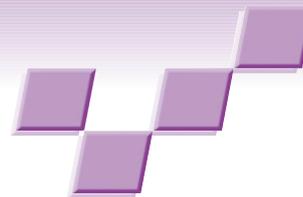


# Visualizing Diversity and Depth over a Set of Objects



Jason Pearlman, Penny Rheingans,  
and Marie des Jardins  
*University of Maryland, Baltimore County*

In many domains, the user is interested not only in including objects with particular desired values, but also in the distribution of values in the set. We refer to the former preference as the user’s depth preference, and the latter as the diversity preference. For example, in college admissions, students with good grades and extracurricular achievements are preferable (depth), but it’s also desirable to have an incoming class with a diversity of personal interests and socioeconomic backgrounds (diversity). In Web searches, a user only wants to see documents relevant to her query (depth), but would ideally like documents from a variety of sources (diversity).

**Our approach for visualizing a set of objects uses glyphs overlaid on a composite representation of the entire set to convey objects’ depth and the set’s diversity. We test and apply this technique to three application domains: analyzing student applicant pools of a particular school or department, building an effective fantasy football team, and analyzing traffic activity on a network.**

Our research aims to provide a visualization that enables a user to understand both the attributes of individual members of the set (depth) and the distribution of attribute values across the set (diversity). Our approach combines glyph-based multivariate visualization techniques with barycentric multidimensional layout methods. We let users select specific attributes of interest to emphasize the information that’s most relevant for their application. For each application, we expect to be able to identify the depth of elements in a manageable set, but also be able to compare those elements with the diversity of

a larger, less-manageable set. We expect to provide the user with a view of the set and additionally provide a view into the set’s individual members without losing the set’s context.

Many existing techniques for visualizing sets result in a loss of information about individual members. Commonly used visual techniques for this problem include bar graphs and parallel coordinates. Although a bar graph adequately conveys information about the collection as a whole, it doesn’t show the diversity of set members, nor does it show dependencies between attributes. Conversely, parallel coordinate representations can be effective at showing diversity and some pairwise

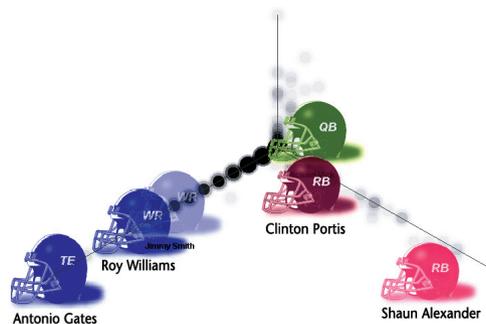
attribute dependencies, but the user can’t easily distinguish individual instances in the display.

We can apply our set of visualization methods (see Figure 1) to a variety of data sets containing objects with multiple attributes. In our preference-modeling project, we’ve been working with several different data sets, including music playlists, scientific image data, and movie recommendation data.

In this article, we apply our techniques to three application domains: graduate admission pools, fantasy football teams, and computer networks. Understanding the changing nature of graduate admissions pools is important for the continued success of computer science departments in the face of changing global conditions. We define the depth (or academic quality) of the admission pool by attributes such as Graduate Record Examination (GRE) scores, country of origin, and gender while the diversity will represent how diverse the entire set is among those attributes.

Fantasy football is an ideal domain because it’s impossible to configure a winning team without considering both the team’s depth (quality and experience of the individual players) and diversity (coverage of different skills).

Analysis of a computer network allows administrators to monitor the amount and types of traffic and use this analysis to determine the network’s security implications. The depth here is the network user’s traffic



**1 Visualization of our techniques applied to a fantasy football team that won a fantasy football league. Position and color represents quality of the player by number of touchdowns scored.**

while the diversity discusses the different types of usage for this network. Furthermore, these data sets represent information where the properties of an individual in the set are important against the distribution of properties across the entire set.

### Three application domains

We're interested in the rather abstract problem of understanding the multivariate characteristics of potential sample sets constructed from a much larger population of possibilities. The driving problem here is the development of techniques to construct sets that exemplify both the quality of individual members (depth) and a spread of other attribute characteristics (diversity). The role of visualization in the larger project is to help understand the differing nature of sets produced by different approaches.

### Applicant pools

In the past several years, applications to graduate programs in computer science in the US have declined dramatically. This decline is easy to see from the raw number of applications. Less easy to understand is potential changes to the composition of the applicant pool in the face of decreasing size. We are interested in issues of applicant quality (depth), as well as the distribution of other characteristics of interest (diversity). In constructing an entering graduate class, we are also interested in diversity in terms of gender, country of origin, and degree sought (MS or PhD).

For the purpose of this investigation, we use GRE scores as a crude measure of applicant quality. Although an undergraduate grade point average would be a much better indicator of quality, lack of a consistent scale across institutions and countries makes this problematic. For instance, US universities tend to use a 4.0 scale, Chinese universities tend to use a 100-point scale, and Indian universities might use either. Worse yet, the same score at even two universities that use the same scale may mean something very different, making it difficult to automatically compare grades.

Specific questions we would like to answer include

- Is the pool of those admitted more or less diverse than those applying?
- How does the student's country of origin impact the applicant's attributes in this study?
- Does an applicant's diversity differ by country?
- As the applicant pool has decreased, has the quality of those admitted gone down?
- How has the geographic distribution of the applicant pool changed over time?
- Has the diversity of the applicant pool changed?
- Has the diversity of admitted students changed?

### Fantasy football

Fantasy football is a game that's based on player statistics gathered from actual football games. A fantasy football league consists of a certain number of teams, usually around 10 or 12. Each team might have a certain number of actual football players on the team, but a player may only belong to one team in the league. Each

team's performance is based on the aggregate performance of the team members in real football games. Fantasy football teams accumulate points, following league rules, where a player's actual performance is translated into a certain number of points. For example, a passing touchdown might be worth four points to the fantasy team that includes the player who threw the passing touchdown. Fantasy teams in the league compete with each other to see which team accumulates the most points over the actual football season.

The goal of the fantasy football visualization for a user might be to determine the team's strengths and weaknesses. These insights would be useful during a fantasy draft, when team owners select players for their fantasy team. The visualizations in this article use US National Football League data for the 2004–2005 season, collected from a fantasy football Web site hosted by Yahoo (see <http://sports.yahoo.com/nfl>). Fantasy football players often use the previous season to build their team for a new season, since the previous year is generally a good predictor of how well a player will perform the following year.

