Journal of Data Science **2**(2004), 87-105

# SEER: A Graphical Tool for Multidimensional and Categorical Data

Chris Chiu[1] and Ronald Fecso[2]
[1]*Law School Admission Council and* [2]*National Science Foundation*

*Abstract*: This paper introduces a visualization technique, SEER, developed for policy makers and researchers to graphically analyze and explore massive amounts of categorical data collected in longitudinal surveys. This technique (a) produces panels of graphs for multiple group analysis, where the groups do not have to be mutually exclusive, (b) profiles change patterns observed in longitudinal data, and (c) clusters data into groups to enable policy makers or researchers to observe the factors associated with the changing patterns. This paper also includes the hash function, of the SEER method, expressed in matrix notation for it to be implemented across computer packages. The SEER technique is illustrated by using a national survey, the Survey of Doctorate Recipients (SDR), administered by the National Science Foundation (NSF). Occupational changes and career paths for a panel sample of 14,901 doctorate recipients are profiled and discussed. Results indicated that doctorate recipients in some science and engineering fields are roughly two times more likely to work in an occupation when it is the discipline in which they received their doctorates.

*Key words:* Categorical data, data augmentation, graphics, hashing, longitudinal data, Survey of Doctorate Recipients (SDR), visualization.

## 1. Introduction

Data visualization is a type of graphical tools used for understanding abstract relationships among variables. It is frequently used to help interpret the meaning of data collected for scientific purposes in many disciplines such as biology, data mining, digital imaging, educational testing, finance, healthcare and medical research, market research, software engineering, statistics, telephone network analysis, cognitive psychology, and survey analysis. Graphical methods compensate for the limits of traditional statistical techniques by displaying massive data points in one or multiple graphs such that its global patterns can be comprehended while several levels of detail can be revealed (Tukey, 1993; Tufte, 1983; Wilkinson, 1999). Data visualization techniques can also be very useful for social and behavior sciences, particularly when categorical data have been collected

in longitudinal research (hereafter, multidimensional data). Recently, many researchers have developed techniques and frameworks for general graphical displays (e.g., Carey, 2002; Chen, 1999; Friendly, 1992; Ligges, 2002; McCulloch and Barnard, 2002; Narasimhan, 2002; Swayne and Lane, 2002; Unwin, 2002). However, displaying categorical data remains a challenge, especially for data with many categories. This is because, as is discussed in research by Blasius (1998), Hofmann (2000), and Friendly (1992, 2000, and 2002), each category in categorical data can be represented as a dimension and, frequently, high-dimensional data can be hard to depict on paper or on computer screens. In this paper, we first discuss the specific challenges faced by survey analysts in displaying multidimensional data. We then propose a method followed by an illustrated example using an occupational survey.

### 1.1 Longitudinal and categorical data

Longitudinal data refers to information collected over time. A special case of longitudinal data is panel data, which are obtained by surveying the same group of people over multiple occasions. The goal of panel data is to prevent historical events from introducing rival explanations to the survey data, thus establishing a stronger cause-effect relationship for phenomena of interest. The Survey of Doctorate Recipients (SDR) is one example of a panel study.

The SDR, managed by the National Science Foundation (NSF), is designed to "... provide demographic and career history information about individuals with doctoral degrees. The results of this survey are vital for educational planners within the Federal Government and in academia (NSF, 2003)." Also, employers in a variety of sectors (education, industry, and the government) may use the results to (a) understand and predict trends in employment opportunities and in salaries for doctorate holders in science and engineering, and (b) evaluate the effectiveness of equal opportunity efforts. Every two years, over 30,000 respondents are surveyed for the SDR. A considerable number of respondents are surveyed repeatedly (Chiu and Fecso, 2003). This type of panel data can answer questions of interest to policy makers in the federal government and academia such as: How do the occupation(s) of doctorate recipients change over time? To what extent are the fields of education (doctorate degrees) related to occupations? Such questions are difficult to answer using traditional numerical methods because of the excessive number of possible career paths one can follow (e.g., there are 10 billion possible career paths assuming a person is surveyed once a year for five years ($100^5$). Friendly (2002) provides a succinct summary of the milestones in the history of data visualization and points out that a milestone for this decade is the development t of methodologies for displaying high-dimensional and categorical data.

For ease of display, some methods are designed to reduce the number of dimensions in data (e.g., Li, 2002; Yin and Cook, 2002). These methods are most appropriate when a few major dimensions dominate the data. However, these techniques become less appropriate in instances where all of the dimensions are equally important or when every one of the dimensions has a substantive meaning. Policy makers, researchers, and analysts of longitudinal data often encounter nominal or categorical data (e.g. gender, occupational titles, discrete performance levels), which cannot be displayed by traditional time-series plots with one axis showing a time-dependent variable (e.g. month, quarter, and year) and a non-categorical, continuous, or quantitative variable on the other axis (e.g. dollars, weight, and height).

### 1.2 Occupational data

Visualization tools that can capture complex relationships for longitudinal and categorical data are particularly in demand, as described by Syverson (1996), who called for research methodologists to develop ways of tracking, visualizing, and interpreting panel data in particular, the job change patterns for doctorate recipients. Syverson stated:

> "If any of you have really clever ways to follow career paths, we would love to look at your survey, so that we can benefit from your experience ... how do you efficiently collect information on all of the possible things that people do in 10 years, we don't have a solution to this yet. If someone in this group has a clever way to do that, a matrix or a way to code this, we would love to see it and to use it."

In light of the development of a graphical method to facilitate data analysis for policy and decision making, we have developed a graphical method, SEER (SEE Repeated data), which:

1. is a process-oriented and interpretable tool for fitting and displaying categorical data with a time-dependent variable;
2. enables multiple-group comparisons but does not require the multiple groups to be mutually exclusive;
3. provide an efficient means for policy makers, analysts, and researchers to visualize the trends in longitudinal data measured at the nominal level or higher;
4. stresses unique combinations of category;
5. is designed to show the dynamics of changes that occur among categories in large complex data sets without requiring a colored plot;
6. does not require extensive knowledge in graphical techniques to understand the context, and;

7. is accompanied by matrix notations to enable its implementation to be software-independent.

