Survey Visualization

Maria Tkatchenko
Master's Student
University of British Columbia

1-604-822-1290

tkatch@cs.ubc.ca

ABSTRACT

This paper addresses the issue of presenting the survey results in a way that would allow them to be interpreted in their original context. I present a tool that displays both the administrative data associated with the survey and the results obtained from respondents, using a variety of interactive displays linked together to create a sense of continuity during exploration. I also present an implementation of an innovative display for categorical data, which is based on a scatterplot.

Several different models of exploration are described, and some real results based on experimenting with the tool and a real data set are presented. Experimentation also allowed me to find other potential uses for this tool. Due to the time constraints placed on this project, much work remains to be done, and so a number of extensions and improvements to the tool are also proposed.

1. INTRODUCTION

Multidimensional data comes in many different flavours, some of which have been explored more than others. Survey results are a form of multidimensional data, but one that seems to have been overlooked by the information visualization community. Looking at the numerical results is often not enough when these numbers are mapped from multiple-choice answers to the survey questions. This is because most researchers are not interested in abstract numerical relationships, but need the context of *what* the question was asking for the analysis to be effective.

The easiest way to gather data from a group of people is to administer surveys individually. This is often accomplished through either in-person or telephone interviews, although on-line surveys have been gaining popularity in recent years. This expansion can be explained by the fact that as Internet access becomes available to more people, the medium of the survey is less likely to skew the results.

Surveys allow one to discern the respondents' opinions on a variety of issues that are of interest to the administrators of the survey. They are administered as a sequence of questions, each with a small range of admissible answers. This highly structured format allows easy analysis, and, more importantly, the raw data can be processed mostly automatically due to its structure. However, looking at the raw data is quite uninformative, as the analyst can only gain insight into a single person's opinions, or the group's responses to a single question. The analysis of trends and outliers is much more interesting. It applies not only to the (relatively) small group of respondents, but, if the sample is suitably random, allows you to make inferences about the general population. Many statistical methods exist to determine both the number and characteristics of respondents necessary to be able to draw generalized conclusions with a given degree of certainty.

Political and economic issues are just some areas where surveys have been used extensively over the past few decades. As political competitions have intensified and politicians have become more worried about the voters' perception of them, surveys have become a very important vehicle to discerning public opinion. Politics is only one of the areas where survey administration and analysis are highly important. Economic decisions, health care policies, even TV network programming are often heavily influenced by the respondents' answers to the carefully-worded questions on the surveys. Writing surveys is itself an art, but in this paper I will concentrate on the kinds of answers that can be gleaned from detailed exploration of the responses to the survey.

Correlation is a commonly used statistical measure to calculate how "close" two sets of points are to each other [4]. Correlation coefficients range from –1 to 1, with negative numbers denoting a negative relationship, where an increase in one variable causes a decrease in the other. Under positive correlation, the results for both variables change (increase or decrease) in the same direction. A correlation around 0 means that there is absolutely no discernible (statistically significant) relationship between the two variables – i.e. the results for one cannot be used to predict the other. A correlation of 1 (positive or negative) signifies perfect correspondence, where for every change in one variable, the other changes proportionately. Weatherburn even goes on to say "the magnitude of r [the correlation coefficient] may be taken as a measure of the degree to which the association between the variables approaches a linear functional relationship"[17].

Other statistical measures are often used to calculate the pair-wise correlation and higher-dimensional regression between a number of variables. The latter especially requires tedious calculations, which are made easier by batch statistical analysis toolkits. However, while these tools are able to perform the calculations flawlessly, I haven't found many that can efficiently organize all of this correlation data for overview by the user. The results for the computations are usually displayed separately, or at best aggregated into a table for comparison, requiring the user to go back to the raw data and perform further calculations if a relationship of interest is found.

This work assumes that what the users are trying to accomplish is investigate trends between the responses to survey questions. The user may want to find sets of questions that are highly correlated, repeating this for sets containing anywhere from 2 to n questions (n being the total number of questions in the survey). If a highly correlated set of many questions is found, that means that a strong relationship exists between those questions. This tool currently only implements the case with sets of 2 questions, for technical reasons whose discussion is delayed until Section 8.

This paper is structured as follows. Section 2 presents a more detailed description of the problem. Section 3 explores some of the related work in the field of multidimensional visualization. Sections 4 and 5 present a description of the solution, first at a more abstract level, and then with a description of the specific implementation tools and techniques used. Sections 6 and 7 first present a description of two exploration scenarios, and then present real results obtained from applying these techniques. Section 8 explores some of the challenges encountered during this project, and Section 9 suggests future work. Section 10 concludes with a brief summary. All the figures referred to throughout the text are in Section 12.

2. DESCRIPTION OF PROBLEM

Surveys range over a number of diverse topics, from political opinions to shopping preferences. As diverse as the subjects that the surveys address are, there are a number features that are common among them. Most surveys involve a questionnaire with dozens or hundreds of questions, where the answers to each question are usually restricted to a set of 2-6 possibilities, with only a small number of free-form answers, if any. The number of respondents may range from a few dozen to tens and hundreds of thousands, and strict organization of data is absolutely necessary for the resulting data sets to be useable.

2.1 Task

Humans often fail when presented with a large set of data in many variables, and faced with analyzing the data to discover trends or outliers. Multiple views are often required in order to discover correlations as well as keep track of relationships between different dimensions of data. To make the situation even more complicated, the data sets can be categorized both in terms of the number of questions and the number of individual respondents.

Once you look at the format of the survey results, it becomes evident why better visualization techniques are needed. A sample is shown in Figure 6. Currently, Excel and mathematical/statistical packages such as MATLAB are used to organize and sort the data. The basic structure of a raw result data set is a matrix of questions by respondents. For correlations, statistical analysis packages return a similar matrix of questions by questions, with the corresponding correlation shown in each cell. Given such structure, it seems that even experts would do better with at least some basic visualization capabilities, although I haven't been able to find any research that specifically substantiates that claim.

The main task this tool aims to support is the exploration of relationships, and in particular the degree of correlation, between the various questions on a survey. The task requires the understanding of each question and its raw data set separately, as well as in the larger context of its relationships with some (or all) of the other questions. The goal of the proposed tool is to provide visual cues to higher-level trends within the data, while allowing the user to control the search through a number of independent criteria. It should also allow the user to bring together all the information relevant to survey exploration, such as the background information on the survey, the phrasing of the actual questions and answer choices, as well as the gathered data.

Depending on the particular task they are faced with, users concentrate on looking for either outliers or general trends within the data. I propose to omit the question of outliers for the time being, and concentrate on helping the user discover the various higher-level trends.

2.2 Data set

The data I would like to explore with this tool comes from the pre- and post-election questionnaires, available from the National Election Studies website [9]. The NES provides, in separate files, such information as:

- textual description of the survey, including any relevant information on how it was conducted, etc.
- text for all the questions on the survey
- listing and description of all the variables and the possible values they may take
- raw data, containing both the meta-data for each respondent as well as their answer to each question.

The NES website releases, for most years, both pre- and post-election surveys. The datasets available cover the period from 1956 to 2002, in two-year intervals. For some years, they claim that nearly 75% of the material on the pre- and post-questionnaires is in common. It would be quite useful to be able to explore the responses to the same question in both pre- and post-election survey for a given year. However, I have not been able to determine whether there is an easy way to automatically match the corresponding questions. This is the case where, if the tool were to allow such comparisons, expert input (such as from the creator of the survey) would be needed to set up equivalencies between questions before analysis could begin.

Another option is the Behavioral Risk Factor Surveillance System from the National Center for Chronic Disease Prevention and Health Promotion [8]. These contain results of annual surveys administered from 1984 to 2003. This program is run on a statewide basis in order to obtain information about behavioural risks. Such information allows better planning of health promotion and disease prevention programs. The program started out with 15 states, but by 1994 all the states participated.

The questionnaires in both databases have on the order of a couple hundred questions. The NES data sets have a few thousand respondents, while the BRFSS data sets have on the order of 100,000. The one advantage the BRFSS survey has is the amount of data available in terms of the number of respondents. So, the best alternative would be to test out the system with the smaller NES data set, and then move on to the larger data. See Section 8 for a discussion of why this turned out not to be feasible for this project.

Testing out the tool with multiple data sets would have the added benefit of demonstrating that the tool scales with respect to the number of respondents, or revealing any potential problems. I do not consider the number of questions as being of large importance to scaling as both of the above data sets are representative of "large" surveys. In general, the number of questions on a survey is limited by the amount of time people are willing to spend filling it out. If this tool were to be used by government or industry researchers, they cannot be expected to have surveys much larger than those provided by the NES or BRFSS. If this tool were to be used by the academic community, for example for the analysis of psychology experiments or user studies, the surveys can be expected to be much smaller.

In the end, I settled for using only the NES 2002 data set, so all the screenshots in Section 12 and experiments in Sections 6 and 7 use this data.

2.3 Suitability of dataset to task

Response options on most surveys are arranged in such a way as to make simple comparative analysis possible. For example, for a set of yes/no/maybe questions, similar answers for different questions are going to be mapped to the same numerical value, which allows the user to interpret the raw data scatterplot more intuitively. Correlation values need to be interpreted in a slightly different way for each survey. For the NES data set a positive correlation means that for a given pair of issues, users who responded a certain way to one of the questions are more likely to respond similarly positively (or negatively) to the other question. For the BRFSS survey this means that there is a strong tie between the behaviours described in the questions being examined, which would help identify the possible health risks of this group of people.

It should be noted that there are two orthogonal axes that should be used to evaluate the suitability of data for this tool. First of all, there is the number of respondents. Since the tool, at the overview level, deals with aggregate correlations, this is mainly an issue in terms of raw computation. Methods are available to deal with these challenges, such as doing pre-computations before user interaction begins and saving the results in a file, or using a more efficient statistical tool as a backend on top of which the interactive GUI can be built.

