Visualizing Document Authorship Using N-grams and Latent Semantic Indexing

Ian M. Soboroff, Charles K. Nicholas, James M. Kukla, David S. Ebert
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
{*ian,nicholas,jkukla1,ebert*}@*cs.umbc.edu*

Abstract

An approach to visualizing authorship and writing style of free-form text documents is described. This approach uses n-grams and latent semantic indexing (LSI) to cluster documents according to usage patterns of related n-gram "terms". Latent semantic indexing distributes documents and terms into a relatively low-dimensional space, which can be viewed graphically using various visualization techniques.

1 Introduction

Determining the authorship of a text is a common problem that not only occurs in scholarly studies of historical documents, but is also of interest when one seeks to classify large numbers of documents. Knowing the author or authors of a text can provide valuable information on its "context" which might otherwise be unknowable from the text itself.

Mathematical techniques have frequently been applied to the problem of document authorship. These approaches analyze such things as frequencies of word usage, sentence structure, and word and sentence length. A typical computational structure used is a matrix of terms versus documents, with each entry a count or frequency of term i in document j. Kjell and Frieder [5] cite several historical studies that use these techniques to "prove" the authors of, for example, the Shakespearean plays and *The Federalist Papers.* The chief difficulty in these early studies was a lack of computational resources available for complex analysis. Since an interesting survey may encompass tens or hundreds of documents, each containing hundreds or thousands of words, issues of time and space complexity made large studies prohibitive.

As computational capabilities grow, it has become feasible to use these statistical approaches on large or otherwise interesting data sets. Two techniques for analyzing free-form text, latent semantic indexing and n-grams, are described below. While these techniques are typically used for information retrieval, they can be applied to the problem of authorship with interesting results. Moreover, latent semantic indexing provides a method for graphical visualization of document relationships. An example using biblical Hebrew texts is presented to illustrate these techniques.

1.1 Latent Semantic Indexing

Deerwester *et al.* [3] discuss a method for automatic topic-based indexing of documents called "latent semantic indexing" (LSI). In a basic approach to text indexing, documents are keyed by literal words in the document. This is problematic because different words are used to mean the same thing, and a single word usually has multiple meanings. Latent semantic indexing takes advantage of semantic relationships between terms and documents. These relationships are derived mathematically from implicit linkings of related words in documents; a typical author will use many words to describe the same idea, and all those words in one or a few contexts. This is what is meant by "latent" semantics.

Specifically, a term-by-document matrix M is first constructed. This matrix is decomposed using the singular-value decomposition $M = T\Sigma D'$ The columns of T and D are orthonormal, and are called the *left* and *right singular vectors.* Σ is a diagonal matrix containing the *singular values* σ, ordered by size. If M is $t \times d$ and of rank r, T is a $t \times r$ matrix, D is $d \times r$, and Σ is $r \times r$. The SVD "projects" both terms and documents into an r-dimensional space; one can choose several interesting dimensions and view documents and terms graphically in that space [2].

1.2 N-grams

In using LSI to determine writing style, we have used *n-grams* rather than discrete words. N-grams are overlapping n-character sequences in a document. Typically, using n-grams in information retrieval applications as opposed to words reduces errors from noisy data, and makes stemming and stop-word lists unnecessary; latent semantic indexing also alleviates these problems. N-grams also have the property that they span adjacent words in the same document. Thus, we achieve an additional layer of granularity in term relationships by capturing local linkages of words into phrases. This provides the mechanism for LSI to discern writing style between documents.

2 Using N-grams to Determine Authorship

In 1992, Kjell and Frieder [5] analyzed the frequency of 2-grams in *The Federalist Papers*, in order to determine the authorship of eleven unattributed papers. In their survey, capitalization, punctuation, and spaces are ignored, so that in the phrase "Hello World", *ow* would be a 2-gram.

Kjell and Frieder compared eleven papers of unknown authorship to two prototype documents, constructed by concatenating a number of papers by two different authors. The cosine similarity was taken between an unknown-author document and both prototypes, and authorship was determined by the highest cosine similarity. Using this measure, Kjell and Frieder arrived at conclusions comparable to a classic scholarly study of *The Federalist Papers.* Additionally, by applying principal components analysis to a set of feature vectors, graphs were produced which demonstrate visually the stylistic relationships between the texts.