Attributes of interest about each player include the number of passing touchdowns, receiving touchdowns, rushing touchdowns, skill position, height, age, and name. Passing touchdowns are passes a player throws that resulted in a touchdown. For each passing touchdown, another player must also receive that touchdown, known as a receiving touchdown. A rushing touchdown is awarded to a player who scores a touchdown without catching a pass before scoring. The skill position refers to which role the player plays in the football game—in this application, we include only quarterbacks, wide receivers, tight ends, and running backs. Because performance is the primary attribute of concern, it's mapped to the strongest visual cues—location and color. Also, the different types of touchdowns are the biggest factor in determining if a team is diverse in its performance.

The specific questions we look to answer are

- Is a fantasy football team diverse in its players?
- Does a fantasy football team have more diversity than the general set of players?
- How likely is a team to have a player that performs well in each touchdown category?
- What is the amount of diversity on teams that perform well in the fantasy league versus teams that don't?
- Does depth in a particular performance area overcome the lack of diversity?

### Network traffic

Network administrators monitor computer networks for various reasons, including security, limitations, stability, and connectivity. Understanding the network aids the administrator in any form of monitoring. Once again, a depth and diversity relationship exists. Understanding the type and amount of traffic throughout the network is a diversity problem while understanding the reasons that traffic exists in the network is a depth problem (because all of the traffic is produced from a

## Related Work

Many researchers have used glyph-based visualizations for multivariate visualization. Healey and Enns<sup>1</sup> describe several types of glyph alterations that can be used for this purpose. They describe an approach for mapping multiple attributes using a color scale. They find that shape and color can interact, that too many visual cues can lead to interference and reduced understanding, and that varying height and density has no effect on a user's ability to identify the colors in a glyph. We used these findings to select our mappings, specifically combining height and color together to map multiple attributes.

Ebert et al.<sup>2</sup> describe 3D glyph drawing techniques that can represent up to eight different scalar values. These techniques include altering the glyph's location, size, color, and opacity. Altering the glyph's shape, color, and transparency in combination was shown to be an effective technique for visualizing multivariate data.<sup>3</sup> In a more generalized sense, Bertin provides analysis of each of the techniques we apply to the glyphs.<sup>4</sup>

Parallel coordinates are a commonly used method for multidimensional visualization.<sup>5</sup> As we mentioned, these techniques aren't appropriate for our goal of visualizing both diversity and depth. In particular, although parallel coordinates can effectively convey the set's diversity, it's difficult to see the individual objects' attributes, so depth is not easily understood. Furthermore, parallel coordinates don't offer a technique to highlight the data's key attributes.

VisDB<sup>6</sup> demonstrates another multidimensional visualization technique, but it is too generalized for our approach. VisDB is an interactive application that combines multiple techniques to let users find a visualization technique that works for their data set.

Kandogan<sup>7</sup> introduces a technique for visualizing trends and outliers. Kandogan uses a star coordinate system, which aligns an axis with each attribute of the data elements, to visualize multivariate data. This technique provides an effective way to find trends, outliers, and overall set diversity. However, like the similar parallel coordinates technique, it's difficult to see the depth of a particular member of the set.

## References

1. C. Healey and J. Enns, "Large Datasets at a Glance: Combining Textures and Colors in Scientific Visualization," *IEEE Trans. Visualization and Computer Graphics*, vol. 5, no. 2, 1999, pp. 145-167.
2. D. Ebert et al., "Procedural Shape Generation for Multidimensional Data Visualization," *Proc. Data Visualization 1999*, Springer-Verlag, 1999, pp. 1-2.
3. C. Shaw et al., "Using Shape to Visualize Multivariate Data," *Proc. Conf. Information and Knowledge Management, 1999 Workshop on New Paradigms in Information Visualization and Manipulation*, ACM Press, 1999.
4. J. Bertin, *Semiology of Graphics*, Univ. of Wisconsin Press, 1983.
5. Y. Fua, M. Ward, and I. Rundensteiner, "Hierarchical Parallel Coordinates for Exploration of Large Datasets," *Proc. IEEE Visualization*, IEEE CS Press, 1999, pp. 43-50.
6. D.A. Keim and H.-P. Kriegel, "VisDB: Database Exploration Using Multidimensional Visualization," *IEEE Computer Graphics and Applications*, vol. 14, no. 5, 1994, pp. 40-49.
7. E. Kandogan, "Visualizing Multi-Dimensional Clusters, Trends, and Outliers Using Star Coordinates," *Proc. 7th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, ACM Press, 2001, pp. 107-116.

single computer on the network). A computer network's specific attributes that are of interest include the amount of traffic, the type of traffic, and the time that the traffic was present in the network. The type of traffic is the key attribute in this application domain and includes Web browsing, file sharing, chat/messaging, and email traffic.

The goal of this visualization is to aid in understanding a network to diagnose problems or security concerns. Because of this, we chose three different types of network traffic to serve as this application domain's key attributes. The visualization will identify the diversity of the types of traffic occurring on the network, which will identify what the network is primarily used for. Knowledge of the network's primary usage can influence decisions made to optimize the network. For example, if there are large amounts of Web browsing traffic, a Web caching solution could limit the amount of traffic that leaves the local network. Furthermore, the amount and type of traffic that each node produces will be apparent, so we could identify a particular user that produced significantly more traffic than others and what that user is doing. Our example data set includes a college's student network use (simulated to create a larger scale visualization). Some of the questions we attempt to answer include the following:

- What type of traffic is this network most used for and least used for?
- What type of activity do the users with the largest amount of traffic participate in?
- Are there particular users who have a large amount of activity associated with them who might be imposing on bandwidth limitations?
- How different is each user in the network in terms of what type of traffic they produce?

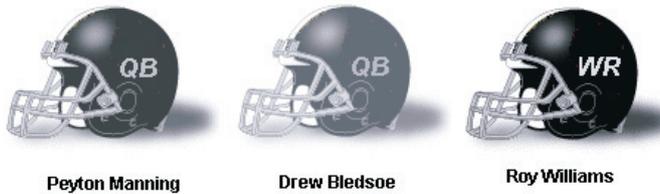
## Approach

In considering the best approach for successful, accurate visualizations, we reviewed a number of previous efforts (see the "Related Work" sidebar for more information). For our research, we determined that the visualization must be able to display both the depth and the diversity of the set of objects. The visualization should also let the user easily identify and understand the individual objects in the set. For instance, it should be apparent to a viewer if a set being visualized has several objects with a high value of one attribute, but no objects with a high value for a second attribute.

We used glyphs to represent each object. The glyph's attributes include size, opacity, color, and location. We designed the glyphs with the application tasks in mind, mapping the most important variables to the most visu-



2 Three players with varying heights.



3 Three different players showing the age to opacity mapping.

ally reliable attributes. We primarily selected glyph attributes that would be separable, such as those that didn't interfere with each other. For example, the visual system processed size and color attributes independently using different subsystems, so values of one have a limited impact on the perceptions of the other.