Graphical representations have a number of advantages over categorical representations (tables). First, they provide a bird's eye view of occupational change over time for all job categories. Every pattern of occupation status is associated with a trend that can be viewed with little effort. Second, it allows cross-classifications – data points can be classified simultaneously in multiple categories. For example, data for survey respondents with multiple concurrent occupations (e.g., several part-time jobs) could be easily incorporated. When survey respondents report that they have more than one primary occupation, they are cross-classified and are listed as an employee in multiple occupations; occupational categories are not mutually exclusive. Third, the graphical representation can capture massive amounts of data all at once. Indeed, we experimented plotting 40,000 cases from a read data set using this approach. The resulting two-dimension plot displayed a clear pattern. In addition, using the zooming features (e.g., zoom in and out) available in many statistical packages (e.g., SAS and SPSS) and a dynamic look-up table (which associates each data point with additional information such as age), we were also able to apply the SEER method interactively.

In the following sections, we first discuss the conceptual framework of the SEER technique with an emphasis on the principles. Second, we provide an example of how the SEER technique can be used by applying it to a longitudinal data set with over three million sparsely filled cells (14,901 cases and 217 variables). Although this example is not extremely large in size, it illustrates the capacity of the SEER technique to display a large number of cases and dimensions. Last, we summarize the technique and discuss its applications in educational testing as well as in market research.

## 2. The SEER Technique

### 2.1 Single-case scenario

Assume that a longitudinal data set is analyzed and each case in the data set is measured or observed four times (e.g., four survey years). Further assume that the variable of interest is a categorical variable that can take on five values (e.g., 1: computer and mathematical sciences, 2: life sciences, 3: physical sciences, 4: social sciences, and 5: engineering). In a hypothetical case where a person is employed as a computer scientist in the first and last survey years but is employed as an engineer in the middle two years, the person would have a categorical vector of data reflecting the occupations reported in the four survey cycles: [1 5 5 1]. We

use a hash function (Maurer and Lewis, 1975; Knuth, 1968, 1969, 1974), Equation (2.1) below, to convert this vector into a binary vector with only zeros and ones. This binary vector is then plotted in a scatterplot using symbols that are easy for visualization — Cleveland (1993) studied the cognitive aspects of graphical displays and found that, when using scatterplots for categorical data, it is much easier to visually process some symbols (e.g., $+, \circ, >, w$) than others (e.g., easily confused letters such as H, F, T, E). To this end, we used 'o' to represent a data point in the scatterplot.

The variable $s$ in Equation (2.1) is a spacing factor, which spaces out each category so that the plot will not be too dense for visual interpretation and examination. For example, assuming $s = 3$, we can apply Equation (2.1) to the second data point (i.e., 5) of the vector [1 **5** 5 1]. Consequently, we can find that it is in the 30th position of a $1 \times 32$ vector because $f(5, 4, 3, 2) = 30$.

$$d = f(c, n, s, t) = (c - 1) \cdot (n + s) + t, \tag{2.1}$$

where $d$ is the index or cell position of a data point in the binary vector; $c$ is the value of a data point in the categorical vector; $n$ is the number of observations or the row size of the categorical matrix; $t$ is the order of an observation, $t = 1, 2, \ldots, n$; and $s$ is a spacing constant, that is, the number of spaces separating each group of nonempty cells in the binary vector.

### 2.2 Multiple-case scenario: Vector hash function

In practice, a data set frequently contains multiple cases as opposed to only one case. It is useful to represent Equation (2.1) in matrix notation, which can handle all observations simultaneously. To this end, we use a matrix $\mathbf{X}$ to represent a data set for the $n$ observations of the $N$ cases. Each element in matrix $\mathbf{X}$ can take on any discrete value between 1 and the maximum number of categories in the data and, for this reason, $\mathbf{X}$ is a categorical matrix. We begin with the following example. Let us assume we have a data matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 5 & 5 & 1 \\ 2 & 4 & 4 & 3 \\ 3 & 4 & 3 & 6 \end{pmatrix}$$

where we have $N = 3$ cases, each with $n = 4$ observations, and each observation has at most 6 possible values:

$$X_1 = (1, \mathbf{5}, 5, 1)$$
$$X_2 = (2, 4, 4, 3)$$
$$X_3 = (3, 4, 3, 6)$$

Now assume that we would insert three spaces to spread out the visual display, $s = 3$, and we would need a 39-bit row vector of 0/1 to represent each $X_i, i = 1, 2, \ldots, N$. The final 3 by 39 "binary matrix" is of the form

$$D_1 = (1001\ 0000\ 0000\ 0000\ 0\mathbf{1}10\ 0000)$$
$$D_2 = (0000\ 1000\ 0001\ 0110\ 0000\ 0000)$$
$$D_3 = (0000\ 0000\ 1010\ 0100\ 0000\ 0001)$$

Note that the bold faced data point 5 in $X_1$ corresponds to the bold-faced data point 1 in $D_1$. This is obvious because the value ($c = 5$) of the data point 5 suggests that the data point be stored in the fifth block of $D_1$ and the position of the data point ($t = 2$) suggests that a 1 be put in the second position of the fifth block. We can see that the second position of the fifth block is equivalent to the 30th position of $D_1$ by using Equation (2.1) as shown in section 2.1. Indeed, what we did in the conversion process is the correspondence

$$\mathbf{X} = (X_1, X_2, \ldots, X_N)^T \iff \mathbf{D} = (D_1, D_2, \ldots, D_N)^T$$

or, in a simpler form,

$$X_i \iff D_i,\ i = 1, 2, \ldots, N$$

treating $D_i$ as a 1 by 39 (= 6 categories $\times$4 observations +5 blocks $\times$3 spaces each block) row vector.