The more interesting of the two axes is the number of questions. This is the real test of the scalability of the tool, as the context display is defined by the questions and *guide-lines* (see Section 4 for a description) between the questions and the relevant correlation values. If the tool cannot handle increases in the number of questions gracefully, it will not scale well. It should be noted, however, that there is a practical limit on the number of questions one is likely to find in a survey. The two data sets that I have presented are, in that respect, at the upper end of the spectrum. Additionally, larger surveys are often divided into smaller self-sufficient sub-sections, which would allow the analyst to break up the dataset into a number of subsets that can be visualized independently.

3. RELATED WORK

There are a number of tools available for the display of multi-dimensional data. Parallel co-ordinates have been quite influential since their inception, and are now even used in some commercial tools, alongside the more traditional display and summary methods [13], [14], [19]. The basic idea of parallel co-ordinates [18] is that instead of ensuring that the axes are placed orthogonally, they are placed parallel to each other. This is a large step forward from the orthogonal axes, where anything beyond three dimensions is nearly impossible for the human mind to visualize since more abstract representations are necessary to maintain the orthogonality between multiple axes. On the contrary, the parallel axes make visualizations of a few dozen dimensions a nearly trivial task. Additionally, the paper's claim is that a number of basic patterns can easily by picked out from the data by a trained user. However, the method still has obvious limitations in the number of dimensions that can be presented concurrently. It is a fine technique for a few dozen dimensions but unsuitable for questionnaires where numbers on the order of 100 questions are more typical.

Multidimensional scaling [7] deals with reducing the dimensionality of data by mapping it to a lower number of dimensions using some reduction function. The authors focus on creating 2-dimensional representations of multidimensional datasets. They also present a layout technique that more faithfully represents the relationships between the points in reduced dimensions.

Hierarchical dimension ordering [20] is another method for dealing with high-dimensional data. In this work, the authors propose a "general approach to dimension management for high-dimensional visualization". They hypothesize that approaches such as multidimensional scaling don't work for many data sets because the user loses track of the original dimensions and is presented with a number of new, derived dimensions. In this work, in addition to being able to filter by some dimension, the dimensions are arranged in a hierarchical structure that is built from the similarity measures between the dimensions. The challenge with my data sets is that the similarity between dimensions (questions) is the correlation measure, and building a hierarchy out of these won't provide much additional value.

A combination of the parallel co-ordinates and hierarchical dimension ordering can be seen in hierarchical parallel co-ordinates [3]. This approach is described as being good for a set with a reasonable number of dimensions, and a very large number of data points. It is seen as a way to organize and bring out patterns in the data that would otherwise be difficult to see due to the overplotting that results from so many observations. In some ways, this approach may be useful for exploring relationships between a small set of questions. However, there are already ways of compacting the information contained in survey answers, such as correlations, means, etc., which seem to me to be even more space-efficient. What's needed is a way of categorizing and highlighting the relationships between the correlation measures, not of condensing patterns in the raw data. Overall, though, I like the authors' approach of aiming to support exploratory data analysis, in particular displays that summarize, manipulate, as well as help uncover structure in the data.

The TableLens [10] is another tool that deals with graphic and symbolic representation of multidimensional tabular data. This tool is a good example of good use of labels and critical layout, which is dependent on the type of data being displayed. The TableLens is another tool that makes it easy to find information that would be nearly impossible to see using a traditional spreadsheet, and it is a valuable resource to learn from.

One fundamental difference between these tools and what I'm trying to accomplish is that these tools assume that the name of each of the dimensions has a well-defined meaning when it is used as a label. For example, if car data is being explored, and each axis is labeled as "Price", or "Mileage" (see [18] for the detailed example), the meaning of these is fairly evident to the viewer. However, if the same were attempted with a questionnaire data set, labels such as "Question 10" would not provide the user with nearly as much information.

4. DESCRIPTION OF SOLUTION

The tool utilizes a combination of an overview and multiple linked views. The secondary views are brought up as the user interacts with the data displayed in these views. A variety of tools on the *Control Panel* enable the user to adjust the overview display to their current needs, through a combination of dynamic queries and space management techniques.

There are two main starting displays: the *Textual View* which is the textual display of survey questions, and the graphical display of questions and correlations, the *Questionnaire View*. The use of each of these will be discussed in more detail in the context of the specific scenarios in Section 6, while their main characteristics are discussed below. For now, suffice it to say that these two views support alternative means of exploration, although the paths taken from each of them will intersect at certain displays.

In the *Questionnaire View*, pairs of questions are linked to the correlations to which they correspond by *guide-lines*, shown and labeled in Figure 1.

The *Textual View*, shown in Figure 3, is simply the display of all the questions in the survey, where for each question it contains such information as the:

- question identifier as used in the data set
- general topic addressed by the question
- text of the question as it was posed to respondents
- answer choices and the numerical values assigned to them in the dataset.

These attributes differ depending on which dataset is being explored. The last two are especially important for putting the raw numerical results into context when the users drill down through the linked views to the more specific information. For example, if the key is similar to Figure 3, then upon viewing the raw-points scatterplot which involves the selected question, the user will be able to determine that the answer of "0" correspond to "NA", and an answer of "5" corresponds to "Disapprove Strongly". The survey adheres to the convention of "0" meaning no answer, and the approval scale going from 1-5, in many places.

The user can scroll through the *Textual View*, looking for questions of interest. Clicking on the identifier for a chosen question will bring up a secondary display, the *Correlation View*, shown in Figure 2. This shows a scatterplot of the correlations this question has with others. It should be noted that the questions are numbered in the order they appear in the survey. The reason for numbering the questions on this axis instead of displaying the textual labels is that the existing graph classes do not permit the latter option.

The user can then devote their attention to this display, or go back to browsing in the textual window if they don't find any correlations of interest. However, if a particular correlation catches the user's attention, clicking on the corresponding point will link to two different displays. One of the resulting frames is the *Questionnaire View* (Figure 1)*,* the main starting display for the other branch of exploration. The point on which the user clicked in Figure 2, as well as the guide-lines from it to the relevant questions, will be highlighted in order to allow the user to keep track of their context, and enable them to retrace their path. The other frame is the *Raw Data View* for the selected correlation, which is discussed a bit later in this section.

At the other starting point, there is the *Questionnaire View,* seen in Figure 1*,* which shows the questions and the corresponding correlations. This display shows the aggregate information for each relationship of interest in the form of a single correlation value. Each point on the semi-circle represents a question in the survey. The semi-circular display was largely inspired by TextArc [15]. The horizontal axis represents the correlations, and accordingly runs from –1 to +1. A slider on the *Control Panel*, shown in Figure 4, dynamically controls this range.

A point on the correlation axis corresponds to a correlation between the results for two questions. The questions are connected to the relevant correlation point by a line, called the *guide-line*. If a correlation or its guide-lines have been selected, the shape of the point will change, and both the point and guide-lines will be highlighted in yellow. A guide-line can also be selected by clicking on it, and will also be highlighted yellow. Clicking on a selected element will de-select it. The statistical aspect of these correlation calculations is discussed in Section 5.2.

The display also provides context in a different way, by enabling the user to explore the questions themselves. Clicking on a question point will result in linking to the *Correlation View*, which shows the relationships that this question has with others. One minor detail that has not been implemented but would be useful is to also link this action to the *Textual View*, so the textual description for the relevant question would also be displayed. This would allow the user to place the numerical values that have been assigned to each response in the proper context. For the time being, the user must scroll down to the question manually.

Clicking on a correlation point will result in a secondary display being brought up. This is the *Raw Data View*, which shows the actual raw data in detail, in the form of a scatterplot. In the future, it may be beneficial to give the users a choice between a scatterplot, parallel co-ordinates, or some other specific visualization. This is also the view for which I had to develop an augmented version of the scatterplot. There are two alternative implementations that serve the same goal, shown in Figures 7 and 8, but without some sort of a user study I cannot confirm which one is better.

The basic idea behind these alternatives to the scatterplot is to provide more information than is normally available. The main drawback of the plain scatterplot visualization is that each point is the same size. Since the answers for each question are (generally) limited to 10 choices, the scatter-plot often ends up with no more than a few dozen points. When the respondents number in the thousands, such a display is not very informative, as the user has no information as to the frequency of each of the response combinations. A higher frequency would mean that more respondents picked that particular combination of answers. My suggestion is for each point to encode the proportion of answers that fall in this category through either its colour (Figure 7) or size (Figure 8). A possible drawback with the use of colour is that I currently use a very simple scaling function to transform the proportion of answers in this category to the amount of the green component(G) of the RGB. A less saturated ("brighter") colour corresponds to a larger frequency, while more saturated ("darker") squares correspond to smaller frequencies. The possible problems with both of these approaches are discussed in Section 9, and solutions are suggested.

The number and kind of relationships displayed can be controlled by the user through a combination of sliders and buttons located on the *Control Panel*, shown in Figure 4, which provide a limited dynamic query capability. These queries are mainly a way to deal with the overwhelming number of guide-lines that appear initially in the *Questionnaire View*. The "Correlations" slider, seen at the top left of Figure 4, allows the users to select a correlation range they would like to see. The horizontal axis then expands to display that range exclusively, as can be seen in the transition between Figure 9 and 10. The guide-lines whose correlation points falls outside that range are filtered out and not displayed in this mode. This may be useful when the user is only interested in exploring high correlations initially, in which case they can immediately move the slider to the upper range.