Also, whenever possible, we leveraged the natural connotations of words. For example, we used glyph height to represent a player's height or glyph color to match the glyphs according to a team's flag colors. In cases where we used integral dimensions, it was generally for item attributes that logically mixed. For example, we used three color components to represent different aspects of football scoring and network traffic, with the three variables mixing together to form an intuitively single property. In both cases, we redundantly mapped those variables to position to preserve the univariate information. Colin Ware provides a more detailed analysis of choosing appropriate glyphs in his book.<sup>1</sup>

In addition to glyphs for individual set members, we present a composite view of the set's population. This view shows fewer attributes, but gives an overview of a population in a much more scalable way. We choose solid icon glyphs that relate to the data set being visualized so that they stand out from the composite view. We use the glyph and composite views together to show the depth and diversity of a small selected set in the context of a much larger population.

**Glyph attributes**

Users have the capability to alter several of the glyph's properties, which helps them more easily identify the object that the glyph represents. In our approach, we modified four glyph properties to suit our needs: scale, opacity, color mapping, and location. We then normalized each technique to the data set, so that the visualization can handle the largest variation of each glyph's property.

**Scale.** The first glyph mapping technique used is scaling. A glyph can be scaled according to one or two attributes of the visualized object. Because scaling can occur on both axes, it's possible to map two attributes to scale—one attribute to x-axis scaling and another attribute to

y-axis scaling. Furthermore, if it's important to keep the glyph's dimensions to its original scale, one attribute can be mapped to scaling the glyph equally along both axes. The formula used to scale the glyph along the y-axis is  $y\ scale_i = (v_i - (v_{min} - 1))/2$ , where  $v_i$  is the value of the attribute for element  $i$  and  $v_{min}$  is the minimum value of the attribute across all the set's elements.

The most intuitive and direct scale application is in the football example visualization where we map the player's height to the glyph's y-axis scaling. In this way, a player's height is mapped to the glyph's height. We selected the value 5.0 as the minimum attribute value, because there's no player in the National Football League who's less than five feet tall. A five-foot-tall player will have a scale factor of 0.5; a six-foot player will be represented by an unscaled glyph ( $y\ scale_i = 1.0$ ).

Figure 2 shows glyphs with height mapped to the y-axis scale for several players in the NFL, with a legend below. The three players that the glyphs represent are Peyton Manning (with a height of 6'5"), Roy Williams (with a height of 6'2"), and Jamal Lewis (with a height of 5'11"). It's straightforward to interpret the glyph height to see the player's height.

**Opacity.** Another attribute value is mapped to the glyph's opacity. To achieve this mapping, a glyph's alpha value is altered according to the represented object's attribute value. The lower the alpha value, the lighter the glyph will appear. When using this technique, it's important to adjust minimum values in the formula so that relevant attribute values are still noticeable in the visualization. If an alpha value of zero is applied to a glyph, the glyph will no longer be visible. The general formula used to calculate the opacity is  $opacity_i = 1 - (v_i - v_{min})/P$ , where  $P$  is the largest value that keeps the opacity scale factor above zero for the entire data set.

For the football data set, age is mapped to the glyph's opacity. As the player gets older, the glyph loses opacity, giving the effect of older players fading out. Younger players will have large alpha values. By mapping age to opacity, a scale factor of one or greater will show the glyph with full opacity so that any player of age 20 (the minimum value, or  $v_{min}$ , for this data set) or less will be displayed with full opacity. For the football visualization,  $P$  was set to 28, meaning that any player less than 48 years will be shown with some opacity. A player aged 48 or older will be too old to be seen in the visualization, but there has never yet been a player that old in the NFL.

Figure 3 shows examples of three different players with their ages mapped to opacity. Peyton Manning was born in 1976. Drew Bledsoe, the oldest player in this example, was born in 1972. Finally, Roy Williams was born in 1981 and is the youngest player in this example. Roy Williams' glyph representation is rendered with a full opacity.

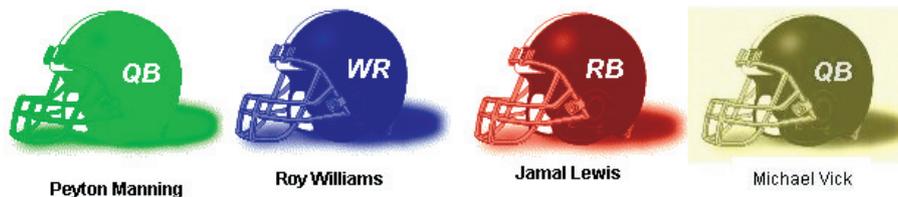
**Color mapping.** Color mapping represents three different attributes. Each of the three attributes alters the value of one of the color values: red, green, or blue. A large attribute value creates a strong value of the respective color. Color and position are the most visually obvious techniques in this article so they should be mapped to important attributes. We normalize the

general formula for mapping each color value to the largest value in the set as  $\text{color}_i = \text{glyph color} + (v_i/v_{\max})$ . The formula is additive in that it doesn't take color away from a pixel value, so a strongly colored glyph would produce poor results. In fact, a base glyph with low color values produces the best results because it allows for a wider range of values to be mapped to it.

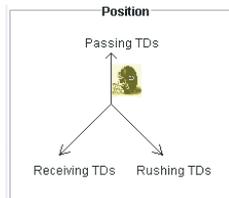
One of the key attributes in the fantasy football visualization is the number of touchdowns. Three types of touchdowns are possible (passing, rushing, and receiving), so we map these to three different colors. Passing touchdowns are mapped to green, receiving touchdowns to blue, and rushing touchdowns to red. For example, the intensity of green in the glyph will increase as the player that the glyph represents gets more passing touchdowns.

Figure 4 shows three different players who each participate primarily in a different touchdown category. Roy Williams, who is represented by a blue glyph, caught eight receiving touchdowns in the 2004–2005 season. Since he got nearly half of the receiving touchdowns of the player who got the maximum number of receiving touchdowns, his glyph representation has been increased by around half of the total blue value available for each pixel. Peyton Manning, represented by the green glyph, receives a maximum green color value for each pixel in the glyph because Manning's 49 touchdown passes in the 2004–2005 season was the most by any player. Jamal Lewis got seven rushing touchdowns in the 2004–2005 season, which is less than half of the maximum number of rushing touchdowns. His glyph representation has a noticeably weaker coloring than Manning's.