Unlike the single case scenario, the multiple-case scenario has more than one case and thus a hash function is also needed to determine the positions of data points along the $y$-axis. Specifically, each row in $\mathbf{X}$ and in $\mathbf{D}$ represents a row in a scatterplot and thus the vertical positions are determined by $Y_i = (1, 2, \ldots, N)^T$.

Conceptually, the above paragraphs show a procedure to help determine the position of a categorical data point in a two-dimensional space. With this procedure, we know that we can place $X_1 = (1, 5, 5, 1)$ and $X_2 = (2, 4, 4, 3)$ in a scatterplot at the following locations, where the $x$- and $y$-coordinates for the four observations in $X_1$ and $X_1$ are, respectively, $[(1, 1), (1, 30), (1, 31), (1, 4)]$ and $[(2, 8), (2, 23), (2, 24), (2, 18)]$. Essentially, what we have accomplished above is to convert the categorical matrix into a binary sparse matrix (Gilbert, Moler, and Schreiber, 1992). The Appendix shows the corresponding matrix notation to implement the conversion process in one step, given a categorical data matrix $\mathbf{X}$.

## 3. Applications: Large-scale survey analysis

### 3.1 Overview and survey respondents

A panel sample of 14,901 advanced degree holders was obtained from the longitudinal survey, Survey of Doctorate Recipients (SDR). The survey was administered biennially. All respondents in the selected sample (a) were under the age 76 in 1999, (b) received at least one research doctorate in science or engineering from a U.S. institution in or prior to 1990, and (c) were residing in the United States on April 15 in four survey years analyzed in the current study (1993, 1995, 1997, and 1999); and (d) were employed in at least one of the four survey years. The panel of 14,901 respondents represented a population of over half a million doctoral degree holders in the United States. For employment research purposes, the occupational titles of the SDR respondents were recorded every two years using a standardized list of approximately 126 occupations. The occupational titles were coded as a categorical variable (e.g., "052" represents Computer System Analysts). These occupations were grouped into six major categories, namely computer and mathematical sciences, life and related sciences, physical and related sciences, social and related sciences, engineering, and non science and engineering. Using the SEER technique, we addressed two substantive research questions.

(1) What is the relationship between educational fields and occupations, or to what extent do Ph.D.s work in disciplines in which they received their doctorates?
(2) What are the job switching patterns for Ph.D.s who were employed in the computer and mathematical sciences?

Having applied the SEER display technique, we created 56 displays to address the above questions. Figure 1 provides a top-down view of the organization of the 56 displays, which are organized into four sets. The first and second sets or examples (a and b) show the unsorted and the sorted SEER displays respectively (the order of sorting is annotated in the figure). Both examples depict all 14,901 respondents employed in the six occupations. The unsorted display (example a) offers a global view of the data whereas the sorted display (example b) provides a more in-depth understanding of the data — it organizes the respondents of the same education field into a group. Also, it further investigates reasons that the respondents did not report their occupations. This is accomplished by breaking down the "logical skip" category into three groups, those who (a) did not respond to the survey, (b) were promoted to become managers, and (c) were retired. Within each of the nine categories, we assigned a mark in an occupation if a respondent was employed in that occupation. The position of the mark corresponds to the year in which the employment occurred. The mark for the most recent employment (1999) is placed in the far left position and the least recent (1993) in the far right position. Also, to enhance the display, we inserted 20 spaces in between two categories. As a result, we obtain a sparsely-filled binary

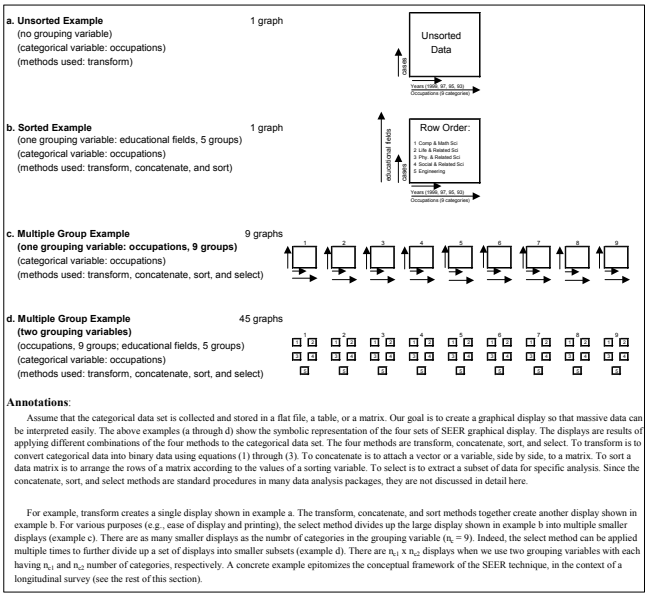


Figure 1: A road map of the SEER displays for career paths of doctorate recipients in science and engineering

matrix depicted in example b. It has a size $14,901 \times 196$, because there were 14,901 cases and 196 binary variables (i.e., 4 years $\times$9 categories +20 spaces $\times$8 blocks of spaces).

### 3.2 Occupations and education: An overview

In Figure 1 (example a), the order of the rows in the binary matrix is intrinsic to the order by which the respondents were surveyed. Thus, it does not show any distinctive patterns. This limitation is improved in example b, where the rows are grouped by educational fields in ascending order, from top to bottom. Figure 2 below shows the SEER display with operational data corresponding to example b in Figure 1. As shown, doctorate recipients with a degree in computer and mathematical sciences are the first group (positioned in the top portion) of the binary matrix; those graduated from life and related sciences come second; physical and related sciences graduates are next, followed by graduates of social and related sciences. Engineering graduates are positioned at the bottom of the graph.