Additionally, in some cases the user may not be interested in the sign of the correlation, only in its magnitude. For this purpose, they can use the "Absolute Value" vs. "Full Range" radio buttons, which will flip the display between these two modes. In the absolute value mode, the correlation axis will run from 0 to 1, and the absolute values of the negative correlations will be plotted in corresponding locations on the positive side. As can be seen in the transition between Figures 11 and 12, the highlighted lines that have high negative correlations flip to the corresponding position on the positive side. The line that has a positive correlation stays where it was.

The buttons on the right-hand side of the control panel can also be used to aid in exploration. Once the user has selected a correlation point, they can request to see either the "Next" or "Previous" correlation point. This will cause the nearest neighbouring correlation point to be highlighted along with its guide-lines, and the *Raw Data View* for the this question will also be popped up.

The "Clear Selections" button can be used to clear both the selected correlations and guide-lines.

One option that is provided can be used as a simple way to mitigate the lack of a multiple correlation display. Upon selecting a number of guide-lines, the user can click the "Cross-correlations" button, which will highlight the correlation points and their corresponding guide-lines for all the correlations in the cross-product of this set with itself. For example, suppose the user has selected the lines between Q1 and Q57, as well as Q42 and Q49. What they will see upon the completion of this operation are all the correlations between the four questions in the set, i.e. they will see correlations between Q1 and Q42, Q1 and Q49, Q42 and Q57, Q49 and Q57, in addition to the original two. The trivial relations such as the correlation between Q1 and itself are omitted, as they can provide no useful information. No raw data graphs are created for any of these correlations to avoid bombarding the user with too many graphs that they did not explicitly request.

It is perhaps worth noting that most of these views are linked together, and clicking on a point in one view will result in some action being taken on one or more of the other views, and the corresponding elements being highlighted in these elements if possible. This approach differs somewhat from the traditional application of brushing and linking, discussed in [1] and [16], but still maintains the basic notion of these concepts. Direct linking is not possible for all views as some are at different level of abstraction than others, but most meaningful connections are supported.

5. HIGH-LEVEL IMPLEMENTATION

The implementation of the tool can be separated into two major categories, each of which will be discussed in turn. One is the actual graphical display, and the back-end that allows the proper updating and maintenance of the display. The other one is the statistical machinery used to carry out the correlation calculations that are then used in the display. While the first is by far the major component of this project, I will nonetheless devote a fair bit of attention to the latter, as it presented a number of interesting challenges.

5.1 Visualization component

Of a number of toolkits that I surveyed during the initial stages of the project, the two that I ended up using are the InfoVis toolkit [2], [5] and the Scientific Graphics Toolkit (SGT) [11].

The InfoVis Toolkit provides easy-to-use native parsers and convenient data structures in the form of tables, with rows and columns recognized as distinct objects. There are also many kinds of graphs (visualizations) provided by the toolkit, which is why I chose it originally. I ended up using it mainly for the data- and file-handling aspect. In the Future Work section, I also suggest using it for some of the more complicated graphs, e.g. the parallel co-ordinates. Of course, these would not be too hard to implement in SGT, as well.

The file parsing facilities in the InfoVis toolkit are a double-edged sword. The disadvantage is in the fact that the toolkit uses its own specialized form of the CSV file format, with extra book-keeping information (like the number of rows and columns, the names of columns, etc.) appended to the front of the file. The file appears with the extension "TQD". This is not a standard file format like CSV (which is fairly widespread), and so the toolkit is not compatible with many standardized data sets. This can be remedied with a simple parser, which will attach the necessary information to the header of the file. Of course, for each file format that you would like the tool to work with, a separate parser needs to be written. However, the ability to parse a file in to an internal table representation remains a major advantage.

The biggest problem with the InfoVis toolkit was the initial difficulty in extending the native classes to do what I wanted, while the SGT provided a lot more flexibility. One inherent advantage of the InfoVis toolkit over SGT was the built-in lenses and zooming, as well as the more complex visualizations, and the apparent ease with which such operations can be extended. However, lacking reasonable documentation, this toolkit is a much less productive contribution to the infovis tool development community than it could be. The main challenge of this project was the fact that none of the visualizations available in any of the toolkits were exactly what would suit my needs, so the more intricate functionality of some views had to be implemented by hand.

The SGT is another toolkit that is provided and maintained by members of the academic community. It was developed for interactive visualization of scientific data, and has been applied mainly to atmospheric and oceanic datasets. Its main selling feature, to me, was a basic graph class that was much more easily extensible, and which I was able to modify for my task. In the *Questionnaire View*, the biggest challenge was the display of the semi-circular line of questions, and the integration of this semi-circular axis with the regular horizontal one. One thing that could be a good addition from this project to a toolkit is an extensible axis that would allow you to change the shape/orientation of the axis beyond the perpendicular x/y currently present. The use of the SGT toolkit, along with standard Java and Swing methods for reacting to user events, allowed me to implement the flexible interactive functionality that I desired.

Finally, the implementation also involves a file converter, which must be run to transforms plain CSV files to the InfoVis toolkit's TQD format before a new data set can be loaded into the tool.

5.2 Statistical Component

Another piece vital to the implementation was the statistical back-end, which was necessary to compute the correlations given the raw results for each question. There are a number of statistical routines, if not full statistical packages, available freely on the web, and I used an implementation of Spearman's rank correlation [12]. There are a number of various routines available for calculating correlation, each applicable to a certain subset of data. Spearman's correlation differs from some of the others in that it does not make an assumption as to the linearity of the trend line [6]. I felt that this was more appropriate for use with real-world results obtained from surveys, as relationships between subsets of results may become quite complicated and shouldn't be expected to conform to a linear relationship.

Originally, I had intended to allow users to deal with multiple correlations (where multiple refers to the number of questions between which the correlation is calculated), but was met with a number of problems due to the nature of various statistical analyses. This is discussed in more detail in Section 8. The tool currently supports only two-way correlations.

Multidimensional scaling is another possibility for showing a more manageable representation of a multidimensional data set. The problem with using this approach is that for the highest-level overview, two orthogonal sets of dimensions need to be reduced – the set where the questions are the dimensions, as well as the one where the results for each independent respondent are the dimensions. If two separate scaling functions are used, the resulting representation is more likely to be further removed from the true value. And a single scaling function that takes into account the variation in dimensions in both sets and reduces all of them concurrently would become too complicated to be used efficiently. For these reasons, I decided that multidimensional scaling was not the optimal approach for this tool, at least in terms of the top-level visualizations. In the future, especially for the *Raw Data View*, it is possible that the user would have a choice of data representation techniques.

6. SCENARIOS OF USE

There are essentially two different exploration paths available when using this tool. One is to start with the *Textual View* of the questionnaire, and progress through, exploring relationships between questions whose description sounds like it would be of interest. The other is to start with the graphical display of questions and correlations in the *Questionnaire View*, and explore exceptional relationships that are pointed out through the interface. I believe that the former approach would be more suitable for someone interested in analyzing particular aspects of the survey, based on specific phrasings of the questions, while the latter method would be of interest to an analyst trying to detect general patterns, and not restricted to looking for something in particular.

6.1 Contextual Analysis

The user is interested in the general subject area that the survey was addressing, but is not familiar with the general layout of the questions, nor are they necessarily looking for any certain numerical correlation. They are simply interested in exploring the topics addressed by the questionnaire, and the relationships between them. However, they do not have a specific set of topics in mind.

In this case, the first path of exploration would be more beneficial. So, the user would start with the *Textual View*, described in Section 4. After scrolling through the textual display and finding a question that interests them, the user may click on the button that corresponds to that question. This action will bring up the *Correlation View*. This plot gives the user the ability to pick out some outlier correlations, if that's what they're interested in, or simply survey the distribution of correlations this question has with others. This graph links back to the main *Questionnaire View*, highlighting the selected correlation in a different colour, to allow the user to easily distinguish it from the sea of other plots. The user may go back to the *Correlation View* and select new correlations, or further explore from the *Questionnaire View*.

6.2 Value-based Analysis

The user is interested in finding exceptional correlations, perhaps with a high correlation value, or within a certain range. Initially, they don't care too much about the context of these relationships.

For this kind of exploration, the graphical display of questions and correlations in the *Questionnaire View* would be a better starting point**.** Clicking on a question-point brings up the *Correlation View*. Like before, this plot is linked back to the main questionnaire graph, and so selecting a point on it will highlight another line on the main graph. Clicking on a correlation-point brings up the *Raw Data View*, which is a scatter-plot of the raw data of the two questions that correspond to this correlation point.

7. RESULTS AND EVALUATION

The main set of data I concentrated my attention on was the 2002 NES survey. I used the two scenario techniques described in Section 6 to explore the data. I found the contextual analysis scenario approach to be more useful than the value-based analysis. However, this actually serves to validate my claim, as I was interested in specific issues, and thus needed the *Textual View* to provide me the necessary initial context.

There are a number of interesting bits of information that I found through my exploration. In the following discussion, excerpts from the survey are placed in quotation marks to distinguish them from my own suppositions or ideas.

As an example, people's approval of congress was not strongly dependent on the strength of their approval or disapproval of President Bush.

Among the individuals who said that the gap between the rich and poor has increased over the two decades, paradoxically enough there were quite a few who also claimed that there should actually be a decrease in pre-school funding for children in black or poor neighbourhoods.

"Opinion thermometers" were applied to a number of prominent political figures. Many more people didn't recognize or couldn't rate Ralph Nader than Laura Bush. Those who rated Nader highly also tended to rate Laura Bush fairly high. This is in slight contrast to those who rated Laura Bush highly, where there was a much larger spread of opinions about Nader, with nearly as many rating on the low end of the scale as on the high. An “opinion thermometer” asks for a judgement as to how favourably you feel towards the person/issue is, on a scale of 0 to 100. A rating below 50 would mean the respondent is unfavourable or cool towards the person in question, whereas a rating above 50 would mean the respondent is favourably impressed with what the person is doing.