Several interesting questions are raised by Kjell and Frieder's work. Are 2-grams large enough to capture writing style? Should inter-word spaces have been included? Is cosine similarity a good measure for determining authorship? Were the feature vectors chosen for visualization good summaries of the documents? We believe that the LSI approach is a better mechanism for deriving feature vectors, which in this case are the actual documents projected into a reduced space. Moreover, the choice of length of n-grams to use is not well understood, and it is probable that preserving word and sentence divisions will improve results.

3 Visualizing Authorship with N-grams and LSI

The example presented here uses latent semantic indexing, with n-grams as terms, to visualize authorship among biblical Hebrew texts. The texts consist of the books of Daniel, Song of Songs, and Ecclesiastes. In this set, which has 32 total chapters, Song of Songs and Ecclesiastes are traditionally attributed to King Solomon. Those books differ significantly in content, and both have a different style than that of Daniel. Moreover, half of the chapters of Daniel are primarily written in Aramaic, a language related to but different from biblical Hebrew. Each chapter has between 680 and 4291 characters. The text was acquired from the Snunit Educational Information System [1], part of the Israeli English Teachers Network. Inter-word spaces are preserved as a single space, and punctuation is converted to spaces, except for periods which mark the ends of verses. In the Snunit text, punctuation is present which is not standard across printed texts, so it is ignored.

Figure 1 shows a plot in the style of Berry *et al.* [2]. The first two columns of the D matrix are scaled by their corresponding singular values, and plotted in x and y respectively. In other words, the two most significant dimensions of document interrelationships, as derived by the SVD, are shown against each other. Note that in the x-direction, corresponding to the first dimension, all chapters of Daniel but one fall to the right of the Solomon texts, and that Daniel falls into two groups. The second dimension, along the y-axis, cleanly divides the books such that half of the chapters of Daniel are separate from the others.

Figure 2 shows the first dimensional value for each chapter in the set. A line dividing the Solomon texts from Daniel is shown for emphasis; no texts attributed to Solomon fall above this line, and only one chapter of Daniel lies below it. We can see that the chapter of Daniel "closer" to the Solomon texts is the last one. In Figure 3, which shows the values in the second dimension, it is easily seen that the isolated chapters in that dimension are the second through the seventh of Daniel, which are the ones written in Aramaic; again, the line drawn along the x-axis is for emphasis.

In other words, in this document set, the most prominent differentiating characteristic found by the SVD is writing style; the second most prominent is language. While the texts with the most outstanding values in the first dimension are the Aramaic chapters of Daniel, only one Hebrew chapter of Daniel intermingles with the Solomon texts. This shows that although language is an influence in the first dimension, it is not the differentiating factor here. The second dimension shows a clear division between those chapters containing Aramaic and those with only Hebrew.

[1] http://www.snunit.k12.il/

These figures were taken from data which uses 3-grams. N-grams of length four or five show similar graphs, with subtle differences. For 4-grams, the first dimension divides more sharply along the Hebrew-Aramaic difference. A plot of 5-grams shows the Solomon texts grouping still more cohesively, and show the chief distinguishing factor to be language. N-grams of length two, similar to Kjell and Frieder but including inter-word spaces, did not show as clean a differentiation for style as did 3-grams.

4 Visualizing Multiple Dimensions

One complexity of LSI is that often more than two or three dimensions are of interest. In the previous figures, while only rendering one or two dimensions, we have added another "pseudo-dimension" by using shapes, or *glyphs*, to highlight known authorship of documents. By combining spatial position with color, shape and size of the glyphs, we can visualize more than three dimensions simultaneously.

Ebert *et al* describe in [4] a method for using shapes to represent continuous data values. This technique is based on superquadrics, a parametrized three-dimensional shape described by Barr [1]. In a superquadric, three parameters are varied to describe size in three dimensions, analogous to how two parameters define the major and minor axes of an ellipse. Two additional parameters are used as exponents.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} a \cos^{e_1} \eta \cos^{e_2} \mu \\ b \cos^{e_1} \eta \sin^{e_2} \mu \\ c \sin^{e_1} \eta \end{bmatrix}$$

In this equation, a, b, and c are the three axial factors. η and μ are the two angles of rotation about which the shape is parametrized. The two exponents e_1 and e_2 cause the shape to be rounded, square, or pointed; from a single equation spheres, ellipsoids, cubes, cylinders, stars, and other shapes can be generated.