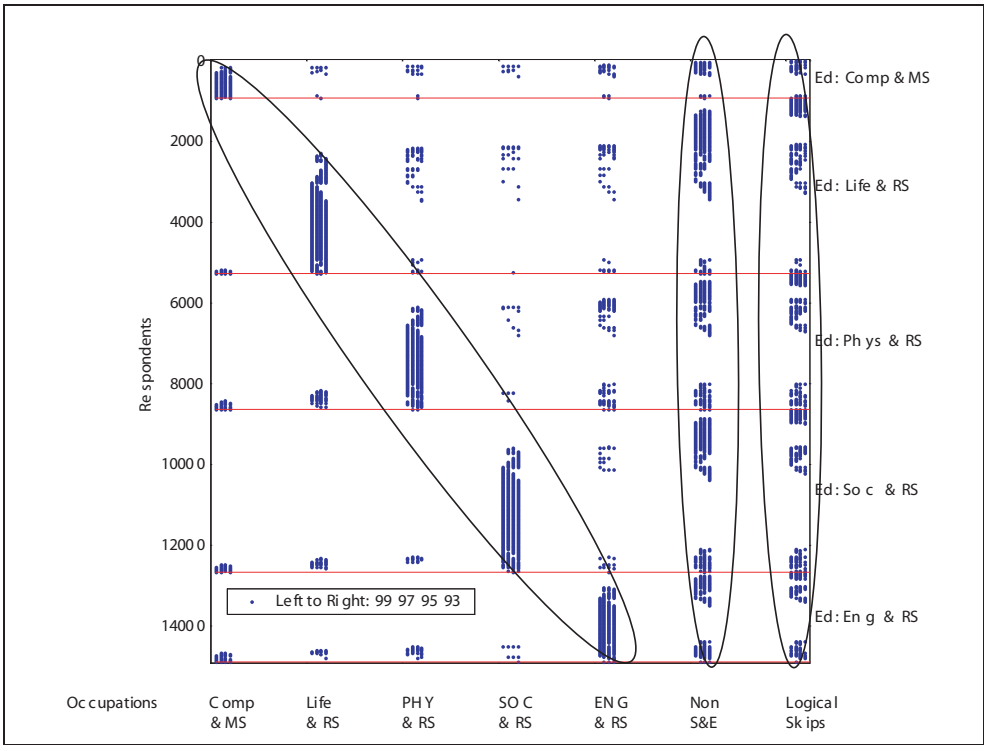


Figure 2: Relationship between occupations and educational fields

Three obvious patterns emerge in the Figure 2. First, a diagonally dominant pattern suggests that many graduates from each of the five S&E disciplines have been employed in occupations related to their doctoral field for multiple years, while the off-diagonal points indicate the cases where respondents report switching to other occupational disciplines. Second, the dense column labeled "non S&E" on the right side of the figure suggests that a noticeable number respondents with of S&E doctorates have worked in what NSF classifies as non-science and engineering disciplines for a good portion of their career — assuming that one biennial report reflects a stable occupation until the next report. Third, the dense column labeled "logical skips" indicates that every one of the five educational fields includes doctorate recipients whose occupations were not recorded due to the noticeable logical skips (see section 3.1 for definitions). Since the number of logical skips was noticeable, we further investigated why respondents did not report their occupations. This was accomplished by separating those who skipped with known reasons (NSF did not require retirees and managers to report their occupations and hence they skipped the occupational question) from those who skipped for unknown reasons. As a result of this separation, we added two categories "retirees" and "managers" to the subsequent plots.

### 3.3 Computer and mathematical scientists: Occupations

To further our understanding of occupational change patterns, we used a sorting scheme to aid the interpretation of data. Below, we first describe the sorting scheme and discuss the visual display of occupational change patterns for doctorate recipients in computer and mathematical sciences.

A set of binary data with $n$ observations has $2^n - 1$ unique switching patterns (e.g., four measurements yields 15 patterns, because $2^4 - 1$ equals 15). We assigned each one of the patterns a numerical value from 1 through 15 to create a sorting scheme, which was subsequently used to sort the rows of the data matrix in ascending order. Since each profession was surveyed four times in an eight-year interval, we used a sorting mechanism to identify those who switch, enter, exit, or return to an occupation. To accomplish this goal, we classified the switching pattern to correspond with the four types of occupational change as follows:

1. Switch: respondents who were in-and-out of an occupation and were employed elsewhere in the last survey, 1999;

2. Exit: respondents who left an occupation and never returned;

3. Enter: respondents who stayed in an occupation once they were employed in the occupation;

4. Return: respondents who were ever employed in an occupation, left, returned, and stayed in the occupation till the end of the longitudinal survey years.

We assigned the 15 unique switching patterns (see table below) to one of the four groups described above. These corresponding data are displayed in Figure 3.

While Figure 2 provides an overview of the relationship between educational fields and occupation, it does not show the dynamics (influx and outflow) of occupational changes, which are frequently of interest to policy makers. To do this, we created nine graphs to capture career paths. Specifically, we plotted a graph for each category of the occupations (i.e., six for the S&E and non S&E occupations, one for logical skips, one for managers, and one for retirees). Because of space limitations, we show only the career patterns of computer and mathematical scientists (the first discipline). In the grand scheme of plots, Figure 3 corresponds to the first plot of example c in Figure 1.

| Acronyms | | Switching Patterns 1999 | 1997 | 1995 | 1993 | Sorting key (v) |
|---|---|---|---|---|---|---|
| | Switch | | | 1 | | 1 |
| S | Switch | | 1 | | | 2 |
| | Switch | | 1 | | 1 | 3 |
| | Switch | | 1 | 1 | | 4 |
| | Exit | | | | 1 | 5 |
| E | Exit | | | 1 | 1 | 6 |
| | Exit | | 1 | 1 | 1 | 7 |
| | Enter | 1 | | | | 8 |
| E | Enter | 1 | 1 | | | 9 |
| | Enter | 1 | 1 | 1 | | 10 |
| | Enter | 1 | 1 | 1 | 1 | 11 |
| | Return | 1 | | 1 | | 12 |
| R | Return | 1 | | | 1 | 13 |
| | Return | 1 | | 1 | 1 | 14 |
| | Return | 1 | 1 | | 1 | 15 |

Note: Each row can be considered as a row vector with four cells. To contrast the visual effect, we did not place zeros in the table. However, one can fill in the empty cells by zeros.