We chose a multivariate color scale to represent the different methods of scoring touchdowns to allow for players who score touchdowns using multiple methods. Generally, a quarterback only produces passing touchdowns, a running back rushing touchdowns, and wide receivers and tight ends receiving touchdowns. However, sometimes a quarterback not only throws for touchdowns, but also rushes for touchdowns. Although rare, this information is valuable for a fantasy team owner, as the fantasy team will receive points for both passing and rushing touchdowns from that player. By using the multivariate RGB scale, it's possible to see a player that receives touchdowns in multiple categories. In an extreme (and very unlikely) case, a glyph that's completely white would indicate maximum color values in each touchdown category because every value of the RGB scale would be maximized. Figure 4 includes a player (Michael Vick) who had both passing and rushing touchdowns in the 2004–2005 season, although he had more passing touchdowns than receiving touchdowns. The brownish tint of the glyph indicates that there is some red value in the glyph, but the overall color still appears more on the green side, indicating that this player's primary method of scoring touchdowns is through the air.



4 Three glyphs representing three different players who each participate in different touchdown categories.



5 Barycentric position legend indicating which attribute pulls the glyph in which direction.

**Location.** Location is another strong visual cue. We designed this project to produce a 2D visualization, but it has three key attributes to visualize. These key attributes are mapped to the RGB color scale. However, to emphasize these attributes more, and to aid in showing diversity among different object sets, these key attributes are redundantly mapped to location and color.

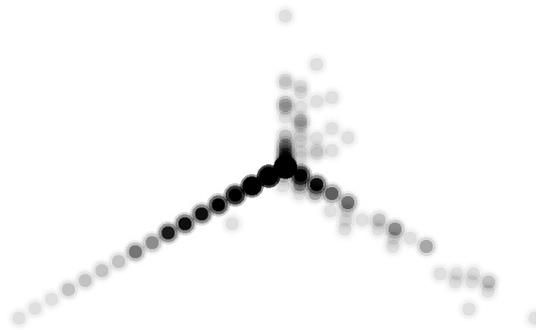
A method for mapping three attributes to a position in a 2D plane is using a barycentric coordinate system. A barycentric coordinate system is based on three vertices of a reference triangle, effectively (1, 0, 0), (0, 1, 0), and (0, 0, 1). Then, each triplet (x, y, z) is mapped within the reference triangle and pulled toward the vertex that corresponds to the attribute mapped to that number that is part of the triplet. There is a limitation to using Barycentric coordinates. The barycentric coordinates (0, 0, 0) would map to the same location as (0.25, 0.25, 0.25), which would both correspond to the center of the reference triangle. However, since the key attributes are mapped to color as well as location, the colors will vary for those two different attribute sets, giving an indication of the absolute value even though the location is relative.

The three key attributes in our fantasy football visualization are doubly mapped to location using barycentric coordinates. Figure 5 shows the barycentric space defined by passing, rushing, and receiving touchdowns.

### Background overlay

To show the diversity of the set being visualized versus the diversity of the entire set of possible objects, the background is overlaid with the entire set's location information. At each point on the barycentric coordinate map, the intensity is determined by the number of elements in the entire set that would be located at that point. Based on the three attributes chosen to map to location, the background shows the set's diversity.

To present a smooth background overlay, we use a Gaussian kernel. For each element in the entire set, we find its barycentric position using the three attributes chosen to represent location. Then, we apply a Gaussian kernel to the position, darkening points around the



**6 Background overlay of the football player data set. The darker the spot on the barycentric coordinate plane, the more players that fall into that category.**

element's position. The further away a point is from the element's position, the smaller the darkening effect of the Gaussian kernel until the effect gradually drops off to nothing. For the example application domains, the background began white, and color was subtracted using the kernel. This is similar to the Graph Splating technique, which allows for smooth visualizations of large data sets.<sup>2</sup>

If appropriate for the data set, we can apply color to the background overlay. Each element of the data set also has a color associated with it using this method. Instead of applying a Gaussian kernel to a point's intensity, we apply the kernel by making the points' color closer to the representative color of the element in the set. Again, for the example application domains, we begin with a white background. Then, for each point on the barycentric coordinate plane, we alter the point and the surrounding points' color to be closer to the color of the element at that point. We do this by subtracting color from the RGB scale based on an inversely proportional function to the element's color. For example, if the element's color was blue, then we would subtract both red and green using the Gaussian kernel from the element's point.

In the fantasy football example, this is useful for measuring the quality of a team versus the quality of all players in the league. Figure 6 shows the background overlay of all skill position players for the 2004–2005 season.

Figure 6 emphasizes one of the widely used strategies of fantasy football veterans: draft a top running back early. Rushing touchdowns favor the southeast axis of the barycentric plane. In the visualization, it's apparent that a small but noticeable set of players fall far from the center along the rushing touchdown axis, meaning there's a small cluster of players that score significantly more rushing touchdowns than other players. Furthermore, outside of that small cluster of elite running backs, there's a fairly large drop-off in touchdowns to the next set of running backs, indicated by the white space between the bottom right cluster and the other data.

Our system's overlay also shows that players with receiving touchdowns experience a gradual drop-off in the number of players as the number of receiving touchdowns increases. Furthermore, of the quarterbacks in the league, those who throw for passing touchdowns cluster near the middle of the passing touchdown axis. Partially, this is because Peyton Manning set the pass-

ing touchdown record in the 2004–2005 season, which was significantly higher than all other quarterbacks' passing touchdowns. However, it does show that many mediocre quarterbacks exist in the league, with only one elite quarterback. This implies that significant effort shouldn't be placed in attempting to obtain one of the better quarterbacks in the league (excluding Peyton Manning), as there are many other quarterbacks who will produce a similar number of passing touchdowns.

### Application results

The application of these techniques produced visualizations in which a user could draw conclusions about the data set. Some of the conclusions were unexpected. For the fantasy football application domain, we can analyze and compare depth and diversity of a team with the depth and diversity of another team to predict a winner. In the student admission application domain, we can visually compare diversity over the set of admitted students from year to year. Finally, in the network traffic domain, we can isolate the primary form of traffic as well as the users who produce the most activity.

### Fantasy football

Once a fantasy team is rendered in the same image, it's fairly easy to see the diversity among the team's players. Five different attributes help show this diversity, but the most obvious one is color. This is important because color equates to the most important statistic for a fantasy football team owner—touchdowns.

The strength of these visualization techniques is that the result displays both the depth of objects in a set, as well as diversity over the set. Figure 7 shows diversity in images of two different object sets.

Figure 7a shows a complete visualization with a legend of all attribute mappings and two fantasy football teams. This is actually visualizing the starters from a fantasy football play-off game for the 2005–2006 season. The team visualized on the left side ended up winning the play-off game. Both teams are reasonably diverse but both also suffer by not having a player with a large number of passing touchdowns, represented by green values in a glyph.

However, the team on the left has players that participate in multiple touchdown categories while the team on the right has players that only participate in one touchdown category. We can figure this out by looking and seeing that the team on the right has all of the glyphs lined up along the axis that the barycentric coordinate plane created.