Using superquadrics, up to five dimensions could theoretically be shown, using the axial parameters and the exponents. In the following examples, we use a single dimension for a, b, and c, and a single dimension for the exponents.

In Figure 4, the data set is shown again. The first three dimensions, scaled by their singular values as in the above plots, correspond to the glyph x, y, and z locations. The seventh dimension controls glyph shape, while the sixth dictates their size. Color is used to show the known authorship of the documents. Thus, this is image shows six dimensions of the data; five from the LSI and one imposed from external knowledge. We can see that several Solomon documents have a similar glyph shape to some chapters of Daniel, for example.

Our tools allow us to dynamically reassign visualization parameters to dimensions of the data. Figure 5 shows a completely different view of the same data, achieved by changing the dimensional assignments. Here, the fifth, sixth, and seventh dimensions are used as positions. The first dimension is size and the second is shape, and color again shows the known authorship. This allows us to keep track of the language and author differences while we explore the document relationships in other dimensions.

5 Conclusion and Future Work

We have shown that using latent semantic indexing with n-grams as terms can group documents by authorship. This method allows elements of language and style to be indexable

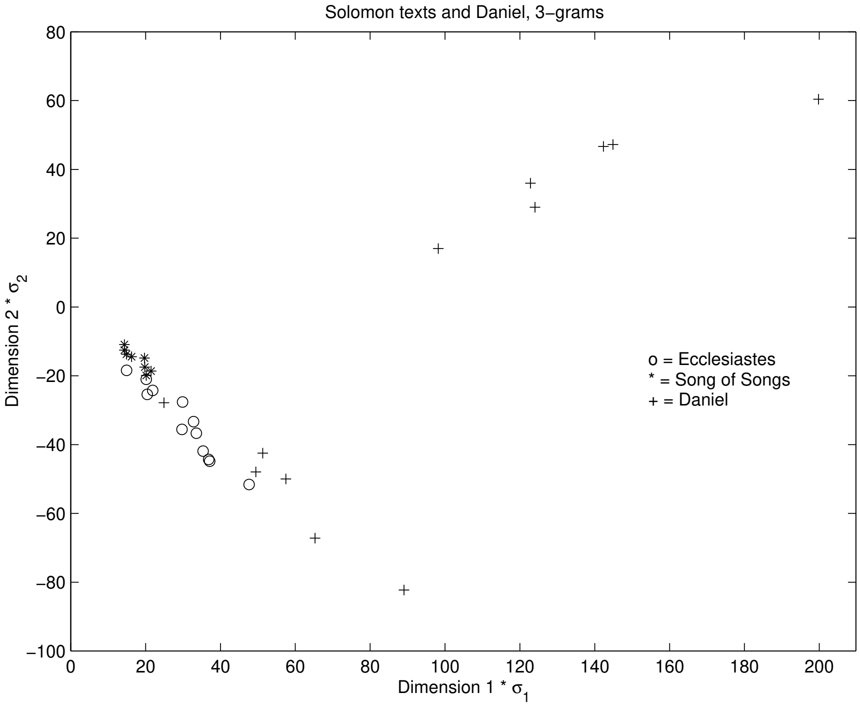

Figure 1: The first two scaled dimensions of the SVD.

