researcher has the option to arbitrarily reposition the nodes (or clusters of nodes) in GLA software by simply moving nodes to allow for a clearer view of nodes and the relationships between them, or to highlight

TABLE 4
Summary Comparison of Multidimensional Scaling (MDS) and Graph Layout Algorithm (GLA)

| | *MDS* | *GLA* |
|---|---|---|
| Dimensionality of data | Excellent for representing data of low dimensionality (less than or equal to 3). As dimensionality increases beyond three, MDS plots are more difficult to interpret and impracticable to represent in print. | Well suited for complex data of dimensionality greater than 2 |
| Dealing with outliers | Outliers can cause distortion in MDS plots and increase levels of stress | Provides researchers with the opportunity to easily identify outliers at different levels of strength and remove them if necessary |
| Transitivity (triangle inequality law) | Intransitivity in a data set causes distortion in MDS plots and increases levels of stress | Well suited to represent intransitive data |
| Symmetry | Unable to represent asymmetrical relationships | Well suited to represent asymmetric relationships at multiple levels of analysis<br>Directionality of relationships is represented using arrowheads |
| Interactivity of tool | Not suited for interactivity | Provides researchers with the opportunity to easily investigate relationships at different levels of strength |
| Precision in visualization | Compromises to produce one visualization that best represents all dyadic relationships and their strength at once | Represents existence and nonexistence of relationships at a precise level completely accurately but provides no additional information about relative "degree" or strength of those relationships |
| Node positioning | MDS represents similarities and differences among a set of items as Euclidean distances in an n-dimensional space | Placement of nodes is not determined by Euclidean distance and does not carry meaning. This provides the researcher the opportunity to reposition nodes to improve the readability of the graph. |

specific relationships. This is possible because GLA-enabled tools rely primarily on lines to represent the presence or absence of a relationship, and positioning can be used for aesthetic purposes. Of course, researchers accustomed to interpreting MDS representations will need to remind themselves of the fact that the relative position of two nodes is optimized for clarity and not to convey raw proximity data.

Table 4 summarizes the major points from the systematic comparison of MDS and GLA representations in the previous sections.

## CONCLUSION

In this article, we have considered two methods for visualizing proximity data: traditional MDS and newer GLAs. For data that have inherently low dimensionality and conform to the symmetry and triangle inequality properties of metric space, MDS provides very rich representations that encode a great deal of detail about the data. As data depart from this ideal, MDS continues to yield insight into broader features of the data set—the global structure—but can be inaccurate and misleading in the details. GLAs sacrifice specificity for a more cartoon-like schematic representation of the data, but one that is always completely accurate. At any given moment, the GLA can only show which pairs of items have a proximity level greater than a given threshold (which constitutes a loss of specificity), but the representation is never inaccurate in representing this. This approach is economical and can render visible aspects of structures that in MDS would require higher-dimensional representations. When working with data that violate metric space properties, we recommend the GLA representations. In addition, we suggest that the interactive capability of GLA tools make them particularly attractive in familiarizing oneself with the full character of a data set.

It is worth noting that although we have concentrated on cultural domain applications—and therefore proximity matrices derived from pile sort and triad tasks—both MDS and GLAs are appropriate for other proximity matrices such as correlation matrices. This is well known for MDS but not for GLAs. In our experience, GLA representations of large correlation matrices are particularly effective ways to visualize the principal patterns.

## REFERENCES

Batagelj, V., and A. Mrvar. 2003. Pajek—Analysis and visualization of large networks. In *Graph drawing software*, ed. M. Jünger and P. Mutzel, 77–103. Berlin, Germany: Springer.

Blythe, J., C. McGrath, and D. Krackhardt. 1994. The effect of graph layout on inference from social network data. In *Proceedings of Graph Drawing Symposium: Lecture notes in computer science*, ed. F. Branderberg, 1027. Passau, Germany: Springer.

Borgatti, S. P. 1994. Cultural domain analysis. *Journal of Quantitative Anthropology* 4:261–78.

———. 1998. Elicitation techniques for cultural domain analysis. In *Enhanced ethnographic methods: Audiovisual techniques, focused group interviews, and elicitation techniques,* ed. J. J. Schensul, M. D. LeCompte, B. K. Nastasi, and S. P. Borgatti, 115–51. Walnut Creek, CA: AltaMira.

———. 2002. *NetDraw: Graph visualization software.* Cambridge, MA: Analytic Technologies.

Borgatti, S. P., M. G. Everett, and L. C. Freeman. 2002. *UCINET for Windows: Software for social network analysis.* Cambridge, MA: Analytic Technologies.

Borgatti, S. P., and P. Foster. 2003. The network paradigm in organizational research: A review and typology. *Journal of Management* 29 (6): 991–1013.

Davidson, R., and D. Harel. 1996. Drawing graphs nicely using simulated annealing. *ACM Transactions on Graphics* 15 (4): 301–31.

Davis, A., B. B. Gardner, and M. R. Gardner. 1941. *Deep south.* Chicago: University of Chicago Press.

Eades, P. 1984. A heuristic for graph drawing. *Congressus Numerantium* 42:149–60.

Freeman, L. 1978. Centrality in social networks conceptual clarification. *Social Networks* 1 (3): 215–39.

Fruchterman, T., and E. Reingold. 1991. Graph drawing by force-directed placement. *Software: Practice and Experience* 21 (11): 1129–64.

Greenwald, A., D. McGhee, and J. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74 (6): 1464–80.

Kamada, T., and S. Kawai. 1991. A general framework for visualizing abstract objects and relations. *ACM Transactions on Graphics* 10 (10): 1–29.

Klovdahl, A. S. 1985. Social networks and the spread of infectious diseases: The AIDS example. *Social Science and Medicine* 21 (11): 1203–16.

Kruskal, J. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29 (1): 1–27.

Kruskal, J., and M. Wish. 1978. *Multidimensional scaling.* Beverly Hills, CA: Sage.

Richardson, D. C., and J. S. Richardson. 1992. The kinemage: A tool for scientific communication. *Protein Science* 1 (1): 3–9.

Roethlisberger, F., and W. Dickson. 1939. *Management and the worker.* Cambridge, UK: Cambridge University Press.

Torgerson, W. S. 1958. *Theory and methods of scaling.* New York: John Wiley.

Weller, S. C., and A. K. Romney. 1988. *Systematic data collection.* London: Sage.

*RICH DEJORDY is a doctoral student in the School of Management at Boston College. His research interests include institutional theory and conformity in and of organizations, social networks, and temporal perspectives of organizational phenomena. He has published in* Group and Organization Management *(forthcoming) and the* Academy of Management Best Paper Proceedings.

*STEPHEN P. BORGATTI is a professor and the chair of the Organization Studies Department at the Carroll School of Business, Boston College. His research interests include social networks and knowledge management.*

*CHRIS ROUSSIN is a doctoral student in the Organization Studies Department at Boston College. His research interests include trust, psychological safety, and leadership dynamics in work teams. He has published in the* Academy of Management Best Paper Proceedings.

*DANIEL S. HALGIN is a doctoral student in the Organization Studies Department at Boston College. His research interests include social network analysis, career trajectories, and research methodology. He has published in the* Academy of Management Best Paper Proceedings.