The team on the left has two players that appear to participate in multiple touchdown categories. Shaun Alexander, found in the bottom right of the left team, has a large number of rushing touchdowns but also has some receiving touchdowns. This is demonstrated by the pink appearance of the glyph as well as the pull to the left of the glyph, off the rushing touchdown axis. This means that this player has more chances to get a touchdown than someone like Willis McGahee, the strongest running back on the right team.

Furthermore, the receivers on the left team are further down the receiving axis as a whole than the receivers on

the right team, meaning the team on the left also has advantage when comparing receivers. Oddly enough, neither championship team has a player who performs relatively well in the passing touchdown category, implying that having a good quarterback isn't so important to field a winning team. This unexpected discovery has implications for standard fantasy football drafting strategies that primarily focus on getting a running back, but also have quarterbacks being taken in the first round.

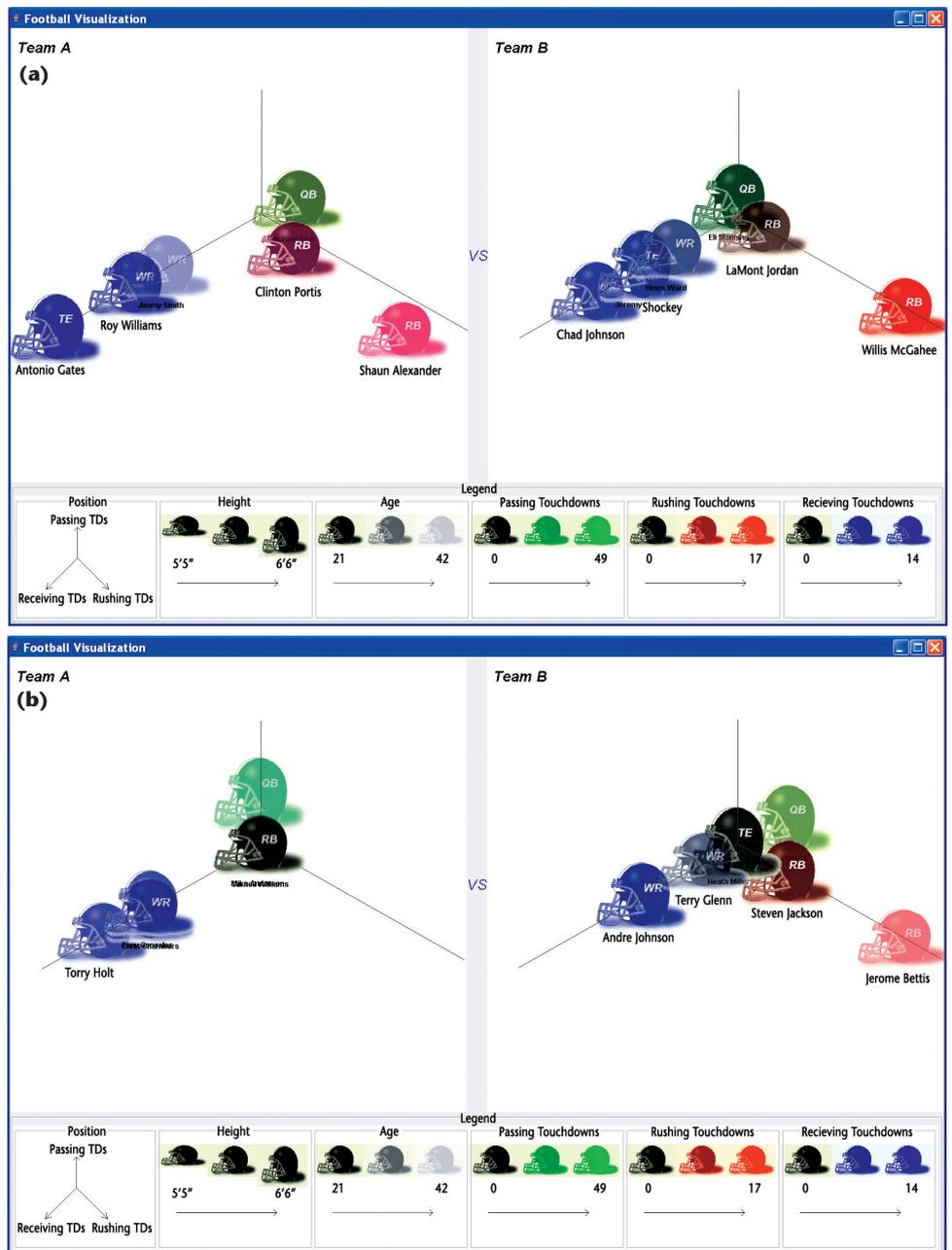
The left set of Figure 7b shows an example of a set which isn't very diverse, as there are no glyphs near the right of the triangle or any red color in the image. For the football data set example, this means the team is lacking rushing touchdowns. Furthermore, it's easy to see depth of an individual object. For example, in Figure 7a, the left image has one glyph that stands out among the other glyphs in the bottom left corner. Because this glyph is in the bottom left of the image and has a high blue value, it indicates that this player scores a large number of receiving touchdowns.

Figure 7b shows another complete visualization of two fantasy teams. The left team represents a fantasy football team in last place. Notice the lack of diversity that the left team has, that could have predicted the team's performance.

In fact, the left team has a lack of rushing touchdowns, represented by no red values in the image and no glyphs moved toward the right triangle's vertex. The left team has a stronger set of players who have larger values of receiving touchdowns but this wasn't enough to keep the team out of last place.

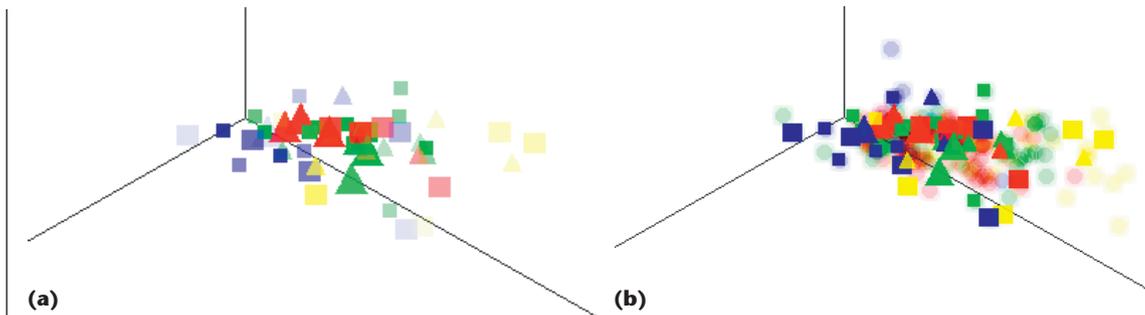
This team would have done well to make some roster moves to find a player with some rushing touchdowns, possibly at the expense of trading some of its receivers (for example, gaining some red and losing some blue). In fact, good running backs are commonly considered the key to a successful fantasy football team, thanks to their consistency and rarity.