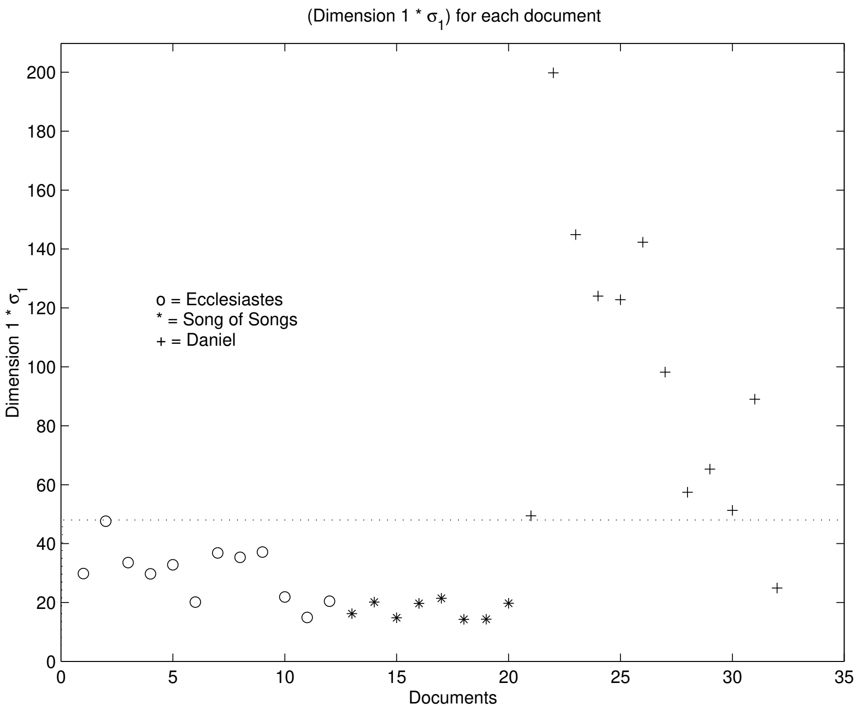

Figure 2: The first dimension values for each book.

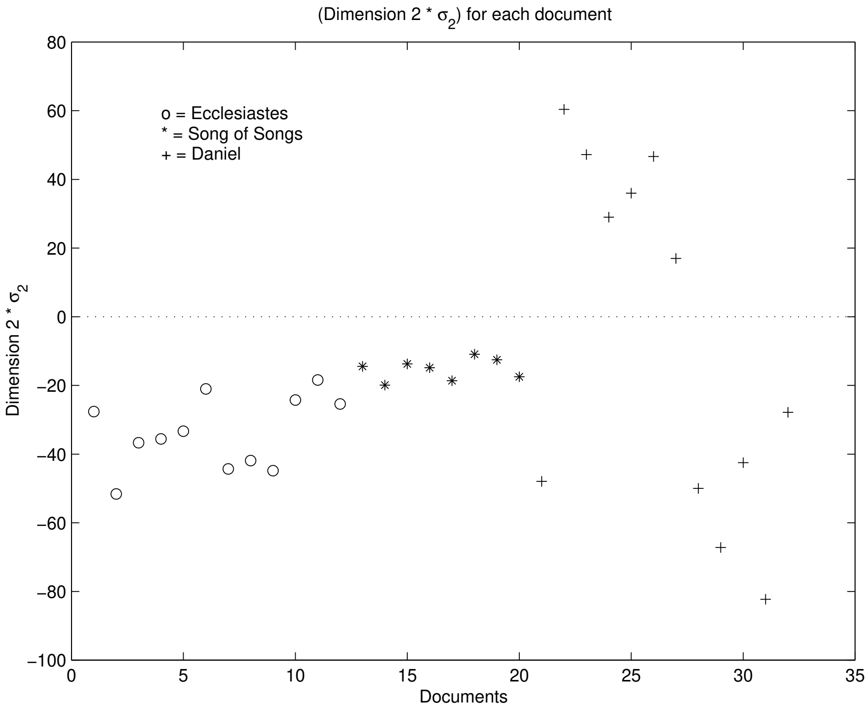

Figure 3: The second dimension values for each book.

attributes of the corpus. Moreover, latent semantic indexing produces several dimensions of possible interest, many of which can be explored simultaneously using visualization techniques.

One significant complication is that the technique is sensitive to length of the n-grams. Varying the length of n can change the LSI space considerably, because not only do the n-grams themselves carry more (and possibly different) information when they are larger, but there are simply more terms to work with for larger n. The choice of a "good" n depends on language and goals. It seems that smaller n are better at capturing style characteristics, because whole words only emerge statistically rather than as literal n-grams. Longer n-grams, especially in Hebrew, can hold one or even two words in some cases.

Understanding of the SVD-generated space is not yet sophisticated to the point that a program could automatically classify documents by writing style. We are able to decide which dimension delineates style because of external knowledge; the SVD creates the space only based on the n-grams from the documents. While algorithms for spatial clustering may be applicable, it is not clear how to algorithmically choose the dimension which highlights writing style. We hope to explore these issues within the scope of visualizing corpus changes over time.

References

[1] A.H. Barr. Superquadrics and angle-preserving transformations. *IEEE Computer Graphics and Applications*, 1(1):11–23, January 1981.