The above SEER plot indicates that the Survey of Doctorate Recipients (SDR) sample has approximately 1,500 doctorate recipients who were ever employed in the computer and mathematical sciences discipline between the 1993 and 1999 survey years. These Ph.D.s were classified into four groups (switch, exit, enter, and return) based on changes in their job titles, major responsibilities, and employers. The SEER display also depicts the extent to which any particular occupational grouping attracts or supplies professionals to computer and mathematical sciences occupations. For example, looking across the SEER plot for the group classified as "enter," the relatively dense clusters tend to associate with engineering, non S&E disciplines (physical and related sciences), and managerial positions. Conversely, relatively few respondents entered the computer and mathematical sciences profession from life and related sciences, nor from social

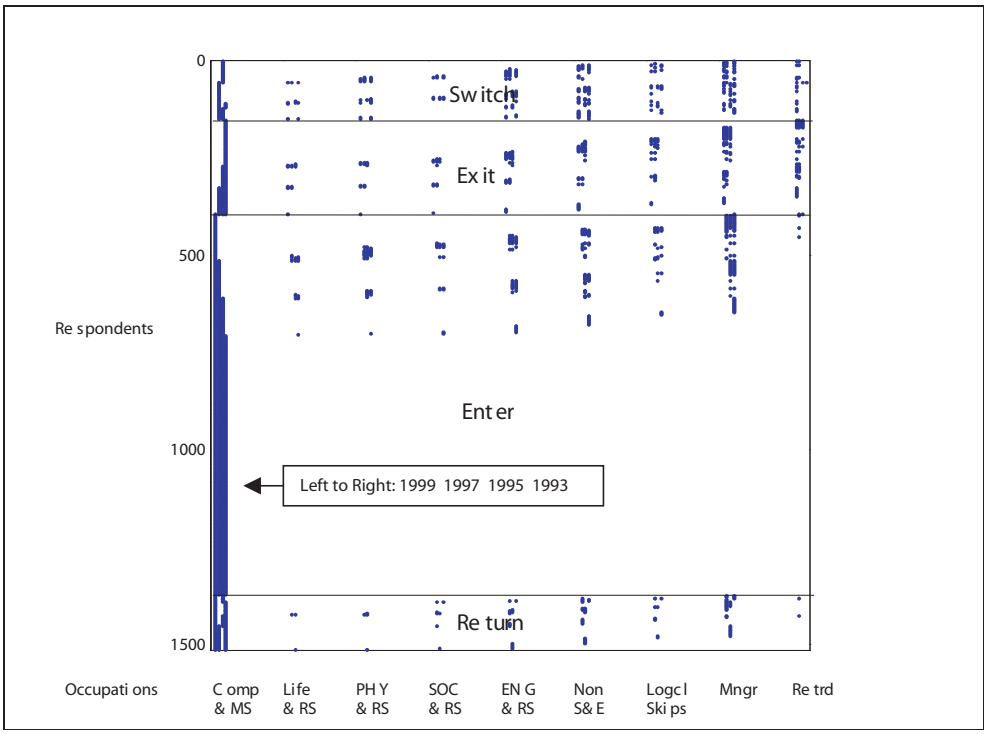


Figure 3: Occupational change patterns for computer and mathematical scientists

and related sciences. In addition, as expected, it is rare for a nonworking retiree to reenter (denoted "enter" in the plot) the workforce in computer and mathematical sciences (see the last column block in Figure 3). Also, looking across the group "exit," a noticeable number of respondents either retired or became managers, which, according to the SDR classification, include individuals managing in a computer and mathematics setting.

### 3.4 Computer and mathematical scientists: Education and occupation

What is the relationship between education and occupations for computer and mathematical scientists? Using a second grouping variable, "educational fields" (the first is "occupation"), we further divided up the cases displayed in Figure 3 into five smaller graphs. Each of the resulting five figures contains data for respondents holding doctorates in a specific field (see Figure 4). For example, Figure 4 shows the career paths for computer and mathematical scientists who held doctorates in computer and mathematical sciences; whereas Figure 5 shows the career paths for computer and mathematical scientists with doctorates in