There was also an interesting relationship that emerged between giving a presentation to a congregation as part of participating in a place of worship, and answers to the question of whether “whites have more in-born ability to learn”. The latter was asked as a possible explanation for differences in employment and salary levels of the “whites” and “blacks”. The majority of people, regardless of their participation in a congregation, said that the above phrase is not at all important as an explanation. However, while those that haven’t given a speech did not select any of the other options, a fairly noticeable number of those who have given a speech said that the statement was either very or somewhat important as an explanation. However, they were also the group that also chose the “statement isn’t true” option with frequency comparable to the other two answers.

Another thing that I noticed is that there are many questions that are linked through multiple levels of indirection. These are often questions of the form “if you answers yes to question X, and if you are a Y, what do you think of ...”. I am not yet sure of the effect such questions have on overall correlations, as for some respondents they will necessarily remain blank. Also, if there were a way to identify such sequences of questions perhaps the tool would be able to extract more meaningful correlations of groups of questions. There isn’t such a method currently, as far as I can tell, at least without extensive parsing.

Finally, I was able to discern that there is even more structure to question identifiers than I had previously thought. For the data set that I explored, it now appears as though the questions beginning with V022 are used for administrative purposes, while those beginning with V023 are the actual multiple-choice questions. If such a separation can be established in any data set, then a number of useful properties emerge. For one, the administrative data can be ignored during the more detailed analysis. However, this also opens up another use of the tool – in this case, not only for researchers interested in the results of the survey, but for survey designers interested in finding out how well the survey was constructed. In particular, they can analyze whether the ordering of questions affected the results, and find ways to fix the survey if the results are skewed in some way.

I would also like to comment on the scalability of the tool. As I haven’t been able to test it with the BRFSS data, I have no experiences with scaling with respect to the number of respondents. However, I do not anticipate that to be a problem, as the majority of the views deal with aggregate data, which is calculated only once, when the tool is loaded. The *Raw Data View* is the only one that deals with the raw numbers, so that is the one possible place where a slow-down might be noticed with a larger data set. As I have already mentioned, the NES data set has a sufficiently large number of questions to test for scalability in that dimension, but during the initial stages I also tested the tool on a much smaller survey (20 questions). The slow-down for the NES data over the small survey is noticeable, it’s small enough that the tool remains interactive.

The major strength of this project is its uniqueness. So far, I have not been able to find any similar tools. The main advantage of my approach is that it provides a visual overview, in addition to details. Both textual and raw-data details are available on demand.

Another strength is the fact that multiple exploration models are supported. Each allows the user to extract different kinds of information from the data set. The fact that the use scenarios for these models actually intersect allows the user to switch search methods halfway through the session if they find that their goals have now changed.

Performance remains one of the major weaknesses. Due to my use of the InfoVis toolkit, some of the data processing is necessarily slow. In addition, I did not do many optimizations to the displays that were implemented by hand, in part due to lack of time and my limited knowledge of graphics in Java. There were also issues with data formats, as discussed in the next section.

Another weakness is that there are few dimensions that can be used to filter data. One filter that I would really like to make available would be based on the questions themselves, and is discussed in the Future Work section.

8. LESSONS LEARNED

There are a number of lessons that I am able to take away from the experience, some more technical than others. During this project, I had to deal with a number of issues. At the forefront were the programming and design issues, such as learning unfamiliar toolkits and designing the overall system. Some of the more challenging questions were specifically infovis questions, such as coming up with the best visualization technique, and then also designing each of the components to fit together properly. Finally, due to the nature of my project, there were also the purely mathematical challenges, dealing with statistical methods that underlie the tool.

8.1 Implementation

Perhaps most importantly, doing the preliminary analysis and choosing a toolkit brought home the abundance of various toolkits. But with such diversity one also pays the price in quality of software. Many toolkits are poorly documented, and the learning curve is fairly high. Once learned, however, they make certain tasks fast and efficient to perform. Additionally, many of these toolkits allow the extension of their native elements in simple and predictable ways, and also allow the creation of new compatible elements.

On a similar note, not having had any experience with graphics prior to this project set me back somewhat, but also made me realize the value of standardized development elements such as the Java AWT and Swing. In retrospect, it would be nice to develop a similar standard for visualization toolkits. At the very least, it seems plausible that, for each toolkit, general guidelines on how to extend each element would simplify the task for new users.

8.2 Statistics

Statistical analysis is much more complicated than my prior knowledge would have led me to believe. There is a marked difference between the analysis of two-way vs. multiple-way interactions. This presented a challenge in incorporating it into the tool, as the presentation needs to be consistent for all interaction relationships. The initial intent was to enable the user to select the "degree" of correlations that they would like to explore – for example, correlations between two questions, or correlations between five questions, etc. The advantage of this kind of interaction is that the user can concentrate on far-reaching relationships, which would expose deeper (and presumably more interesting) trends.

The problem lies in the way that general statistics are used to handle calculations such as these. For the two-question case, there is no problem as the calculation is "directionless". However, as soon as the number of questions increases above two, the correlation becomes "directional". For all of these cases, the correlation is calculated as the effect of the (n-1) questions on the remaining one, if you are trying to find the correlation among n questions. The number of correlations calculated for a given question would increase both with the number of questions in the survey and the number of questions one wishes to correlate at a given time, as all (n-1)-question sets would have to be considered. I was not able to find any way to derive a single correlation measure for the n questions from these partial correlations. For this reason, I decided that multiple correlations were not something that could be efficiently supported by this tool at this time.

8.3 File Formats

I was met with a number of challenges, which resulted in not all of my original exploration objectives being completed.

The first challenge is the varied format in which the results are stored. Even within the NES results, the datasets from different years are stored in slightly different formats. This comment applies to both the actual data obtained from respondents, as well as the textual descriptions of the questionnaire and answer variables. In particular, such things as

- location and format of variable names
- question names
- separators between question descriptions, as well as
- separators between answers to questions

varied significantly. In general, this is not an insurmountable problem, as these are still highly structured formats, and parsers can be written to transform them into the format the tool currently works with. However, due to the time limitations of this project, I chose to concentrate on the information-visualization aspects, instead of writing parser-transformers.

9. FUTURE WORK

Due to the time constraints on the project, as well as some of the challenges discussed in Section 8, there are a number of elements of the visualization that were not implemented. In addition, through using the tool to explore the existing data set, I was able to identify a number of areas for improvement, as well as features that would be useful, but hadn't been anticipated.

A simple addition to one of the secondary visualizations is depicted in Figure 5. This would be applied either to the *Questionnaire View*, or to the *Correlation View* for a given question. In this way, the user would be able to more precisely determine the proportion of relationships within a given range of correlations, thus perhaps directing their search more efficiently. This would be especially relevant in the initial stages of exploration with the *Questionnaire View*, where the number of guide-lines in the graph is simply too large to make much sense of any existing patterns.

My knowledge of the administration of surveys isn't extensive enough to be able to tell if there are any expectations as to the distribution of correlations for a given kind of survey. Assuming there are, there would be another use for the graph proposed in Figure 5. For this, the bar-graph should show the correlations over the whole range. If, for example, a normal distribution of correlations is expected, and the resulting graph deviates from it in some places, those might be the ones where the analyst would concentrate their attention.

In the discussion of the *Raw Data View*, I mentioned the augmented scatterplot. There are a number of problems with both of the approaches suggested. When size is used to encode frequency, larger points tend to overshadow smaller ones, especially in a dense area. When colour is used, the differences between the darker points are much harder to distinguish. To deal with colour, a rainbow spectrum can be used instead of the saturation, or perhaps the logarithm of frequencies can be used to determine the saturation. However, in the latter case the colours would not be exact representations of the values.

There are a number of possible improvements to the *Questionnaire View*. The display that was implemented is one of the possibilities I considered. Another was to eliminate the *Textual View* and display the text corresponding to each question in the *Questionnaire View*, perhaps as a mouse-over when the user hovered their mouse over each question. The reasoning for the alternative approach was that if it was implemented unobtrusively, e.g. with a mouse-over pop-up as suggested before, it would offload some of the cognitive effort from the user. The information would now be available in one tool, instead of having to jump between multiple views. However, this implementation also has the drawback of obscuring a large portion of the main display, thus eliminating one kind of context while providing another. I would argue that it is more important for the user to see the display of questions and guide-lines, and with the proper positioning of the panels it little effort would be expended to glance at the appropriate question in a separate display. However, in the future it would be interesting to implement this alternative, and run a test study to discover which is preferred by users.

Another proposed change involves the question points being located on a full circular axis, instead of a semi-circular one as the are now. It would be interesting to see which alternative the users prefer. My personal opinion is that the original version, shown in Figure 1 provides more of a separation between the question-points and the correlation-points, while this alternate representation may get so cluttered by the large number of guide-lines that the distinction would be lost.

As mentioned before, some of the functionality was limited due to the difference in formats. This problem can be alleviated through the use of extensive text and pattern matching facilities. This is not something that I had the time for, but it would be one of the first major additions that I would consider if I kept working on this project. In addition to simplifying the loading of the files, this would also allow for more sophisticated search facilities, as well as filtering of the questions based on the actual content of the question. This was one of the features that I felt were missing from this version of the tool, where oftentimes the display of all the question-points was unnecessary, but I had no way of dealing with it at this time. With this addition, the users could dynamically filter questions based on some issue of interest, and the irrelevant questions (and the corresponding guide-lines, which are the main source of clutter) would no longer be displayed.

10. CONCLUSION

In this paper, I have presented a new tool for visualizing both surveys and the corresponding results obtained from polling respondents. Displays for showing both the survey text and the raw and processed results are presented, linked in a way that makes exploration more continuous. The tool is used to explore a real dataset, with some interesting results discovered. A number of implementation challenges are discussed, and desirable interaction features are suggested as potential improvements to the tool.