[2] Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, December 1995.

[3] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, September 1990.

[4] David S. Ebert, James M. Kukla, Christpher D. Shaw, Amen Zwa, Ian Soboroff, and D. Aaron Roberts. Automatic shape interpolation for glyph-based information visualization. In *IEEE Visualization 97 Late Breaking Hot Topics*, Phoenix, AZ, October 1997.

[5] Bradley Kjell and Ophir Frieder. Visualization of literary style. In *IEEE International Conference on Systems, Man, and Cybernetics*, October 1992.

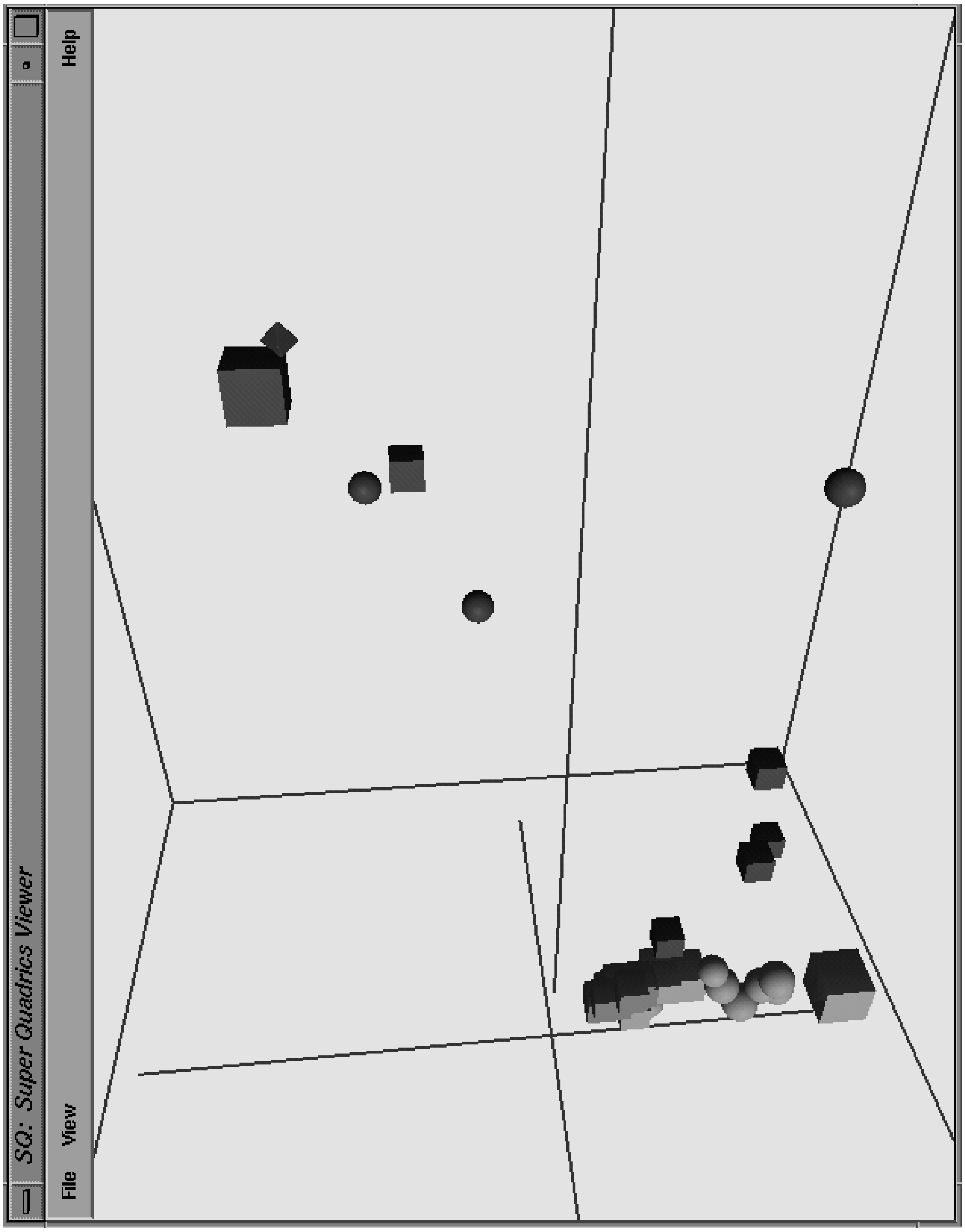

Figure 4: The Solomon/Daniel data set, in a multidimensional visualization which uses color, position, shape, and size to convey information. Here, the positional information in X and Y is the same as in Figure 1.