The right team represents a more diverse team, but you'll notice it has only a few deep players, who have large numbers in a particular touchdown area. The team also has a good example of an older player (Jerome Bettis) who still performed well; see the bot-

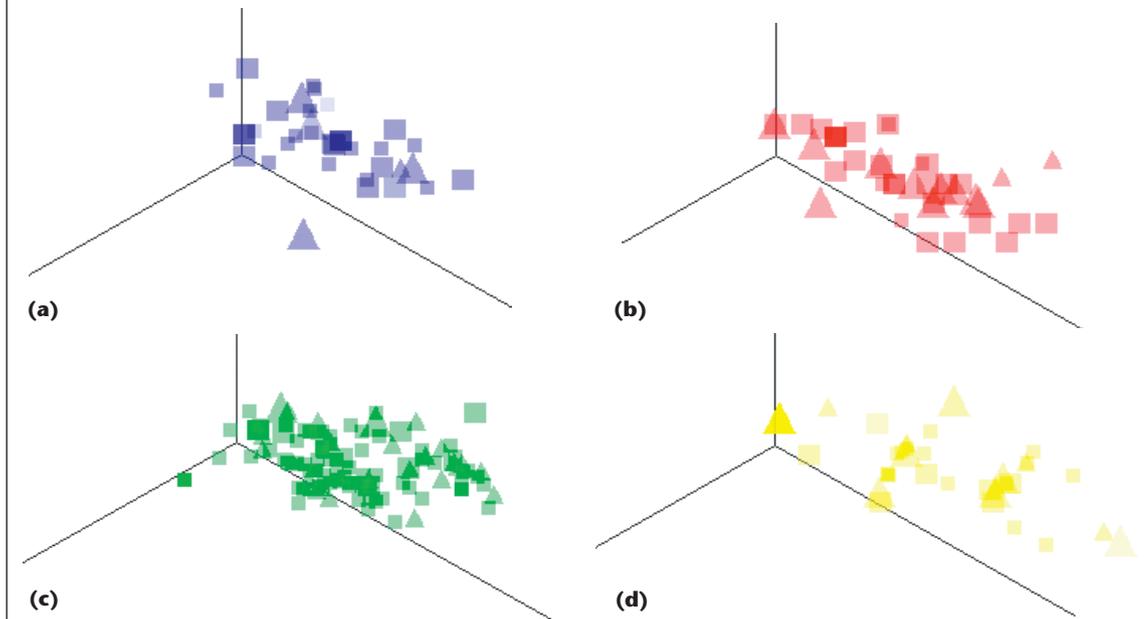


**7 (a)** Complete visualization showing a legend of all attribute mappings of two fantasy teams. The team on the left ended up winning first place in a fantasy football league. **(b)** Complete visualization of two other fantasy teams. The team on the left represents a fantasy football team in last place.

tom right corner of Figure 7b. In fact, Bettis is the deepest player on this team, but his age is catching up with him, and he's currently not performing as well in the 2005–2006 season as he did in the 2004–2005 season, which this visualization shows. The team on the right did outperform the team on the left over the season, and they're looking at a finish somewhere in the middle of the league. Perhaps recognizing that the deepest player on the team is getting old by looking at this visualization could alert this team to look to the future by finding younger players who could eventually perform well.



**8** Students admitted in Fall 1999. (a) Glyphs for admitted students. (b) Background shows the entire applicant pool.



**9** Characteristics of applicants by country of origin (same mappings as in Figure 8). (a) Applicants originating from the US, (b) China, (c) India, and (d) other countries.

**Applicant pools**

The glyph and overlay views help answer questions about changing depth and diversity in the applicant pool. Figure 8a shows all students admitted in Fall 1999. The system represents each admitted student with a glyph: squares for men and triangles for women. Glyphs for PhD students are larger than those for MS students. Each student glyph is colored by the students' country of origin: blue for the US, red for China, green for India, and yellow for all others. Applicants' GRE scores are mapped to barycentric coordinates: southeast for quantitative, southwest for verbal, and north for analytic. Students with higher composite GRE scores are more opaque.

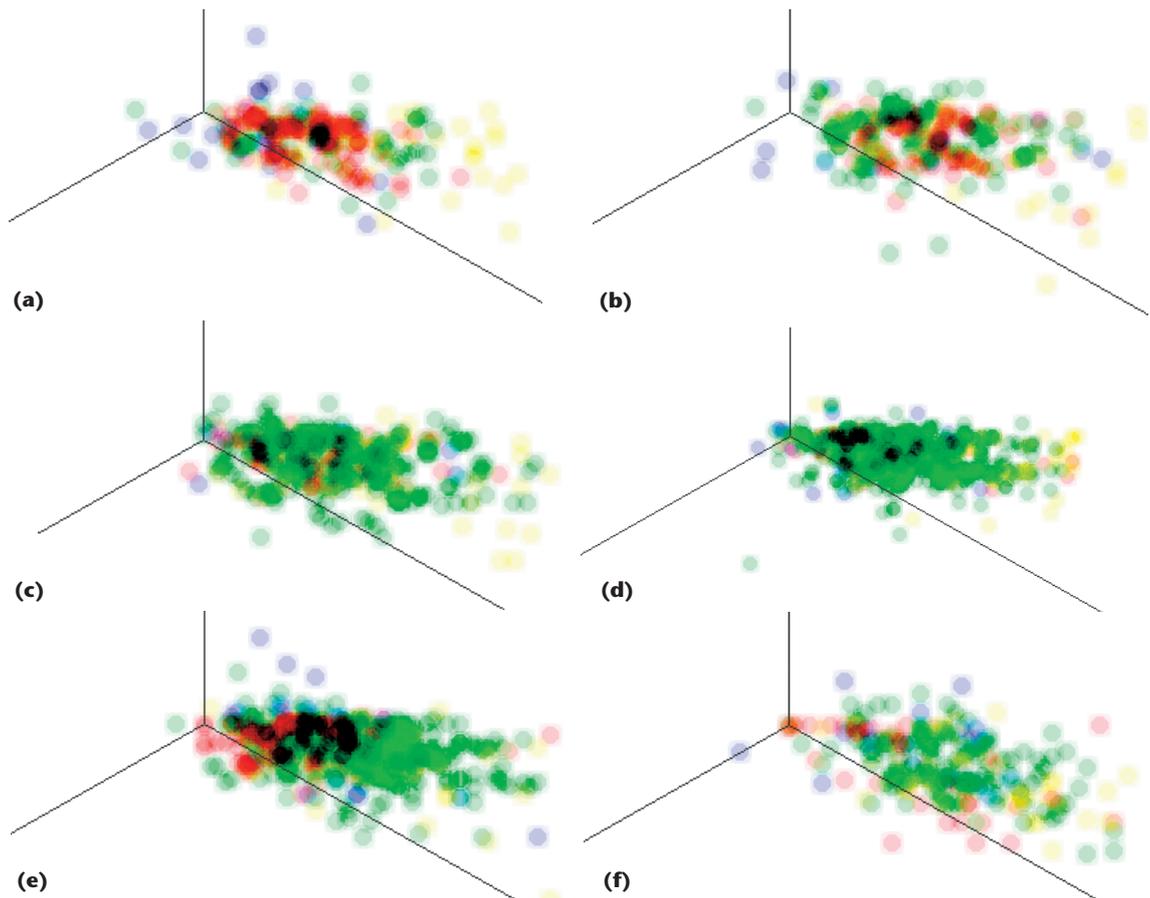
The glyph cluster in the center of the display represents a group of domestic applicants with balanced scores on the three parts of the GRE. Students from China (red), India (green), and especially other countries (yellow) are more likely to have higher quantitative GRE scores. Male students predominate through this group, particularly among domestic students. Figure 8b shows that same group of admitted students against the backdrop of all applicants. The characteristics of this group generally mirror those of the admit-