In addition to being a good exercise in applying information visualization techniques to a new area, this project was also a good experience with project management, in terms of the degree of organization that is expected for such a significant project.

11. REFERENCES

[1] Becker, R. A., Cleveland, W. S. Brushing scatterplots. Tehcnometrics 29, 1987.

[2] Fekete, Jean-Daniel. The InfoVis Toolkit. Proceedings of the 10th IEEE Symposium on Information Visualization, IEEE Press, 2004.

[3] Fua, Ying-Huey, Ward, Matthew O., Rundensteiner, Elke A. Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets. IEEE Visualization, 1999.

[4] Gnanadesikan, R. Methods for Statistical Data Analysis of Multivariate Observations. John Wiley & Sons, Toronto, 1977.

[5] InfoVis Toolkit, http://ivtk.sourceforge.net//

[6] McPherson, Glen. Applying and Interpreting Statistics: A Comprehensive Guide, 2nd Edition. Springer-Verlag, New York, 2001.

[7] Morrison, Alistair, Ross, Greg, Chalmers, Matthew. Fast Multidimensional Scaling through Sampling, Springs and Interpolation. Information Visualization 2(1). March 2003.

[8] National Center for Chronic Disease Prevention and Health Promotion, http://www.cdc.gov/BRFSS/technical_infodata/surveydata/2003.htm

[9] National Election Studies, http://www.umich.edu/~nes/studyres/download/nesdatacenter.htm

[10] Rao, Ramana, Card, Stuart K. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. SIGCHI, 1994.

[11] Scientific Graphics Toolkit, http://www.epic.noaa.gov/java/sgt/

[12] Spearman's Rank Correlation, Stats 101- Elementary Statistics Applets, http://intrepid.mcs.kent.edu/~blewis/stat/index.html

[13] Spotfire, http://www.spotfire.com/

[14] Tableau: The Visual Spreadsheet. http://www.tableausoftware.com/

[15] TextArc, http://www.textarc.org/

[16] Ward, M. O., Martin, A. R. High dimensional brushing for interactive exploration of multivariate data. In Proc. IEEE Visualization 1995.

[17] Weatherburn, C.E. A First Course in Mathematical Statistics. Cambridge University Press, London, 1957.

[18] Wegman, Edward J. Hyperdimensional Data Analysis Using Parallel Coordinates. Journal of the American Statistical Association, Vol. 85, No. 411. September 1990.

[19] XmdvTool. http://davis.wpi.edu/~xmdv/

[20] Yang, Jing, Peng, Wei, Ward, Matthew O., Rundensteiner, Elke A. Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration Of High Dimensional Datasets. Proc. InfoVis, 2003.

12. APPENDIX

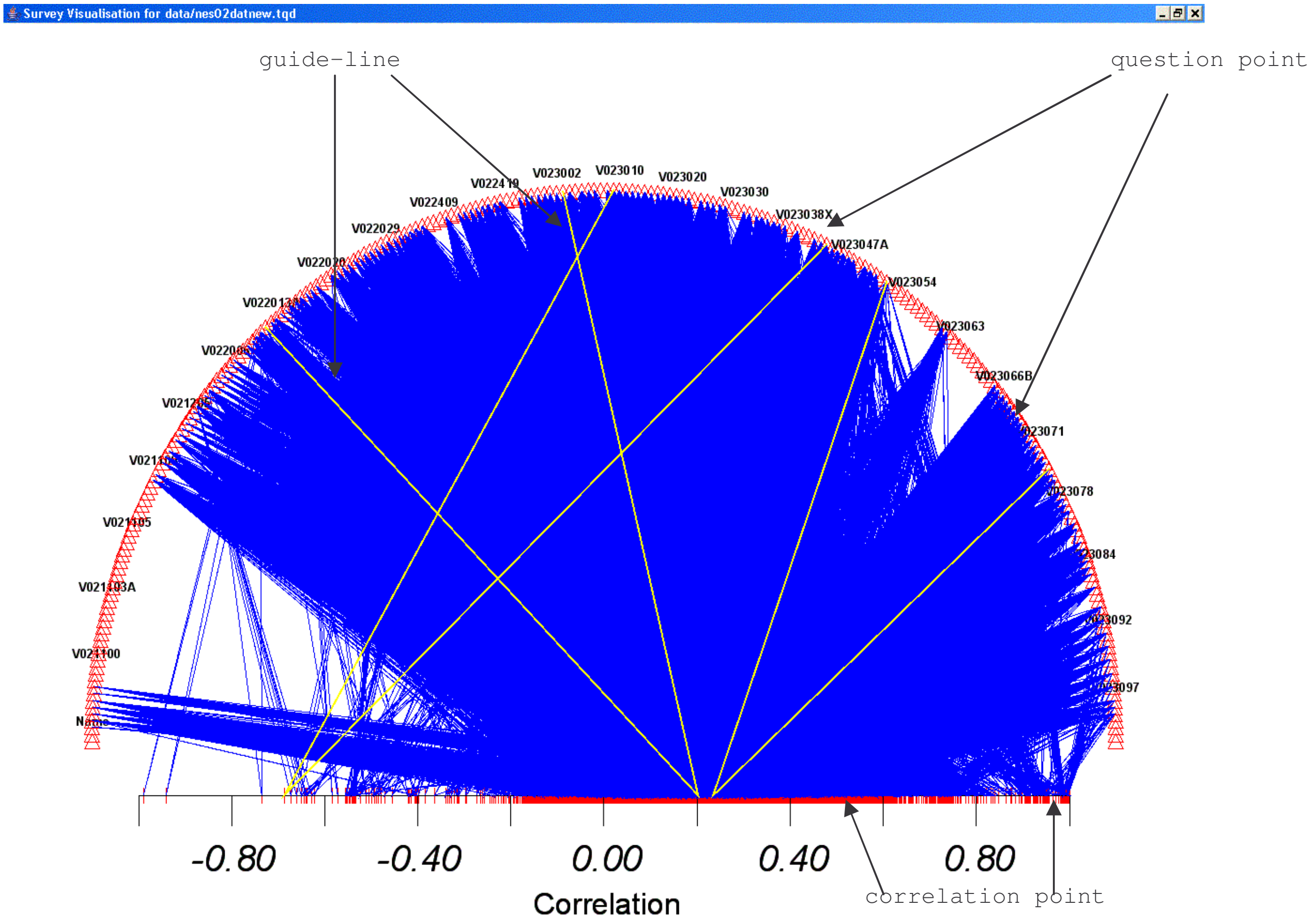


Figure 1: The Correlation View. The major elements are labeled on the screenshot

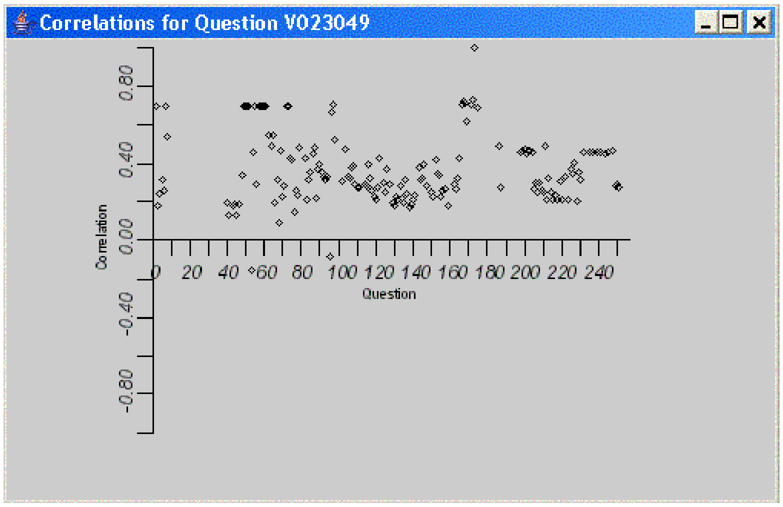
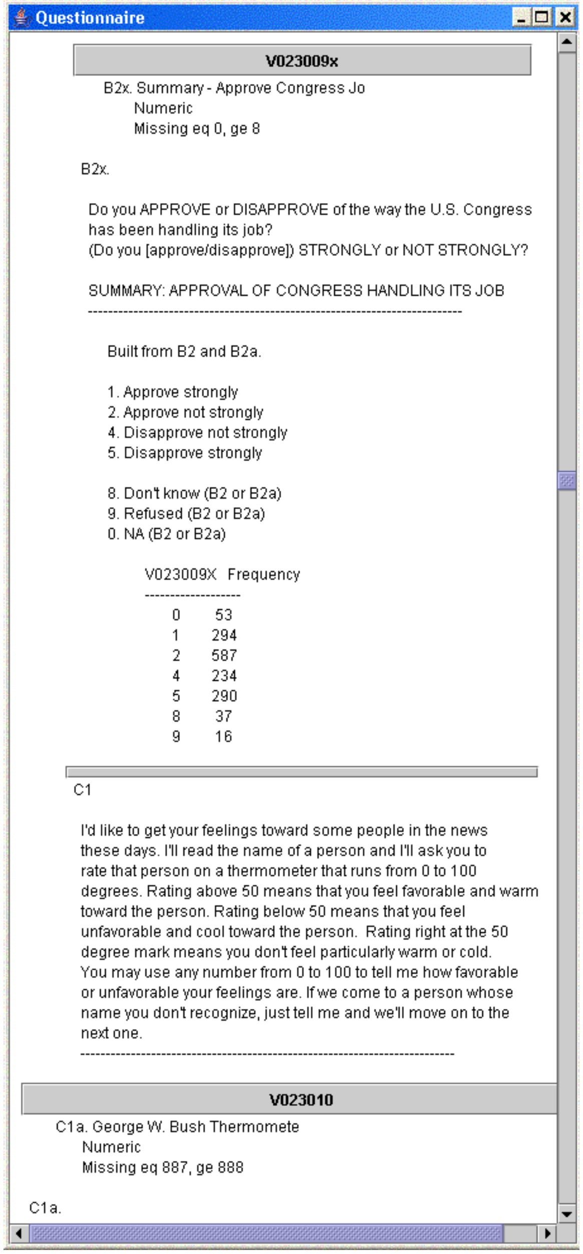
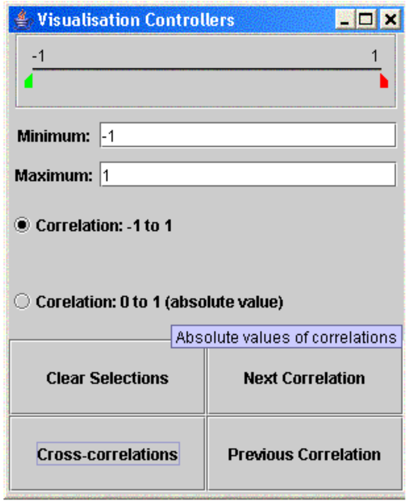