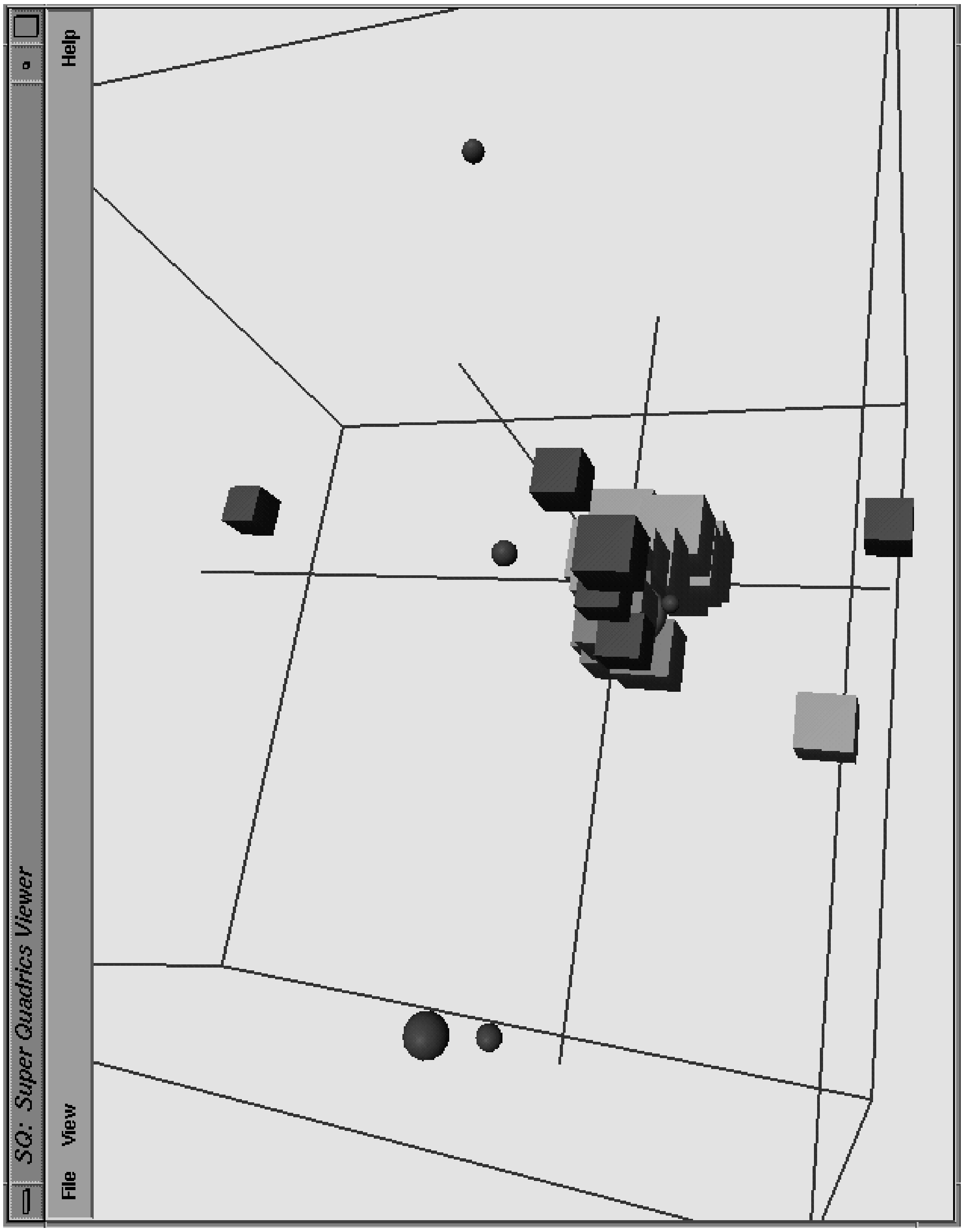

Figure 5: The Solomon/Daniel data set, with a different assignment of parameters.