ted group, but applicants with extremely unbalanced scores seem less likely to be admitted.

Figure 9 shows differing depth and diversity characteristics of applications from different countries of origin. Applicants from the US (blue) are mostly men, but they're a mix of MS and PhD applicants. Applicants from China (red) almost invariably apply to the PhD program (big glyphs). Applicants from India (green) include a larger percentage of women and almost always apply to the MS program. Applicants from other countries (yellow) vary widely in depth (composite GRE scores are mapped to opacity).

Figure 10 shows the changing nature of the applicant pool during the 1999–2004 period. In the first year, applicants from China (red) are well represented. In the middle years, applicants from India (green) predominate. Toward the end of this window, applicants from the US (blue) become more numerous.

Figure 11 on page 44 shows the changing nature of the set of admitted students from the period of 1999–2004. The set of admitted students tips from China, to India, to the US. During the years India dominates, students were comparatively more likely to be MS-seeking and more likely to be women.



**10** Changes in the applicant pool from (a) 1999, (b) 2000, (c) 2001, (d) 2002, (e) 2003, and (f) 2004. The mappings are the same as in Figure 8.

In 2004, the set is smaller, but contains a greater percentage of PhD students. Although the number of students admitted decreases, the number of weaker students (less opaque) decreases even more quickly. However, the overall trend in this visualization is unexpected. As the number of accepted students decreased over the years, the diversity of the entire set doesn't change much. There's still representation from each of the countries, still a varied number of PhD students and MS students, and most surprisingly, a diverse set of GRE scorers in both the total score and subject focus. This is unexpected, because as the number of admitted students went down, we would expect the admitted students would be a less diverse set of GRE scorers, all with higher scores than previously admitted classes where more students were accepted.

### Network traffic

The visualization techniques applied to a network traffic data set reveals depth and diversity information about the traffic. For this application, we chose the type of network traffic as the key attribute, and doubly mapped its features to the barycentric coordinate system and the glyph's color. While nearly limitless types of traffic exist, we break down traffic type into three broad sets. Red and the north axis of the barycentric coordinate graph means an increased amount of Web traffic. Blue and southeast corresponds to chat-based

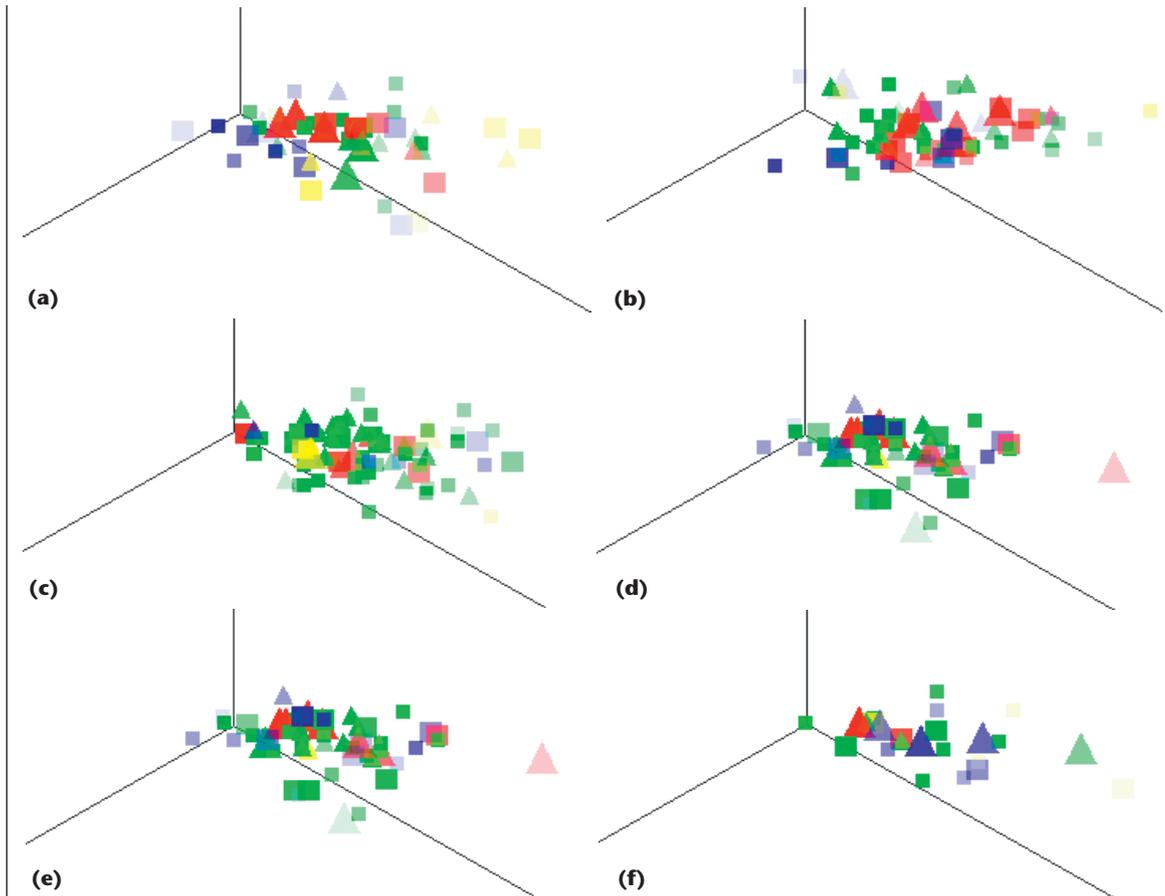
traffic, while green and southwest corresponds to any other type of traffic. We defined chat traffic as Internet Relay Chat or AOL Instant Messenger traffic, both of which are commonly used chat and messaging services.