Figure 2: The Correlation View. The points represent the correlations question V023049 has with all the other questions in the survey.



```
V023009x
    B2x. Summary - Approve Congress Jo
        Numeric
        Missing eq 0, ge 8

B2x.

 Do you APPROVE or DISAPPROVE of the way the U.S. Congress
 has been handling its job?
 (Do you [approve/disapprove]) STRONGLY or NOT STRONGLY?

 SUMMARY: APPROVAL OF CONGRESS HANDLING ITS JOB
 ------------------------------------------------------------------------

      Built from B2 and B2a.

      1. Approve strongly
      2. Approve not strongly
      4. Disapprove not strongly
      5. Disapprove strongly

      8. Don't know (B2 or B2a)
      9. Refused (B2 or B2a)
      0. NA (B2 or B2a)

           V023009X   Frequency
           -------------------
               0       53
               1      294
               2      587
               4      234
               5      290
               8       37
               9       16

C1

 I'd like to get your feelings toward some people in the news
 these days. I'll read the name of a person and I'll ask you to
 rate that person on a thermometer that runs from 0 to 100
 degrees. Rating above 50 means that you feel favorable and warm
 toward the person. Rating below 50 means that you feel
 unfavorable and cool toward the person.  Rating right at the 50
 degree mark means you don't feel particularly warm or cold.
 You may use any number from 0 to 100 to tell me how favorable
 or unfavorable your feelings are. If we come to a person whose
 name you don't recognize, just tell me and we'll move on to the
 next one.
 ------------------------------------------------------------------------

V023010
    C1a. George W. Bush Thermomete
        Numeric
        Missing eq 887, ge 888

C1a.
```

Figure 3: The Textual View. Sample survey question with multiple choice answers.



Minimum: -1

Maximum: 1

● Correlation: -1 to 1

○ Corelation: 0 to 1 (absolute value)

Figure 4: The Control Panel.

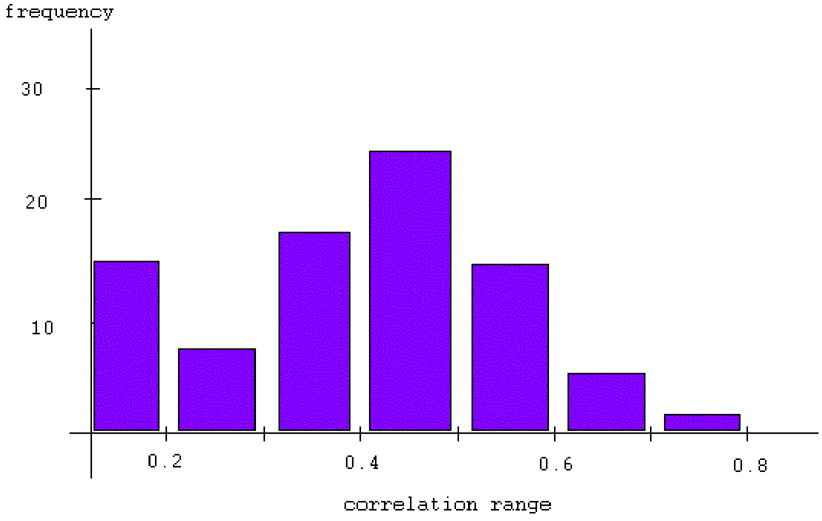
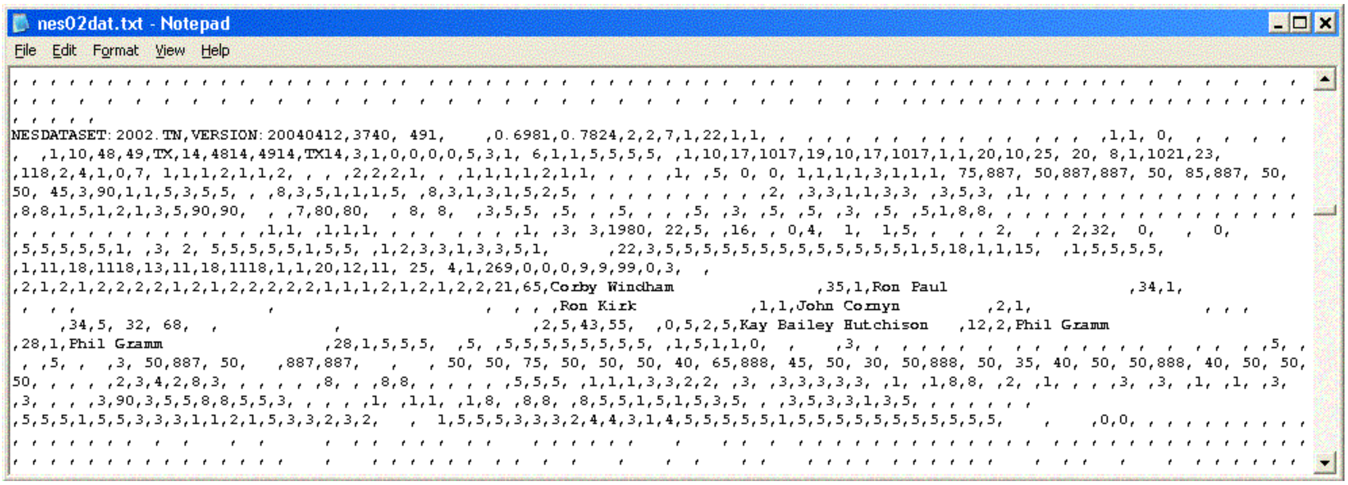


Figure 5: Proposed summary view for the exploration of correlations



```
, , , , , , , , , , , , ,  , , , , , , , ,  , , , , , , , ,  , , , , , , , ,  , , , ,  ,  ,  , , , , , , , , , , , , , , ,  ,  ,  ,  ,  ,
, , ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  , , , , , , , , , , , , , , , , , , , , , , ,
, , , , ,
NESDATASET:2002.TN,VERSION:20040412,3740, 491,     ,0.6981,0.7824,2,2,7,1,22,1,1, ,  , , ,  , ,  , , ,  , ,  , , ,  ,1,1, 0,  ,  ,  ,  ,
,  ,1,10,48,49,TX,14,4814,4914,TX14,3,1,0,0,0,0,5,3,1, 6,1,1,5,5,5,5, ,1,10,17,1017,19,10,17,1017,1,1,20,10,25, 20, 8,1,1021,23,
,118,2,4,1,0,7, 1,1,1,2,1,1,2, , , ,2,2,2,1, , ,1,1,1,1,2,1,1, , , , ,1, ,5, 0, 0, 1,1,1,1,3,1,1,1, 75,887, 50,887,887, 50, 85,887, 50,
50, 45,3,90,1,1,5,3,5,5, , ,8,3,5,1,1,1,5, ,8,3,1,3,1,5,2,5, , , , , , , , , , ,2, ,3,3,1,1,3,3, ,3,5,3, ,1, , , , , , , , , , , , , , ,
,8,8,1,5,1,2,1,3,5,90,90,  , ,7,80,80,  , 8, 8,  ,3,5,5, ,5, , ,5, , , ,5, ,3, ,5, ,5, ,3, ,5, ,5,1,8,8, , , , , , , , , , , , , , , ,
, , , , , , , , , , , , , ,1,1, ,1,1,1, , , ,  , , , ,1, ,3, 3,1980, 22,5, ,16, , 0,4,  1,  1,5, ,  , , 2,  , , 2,32,  0,   ,  0,
,5,5,5,5,5,1, ,3, 2, 5,5,5,5,5,1,5,5, ,1,2,3,3,1,3,3,5,1,      ,22,3,5,5,5,5,5,5,5,5,5,5,5,5,5,1,5,18,1,1,15,  ,1,5,5,5,5,
,1,11,18,1118,13,11,18,1118,1,1,20,12,11, 25, 4,1,269,0,0,0,9,9,99,0,3,  ,
,2,1,2,1,2,2,2,2,1,2,1,2,2,2,2,2,1,1,1,2,1,2,1,2,2,21,65,Corby Windham              ,35,1,Ron Paul                   ,34,1,
 ,  , ,                    ,                         , , , ,Ron Kirk              ,1,1,John Cornyn         ,2,1,                    , , ,
     ,34,5, 32, 68,  ,               ,                    ,2,5,43,55,  ,0,5,2,5,Kay Bailey Hutchison   ,12,2,Phil Gramm
,28,1,Phil Gramm                  ,28,1,5,5,5,  ,5, ,5,5,5,5,5,5,5,5, ,1,5,1,1,0,  ,    ,3, ,  , , , ,  ,  , , ,  , , , , ,  ,  , ,  , ,5, ,
 , ,5, ,  ,3, 50,887, 50,   ,887,887,    ,   , 50, 50, 75, 50, 50, 50, 40, 65,888, 45, 50, 30, 50,888, 50, 35, 40, 50, 50,888, 40, 50, 50,
50, , , , ,2,3,4,2,8,3, , , , , ,8, , ,8,8, , , , , ,5,5,5, ,1,1,1,3,3,2,2, ,3, ,3,3,3,3,3, ,1, ,1,8,8, ,2, ,1, , , ,3, ,3, ,1, ,1, ,3,
,3, , , ,3,90,3,5,5,8,8,5,5,3, , , , ,1, ,1,1, ,1,8, ,8,8, ,8,5,5,1,5,1,5,3,5, , ,3,5,3,3,1,3,5, , , , , , ,
,5,5,5,1,5,5,3,3,3,1,1,2,1,5,3,3,2,3,2,   ,  1,5,5,5,3,3,3,2,4,4,3,1,4,5,5,5,5,5,1,5,5,5,5,5,5,5,5,5,5,5,   ,    ,0,0, , , , , , , , , ,
, , , , , , ,  ,  ,   , ,    , , , , , , ,  , , , , , ,   , , , , , , ,  ,  , , ,  ,  , , , , , , , , , , , , , , , , , , , , , , , , , , , ,
, , , , , , , , , , , , , , ,   ,    , , , , , , ,  ,  ,  ,   ,   ,   ,  ,  ,  , , , , , , , , , ,   , , ,   ,    ,  , , , , , , ,
```