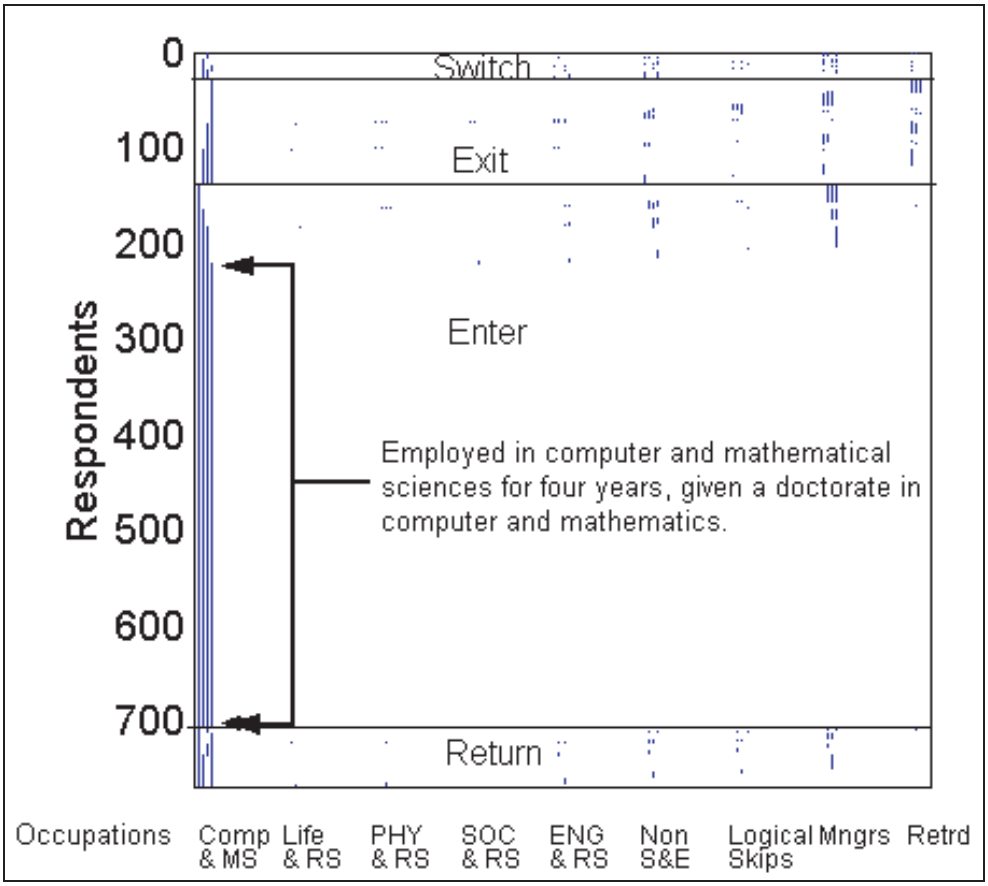


Figure 4: Same occupation and education (approximately 800 respondents)

engineering and related sciences. As indicated by the callout labels in Figure 4 and Figure 5, those with a doctorate in computer and mathematical sciences are much more likely to stay in computer and mathematical sciences for all four survey years. For example, over 60% of doctorates in computer and mathematical sciences stayed in that field, whereas only about 25% of engineering doctorate stayed. This finding indicates a positive relationship between education and the workforce. From a methodological viewpoint, the findings in the two figures (4 and 5) illustrate how a large SEER plot (Figure 2) in conjunction with smaller plots (Figure 4 and Figure 5) to view relationships in categorical data with many dimensions.

As discussed earlier, it is impossible to see these relationships by examining only the overall SEER plot (Figure 2) or the occupational display (Figure 3).

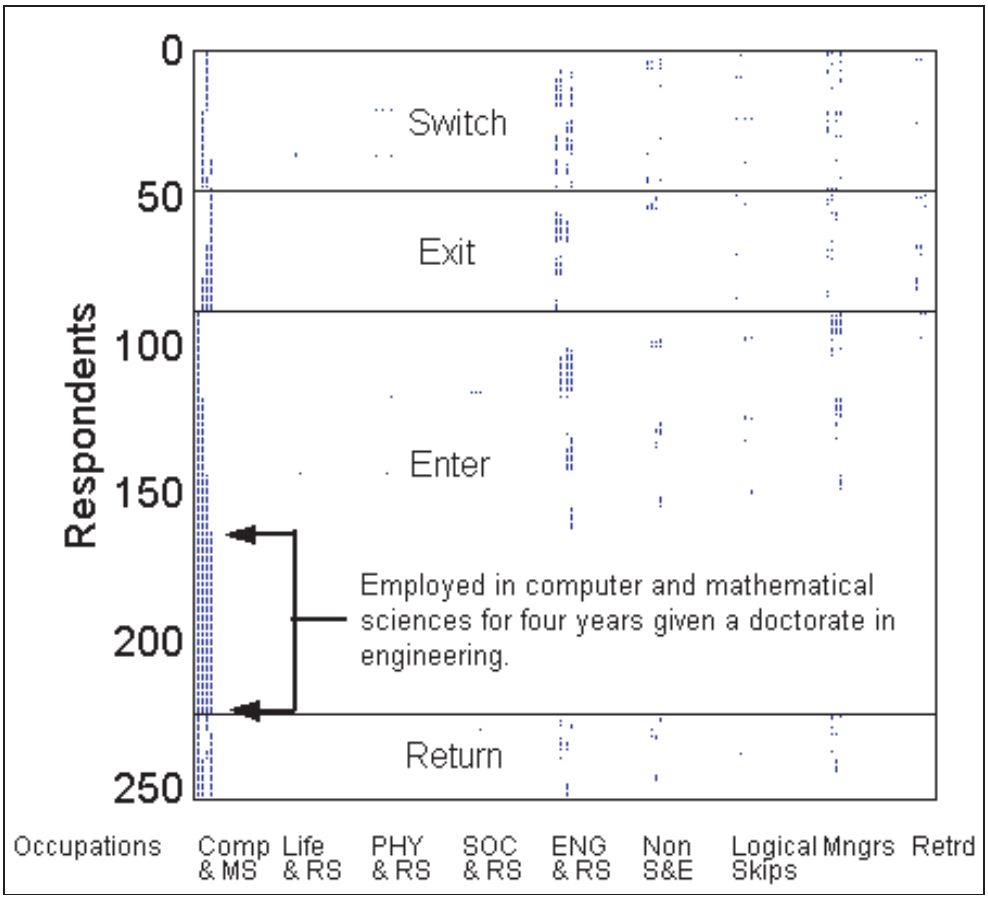


Figure 5: Occupation different from education (approximately 250 respondents)

## 4. Conclusion and Discussion

Frequently, researchers in social, behavioral, economical, psychometrical, and marketing sciences need a tool for exploring the patterns in longitudinal survey data with nominal scales. To this end, we developed the SEER technique, which can display massive amounts of categorical data all at once. The technique is equipped with a feature to create a panel of visual displays suitable for presenting data with a large number of categories (over 100 categories). This feature is especially useful when one graph is not capable of displaying all variables. In addition, the SEER technique can summarize cross-classified longitudinal data (i.e., a data point has multiple group memberships such as occupations).

Having applied the SEER method in a national sample, we visually found that (a) a considerable number of doctorate recipients in science and engineering tend to work in their degree field (see Figure 2), and (b) in some disciplines, such as computer and mathematical sciences, doctorate recipients are much more likely (approximately two times as likely) to work in their discipline when their degree field matches their occupation (see Figure 4).