We break down chat and Web traffic into separate groups because these two categories, along with file sharing, make up the majority of traffic expected from a set of users in a dorm. In this case, the other traffic category will likely consist of large amounts of file-sharing traffic.

Figure 12 shows a display of a small simulated dorm network. The lack of nodes on the southern side of the layout indicates that this network consists primarily of Web traffic, because most users are producing more Web traffic than any other type of traffic. Also, the large cluster of low-activity users in the north of the visualization identify a large number of users that only produce Web traffic and that amount is relatively small.

The node that appears in the east of the layout produces more chat-based traffic than anything else, and this size indicates a large amount of chatting. Finally, the nodes on the west produce traffic that isn't Web or chat traffic, and is likely email and local file-sharing traffic. The nodes with the lowest alpha values have been inactive the longest.

The most surprising element of this visualization is that the Web and chat traffic nearly always outweigh the amount of "other" traffic. This means that any type of gaming or file sharing is outweighed by the amount



**11** Changes in the admitted class from (a) 1999, (b) 2000, (c) 2001, (d) 2002, (e) 2003, and (f) 2004. The mappings are the same as in Figure 8.

of chatting and Web browsing that most users do. The few nodes that are dominated by other traffic do produce a high amount of activity, which is common when users are sharing larger files (such as audio or video). With peer-to-peer file sharing lawsuits making national news in the past few years, we would expect to see larger amounts of file sharing, which will push more nodes into the other traffic section.

**Conclusions and future work**

Because each data set was of interest to at least one of the authors, we self-evaluated the results of each

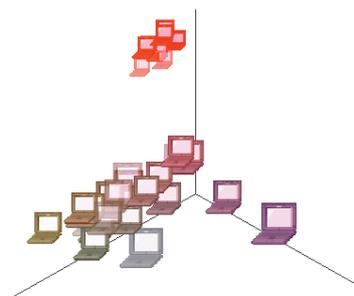
application domain. While we were able to draw conclusions and identify unexpected results in our data sets, a formal evaluation would be more effective in showing the usefulness of our techniques.

What we did succeed in doing with this project is visualize both depth and diversity in each application domain. In the fantasy football application, we were able to infer a team’s quality based on the depth of individual players and the diversity of the set of players in the league. Furthermore, viewing both the winning and losing teams’ results revealed trends that could help a fantasy football owner determine the best strategy.

The student application visualization showed trends over time. This helped us visualize data in multiple areas, including diversity by country, gender, the applicants’ scores, and the quality of admitted applicant’s scores.

The network traffic application gave insight into what a network is being used for and what particular users are doing. Finally, in each application, we found unexpected discoveries by visualizing both the depth and diversity of a set, which someone probably wouldn’t see or notice using standard methods.

The limitations of this visualization are scalability, occlusion, and technique interference. In Figure 7b, some glyphs occlude the others. While the occlusion doesn’t affect seeing the set’s diversity, it’s difficult to see an occluded glyph’s depth. Also, if glyphs are occlud-



**12** Network visualization of depth and diversity. Computer glyphs represent each node. The color and location of the glyph describe the types of traffic produced by the node. The glyph size is the amount of traffic; the opacity represents how recently activity occurred.

ed, the importance of their depth is less relevant, because whichever glyph occluded it has nearly the same properties. Occlusion is related to scalability because as the object set grows, so will the amount of occlusion.

We should also note that we found that techniques such as color and opacity can interfere with each other. The lower the glyph's alpha value, the more difficult it is to see varying color values.

A future extension is to apply focus + context techniques to improve the result's scalability and reduce occlusion. Yang et al.<sup>3</sup> saw similar scalability and occlusion issues in their value and relations work and used various strategies to address those concerns. We could apply their interactive navigation and selection tools to this work to address the same concerns.<sup>3</sup> Finally, we believe that producing interactivity in the glyphs could give users access to more information.<sup>4</sup> ■



**Jason Pearlman** is a graduate student at the University of Maryland, Baltimore County (UMBC), where he recently received an MS. His current research interests include multivariate and network security visualizations. Contact Pearlman at [jpearl1@umbc.edu](mailto:jpearl1@umbc.edu).



**Penny Rheingans** is an associate professor of computer science at UMBC. Her current research interests include volume rendering, information visualization, perceptual and illustration issues in visualization, and nonphotorealistic rendering. Rheingans received a PhD from the University of North Carolina, Chapel Hill. Contact her at [rheingan@cs.umbc.edu](mailto:rheingan@cs.umbc.edu).



**Marie des Jardins** is an associate professor of computer science at UMBC. Her research is in artificial intelligence, focusing on the areas of machine learning, multiagent systems, planning, interactive artificial intelligence techniques, information management, reasoning with uncertainty, and decision theory. Des Jardins has a PhD from the University of California, Berkeley. Contact her at [mariedj@cs.umbc.edu](mailto:mariedj@cs.umbc.edu).

For further information on this or any other computing topic, please visit our Digital Library at <http://www.computer.org/publications/dlib>.

## References

1. C. Ware, *Information Visualization: Perception for Design*, Morgan Kaufmann, 2000.
2. R. Van Lier and W. De Leeuw, "Graphsplatting: Visualizing Graphs as Continuous Fields," *IEEE Trans. Visualization and Computer Graphics*, vol. 9, no. 2, pp. 206-212.
3. J. Yang et al., "Value and Relation Display for Interactive Exploration of High-Dimensional Datasets," *Proc. IEEE Symp. Information Visualization*, IEEE CS Press, 2004, pp. 73-80.
4. H. Siirtola, "The Effect of Data-Relatedness in Interactive Glyphs," *Proc. 9th Int'l Conf. Information Visualization*, IEEE CS Press, 2005, pp. 869-876.

Sign Up Today



For the  
IEEE  
Computer Society  
Digital Library  
E-Mail Newsletter

- Monthly updates highlight the latest additions to the digital library from all 23 peer-reviewed Computer Society periodicals.
- New links access recent Computer Society conference publications.
- Sponsors offer readers special deals on products and events.

Available for FREE to members, students, and computing professionals.

Visit [http://www.computer.org/services/csdl\\_subscribe](http://www.computer.org/services/csdl_subscribe)