Figure 6: Screenshot of a file containing the result data set from one of the surveys.

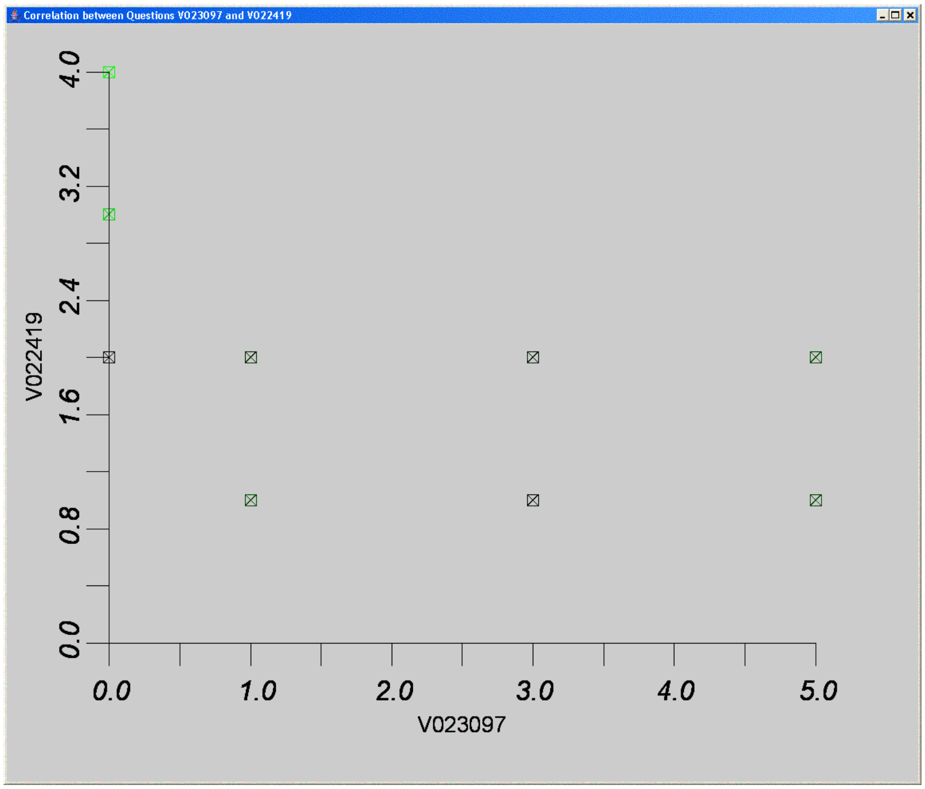


Figure 7: Frequency of answers in the augmented scatterplot as represented through colour.

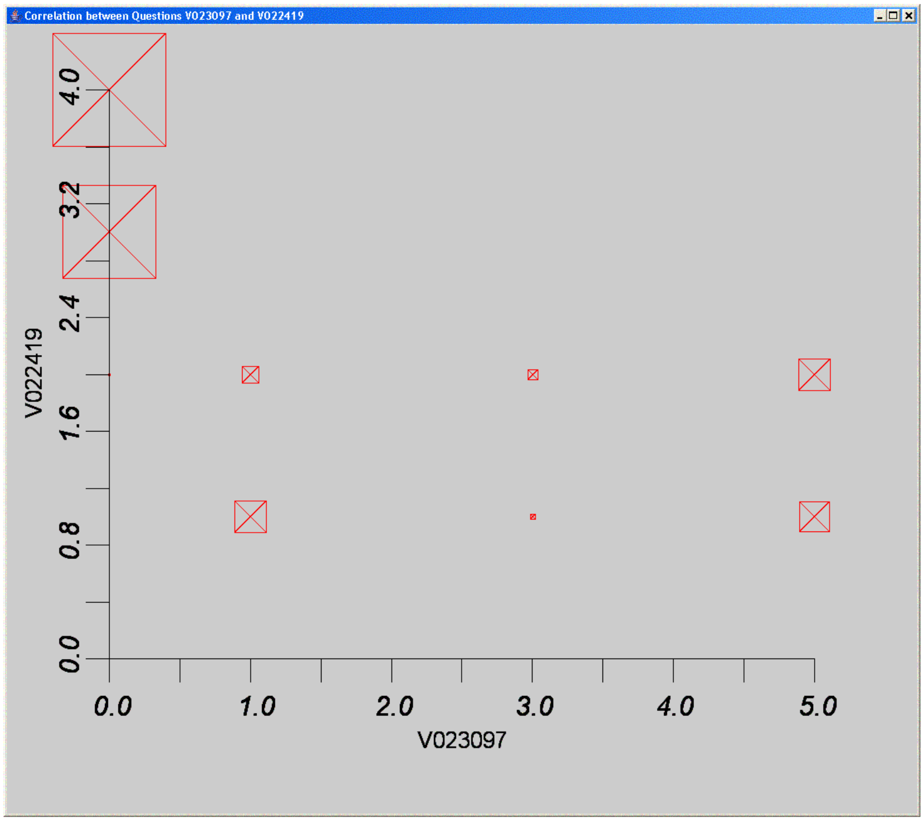


Figure 8: Frequency of answers in the augmented scatterplot as represented through size.

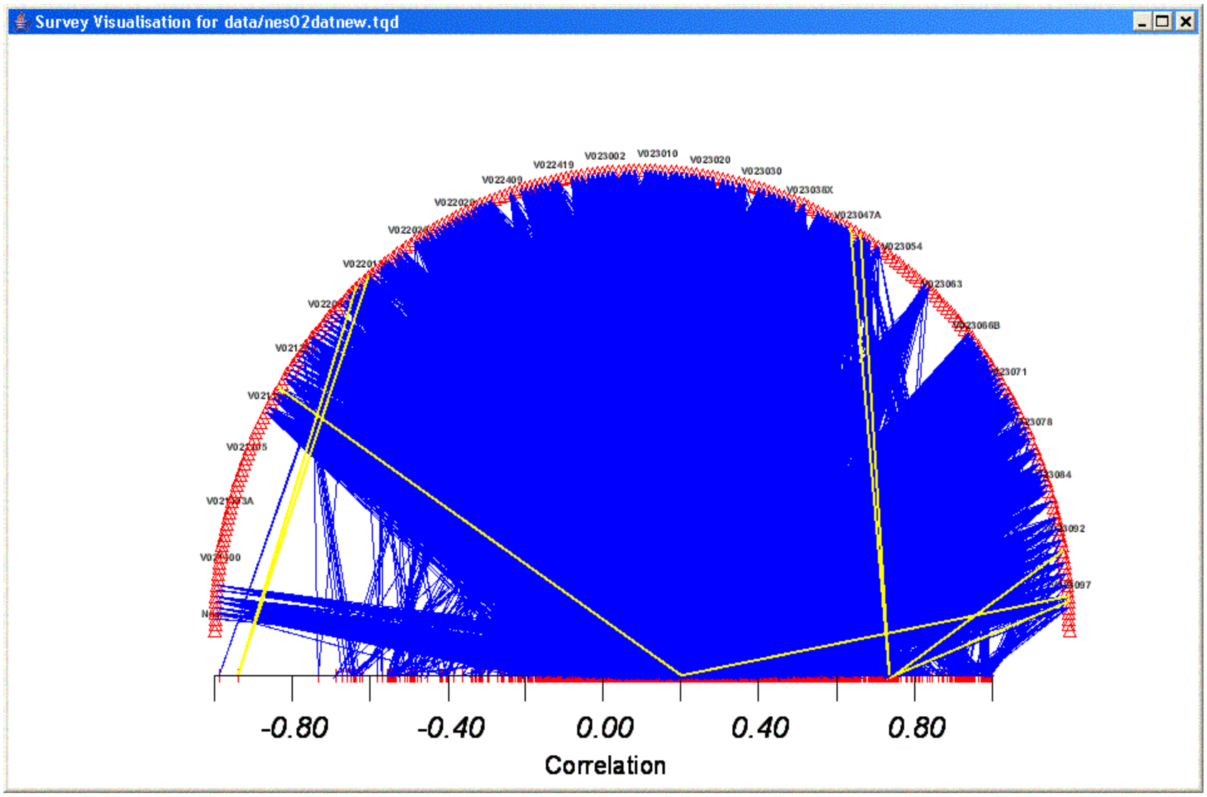


Figure 9: Unfiltered display. Multiple guide-lines are selected.