Designed as a companion for numerical methods, the SEER technique empowers researchers to visually explore the data that can be clustered in a meaningful way to address specific research interests. Indeed, it is flexible enough that one can implement the method as a Multidimensional Online Analytical Processing tool (MOLAP) for data explorations. In this paper, we have described the technique with matrix notations and illustrated the features of the SEER technique through a concrete example. This example used data collected for a longitudinal survey — the Survey of Doctorate Recipients (SDR)— administered by the National Science Foundation (NSF). In order to limit the size of our chapter, we presented only the four most important displays of the 56 SEER plots to examine doctorate recipients' career paths and their transitions between education and the workforce.

Despite the fact our example used survey data to demonstrate the SEER technique, one can apply the technique to virtually any categorical data with repeat measures. For instance, it can be used (a) in market research to determine customers' buying preferences, over time, among different products, and (b) in national testing programs such as the National Educational Assessment of Progress (NAEP or the national report cards) to understand students' performance and to provide diagnostic information for elementary and high school students, and last but not least (c) in high-stake examinations such as the Law School Admission Test (LSAT) to detect irregular test scores for repeat test takers (Chiu and Fecso, 2002). We can provide only a succinct outline in this paper; we cannot describe each and every one of the applications due to space limits. In addition to categorical data, the SEER method can also be applied to display continuous data such as test scores. This is feasible because one can consider continuous data as a special case of categorical data when continuous data are divided into fine intervals.

The question arises regarding overplotting or screen resolution as a possible barrier to the SEER method, because one may question how the SEER method shows, for example, 14,901 rows on a screen with less than 1,000 rows of pixels. We view this as more of a computer hardware challenge, which can be overcome by subdividing a large plot into smaller ones. Indeed, we have shown in the current study how one can use smaller SEER plots as supplements to a large SEER plot when finding and verifying associations among categorical data (see

Section 3).

In this study, we discussed issues related to displaying categorical data in longitudinal surveys. Frequently, it is a challenge to present high-dimensional categorical data in a systematic way while keeping all the desired properties discussed in section 1.2. Just like one might perceive that Trellis plots (Becker, *et al.*, 1996) to be a special case of Mosaic plots for categorical data (Friendly, 1999; Hofmann, 2000), we can probably view the SEER method as a special case of other graphical methods and vice versa. The major contribution of the current study is that we developed a one-step function to convert a large number of data points from a multidimensional space to a two-dimensional space. The conversion process is independent of the size of the data set and thus is efficient.

Graphical displays or visualization techniques are companions for numerical and statistical methods. Future research should focus on the connections between the SEER technique and other numerical methods such as survival functions (Ureta, 1992), stability analysis and generalizability theory (Brennan, 2001; Chiu, 2001), and logistic regression (Hosmer and Lemeshow, 2000), to name but a few. Indeed, when combined with the generalizability theory (Brennan, 2001; Cronbach *et al.*, 1972), the SEER technique provides a powerful methodology for quickly grasping complex interrelationships, while capturing the exact information that is more precisely presented in numerical methods. One example of such a combination is the detection of errors when human judgments are involved in survey analysis or educational testing (e.g., computer aided telephone interviews and human scoring essays in large-scale testing programs).

## Acknowledgement

The authors thank the editor, three anonymous reviewers, and Susan Dalessandro for their positive comments and careful reading. The research in this article does not represent the official positions of the funding organizations.

## References

Becker, R.A., Clevland, W. S., and M. Shyu (1996). The visual design and control of trellis display. *Journal of Computational and Graphical Statistics* **5**, 123-155.

Blasius, M. G. (1998). *Visualization of categorical data*. Academic Press.

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Carey, V. (2002). R graphics: Overview and desiderata. Paper presented at the annual conference of the 2002 Joint Statistical Meeting (American Statistical Association), Aug 10-15, 2002. New York.

Chen, C. H. (1999). Extensions of generalized association plots (GAP). Paper presented to the 1999 Annual Meeting of the American Statistical Association, August 8-12, Baltimore, MD.

Chiu, C. W. T. (2001). *Scoring performance assessments based on human judgments: Generalizability theory*. Kluwer Academic Publisher.

Chiu, C. W. T. (2002). Plotting multidimensional data onto a two dimensional display: A graphical method. Paper presented to the 2002 Annual Meeting of the Psychometric Society, June 20. Chapel Hill, NC.

Chiu, C. W. T. (2003a). Visualizing disagreements in expert rating for evaluations and performance assessments: A graphical method (SEER) Paper presented to the 2003 Annual Meeting of the American Statistical Association, Aug 2-7, San Francisco, CA.

Chiu, C. W. T. (2003b). A graphical model for comparing test equating methods. Paper presented at the 2003 Annual Meeting of the Psychometric Society, July 7-10. Sardinia, Italy.

Chiu, C. W. T. and Fecso, S. R. (2002). Visualizing and mining repeated measure data using generalizability theory: Capturing career paths and test scores. Paper presented to the 2002 Annual Meeting of the American Statistical Association, Aug 11-14. NY.

Chiu, C. W. T. and Fecso, S. R. (2003). Incorporating sampling weights into generalizability theory for large-scale analyses. *Journal of Modern Applied Statistical Methods* **1**, 108-127.

Cleveland, W. (1993). A model for studying display methods of statistical graphics. *Journal of Computational and Graphical Statistics* **2**, 323-343.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley.

Friendly, M. (1992). Graphical methods for categorical data. Proceedings of the SAS User's Group International Conference, 17, 1367-1373.

Friendly, M. (1999) Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics* **8**, 373-395.

Friendly, M. (2000). *Visualizing Categorical Data*. SAS Institute.

Friendly, M. (2002). Milestones in data visualization. Poster presented to the 2002 Annual Meeting of the American Statistical Association, August 11, New York.