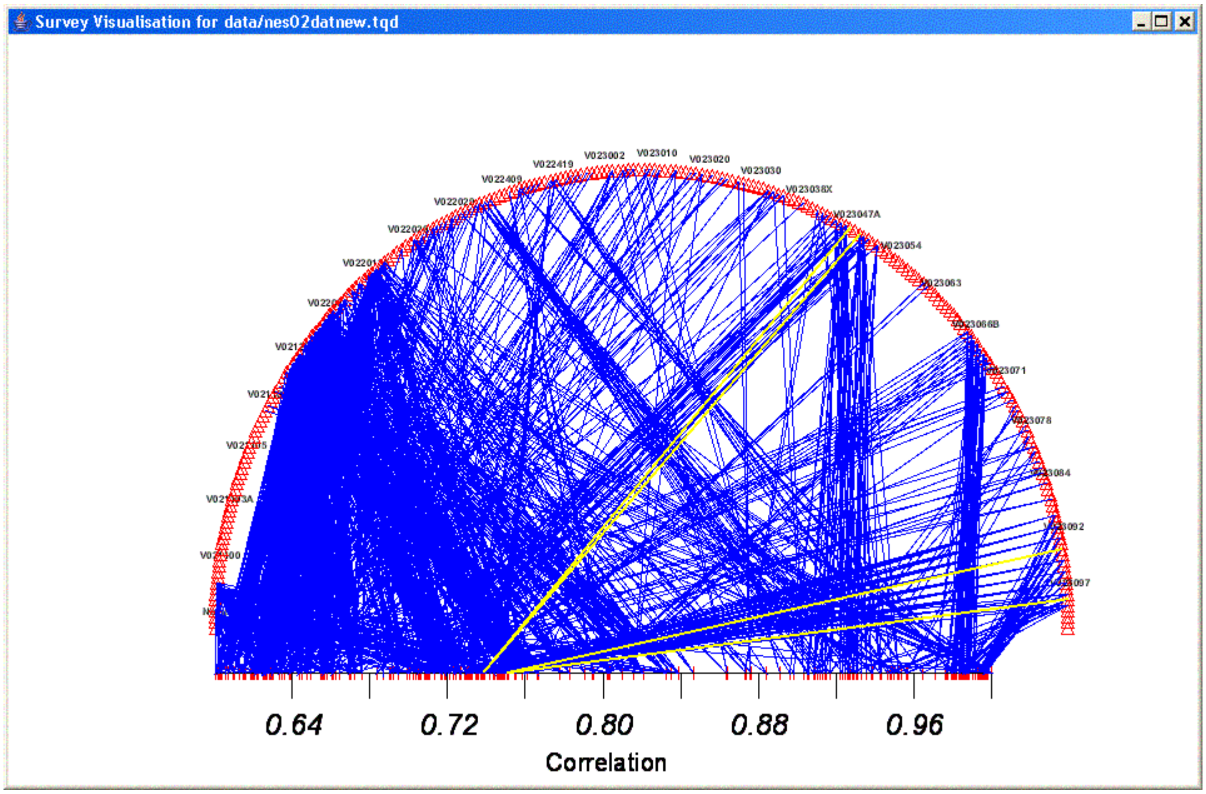


Figure 10: A filtered display of Figure 9. The guide-lines that fall outside the selected range are not displayed.

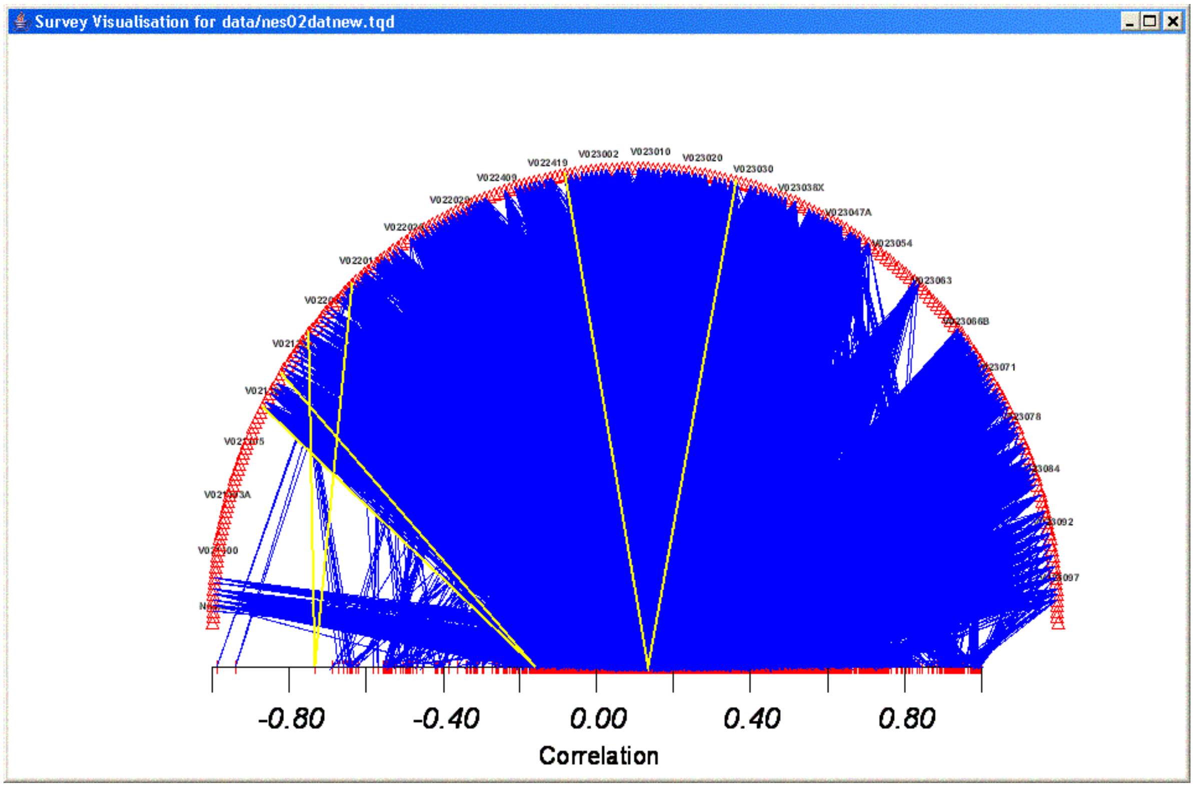


Figure 11: Full range of correlations.

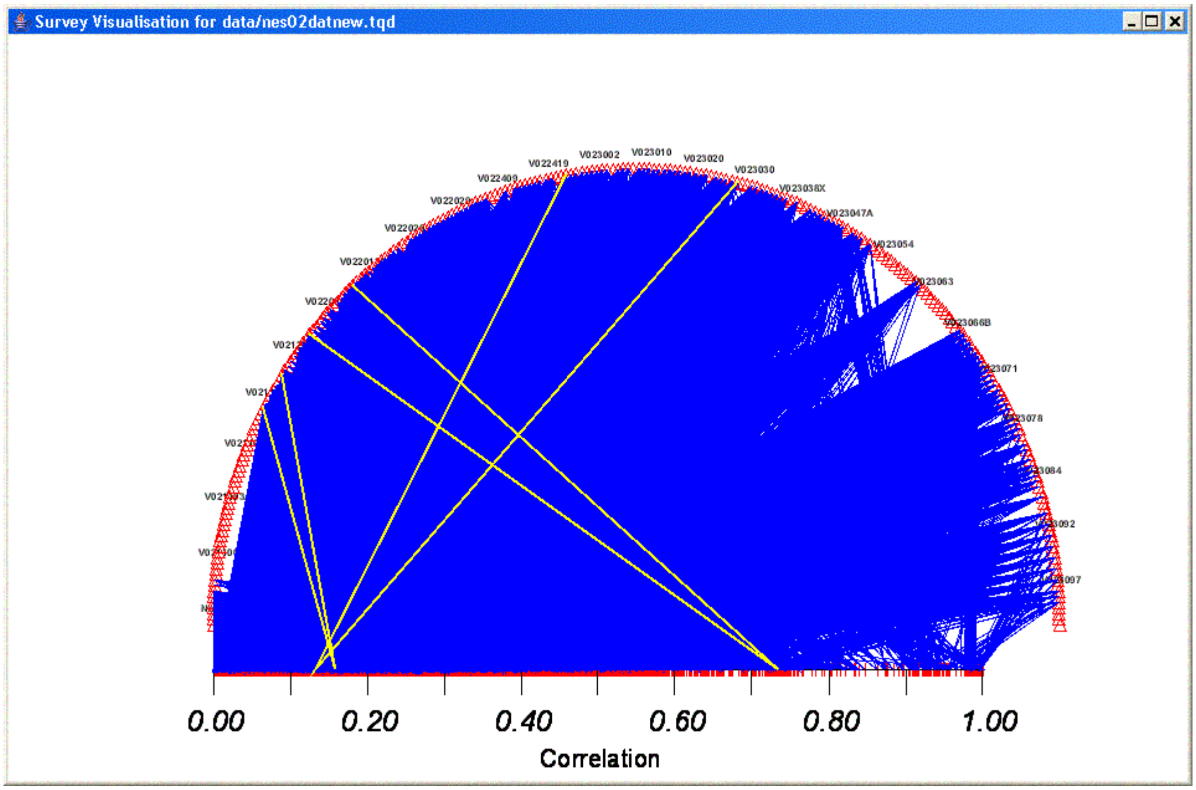


Figure 12: Absolute correlations. Guide-lines for the negative correlation points in Figure 11 flip to the other side.