Gilbert, J., Moler, C., and Schreiber, R. (1992). Sparse matrices in MATLAB: Design and implementation. *SIAM Journal on Matrix Application* **13**, 333-356.

Hofmann, H. (2000). Exploring categorical data: interactive mosaic plots. *Metrika* **51**, 11-26.

Hosmer, D. W. J., and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley.

Knuth, D. E. (1968). *Fundamental Algorithms*. Addison-Wesley.

Knuth, D. E. (1969), *Seminumerical Algorithms*. Addison-Wesley.

Knuth, D. E. (1974). Computer programming as an art. *Communications of the ACM*, **17**, 667-673.

Li, B. (2002). Statistical inference for the central mean space. Paper presented at the annual conference of the 2002 Joint Statistical Meeting (American Statistical Association), Aug 10-15, 2002. New York.

Ligges, U. (2002). R Graphics. Paper presented at the annual conference of the 2002 Joint Statistical Meeting (American Statistical Association), Aug 10-15, 2002. New York, NY.

Maurer, W. D., and Lewis, T. G. (1975). Hash table methods. *ACM Computing Surveys*, **7**, 5-20.

McCuloch, R., and Barnard, J. (2002). Object oriented and graphical methods for Bayesian models. Paper presented at the annual conference of the 2002 Joint Statistical Meeting (American Statistical Association), Aug 10-15, 2002. New York.

Narasimhan, B. (2002). Scalable vector graphics. Paper presented at the annual conference of the 2002 Joint Statistical Meeting (American Statistical Association), Aug 10-15, 2002. New York.

National Science Foundation (2003). Descriptions of the Survey of Doctorate Recipients (SDR). [Online]: http://www.nsf.gov/sbe/srs/ssdr/start.htm

Syverson, P. (1996). Data needs for assessing the relationship between graduate education and evolving trends in the S&E labor market. Unpublished memo. National Science Foundation.

Swayne, D., and Lane, D. T. (2002). Extensible statistical graphics with GGobi and R. Paper presented at the annual conference of the 2002 Joint Statistical Meeting (American Statistical Association), Aug 10-15, 2002. New York.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information.* Graphics Press.

Tukey, J. W. (1993). Graphic comparisons of several linked aspects: alternatives and suggested principals. *Journal of Computational and Graphical Statistics* **2**, 1-33.

Ureta, M. (1992). The importance of lifetime jobs in the U.S. economy, revisited. *American Economic Review* **82**, 322-35.

Unwin, A. (2002). Interactive interfaces for statistical software – the Augsburg Impressionist projects. Paper presented at the annual conference of the 2002 Joint Statistical Meeting (American Statistical Association), Aug 10-15, 2002. New York.

Wilkinson, L. (1999) *The Grammar of Graphics*. Springer-Verlag.

Yin, X., and Cook, D. (2002). Information extraction: A dimension reduction technique based on information theory. Paper presented at the annual conference of the 2002 Joint Statistical Meeting (American Statistical Association), Aug 10-15, 2002. New York..

## Appendix: Matrix Implementation of the Hash Function for Multiple Cases

Equation (2.1) in section 2.1 shows the mathematical form of the hash function, for a single case, used in the SEER method. Frequently, researchers have data from multiple cases and they may wish to have a computational form for implementing the method in one step, given the categorical data matrix $\mathbf{X}$. For this purpose, we present the matrix notation in equation (A.1), which determines the horizontal location ($x$-coordinates) of all data points in $\mathbf{X}$. In general, the equation has a linear form $\mathbf{o}_x = k\mathbf{x} + \mathbf{b}$ where $\mathbf{x}$ and $\mathbf{b}$ are vectors for intermediate computations and $k$ is a scaling constant (i.e., $x$ is a vectorized form of the data matrix $\mathbf{X}$ and $\mathbf{b}$ is a shifting vector). More generally, $k\mathbf{x}$ represents a between-category shifting factor and $\mathbf{b}$ a within-category shifting factor for the location of the categorical data.

$$\mathbf{o}_x = f_x(\mathbf{X}, N, n, s) = (n+s)\cdot([\mathbf{x}_1^{'}, \mathbf{x}_2^{'}, \ldots, \mathbf{x}_n^{'},] - 1) + [\mathbf{1}_N^{'}, \mathbf{2}_N^{'}, \ldots, \mathbf{n}_N^{'},] \quad \text{(A.1)}$$

where $\mathbf{o}_x$ is an $nN \times 1$ vector of indicies, repressenting the hroizontal locations; $\mathbf{X}$ is a $N \times n$ matrix, representing all observations for all data points in a categorical matrix; $\mathbf{x}_j$ is a $N \times 1$ vector, representing the data point for the $j$-th observation ($j = 1, 2, \ldots, N$); and $\mathbf{k}_N$ is an $N \times 1$ vector with $k$ in each cell ($k = 1, 2, \ldots, n$).

Equation (A.2) shows the vectorized hash function for the $y$-coordinates.

$$\mathbf{o}_y = f_y(\mathbf{X}, N, n) = \mathbf{1}_n \otimes [\mathbf{1}_N] \quad \text{(A.2)}$$

where $\mathbf{o}_y$ is a $nN \times 1$ dummy vector, representing the $y$-coordinate of aqll measurements for all data points; $\mathbf{1}_n$ is an $n \times 1$ vector with 1 in each cell; and $[\mathbf{1}]_N$ is an $N \times 1$ vector with a set of consecutive numbers: $1, 2, \ldots, N$, e.g., $[\mathbf{1}]_3 = (1, 2, 3)^T$; and $\otimes$ is the Kronecker product.

Received June 20, 2002; accepted March 23, 2003

Chris Chiu
Research Scientist
Test Development and Research
Law School Admission Council
Newtown, PA 18940, U.S.A.
chiuwing@msu.edu

Ron Fecso
Chief Statistician
Division of Science Resources Statistics
National Science Foundation
4201 Wilson Blvd., Suite 965
Arlington, VA 22230, U.S.A.