Extending the Heatmap Matrix: Pairwise Analysis of Multivariate Categorical Data

David Trye
Computer Science
University of Waikato
Hamilton, New Zealand
dgt12@students.waikato.ac.nz

Mark Apperley
Software Engineering
University of Waikato
Hamilton, New Zealand
mark.apperley@waikato.ac.nz

David Bainbridge
Computer Science
University of Waikato
Hamilton, New Zealand
david.bainbridge@waikato.ac.nz

***Abstract*—Analysts are often interested in understanding the association between variables within a dataset. This paper describes a set of techniques for augmenting the Heatmap Matrix, which represents pairwise intersections of categorical variables. The proposed extensions include adapting the design and layout of the matrix to enhance its readability, expanding the number of metrics that can be presented, displaying matching records in a coordinated table view, and embedding the Chi-square test of independence. These features are demonstrated on two datasets using the empirical prototype that has been developed.**

***Index Terms*—categorical data, discrete data, heatmap matrix, multidimensional data, contingency tables, cross-tabulation**

I. Introduction

Categorical variables are widely used in real-world datasets across a multitude of domains, ranging from business to biomedical science [1]. However, visualisation techniques for categorical variables have received limited attention compared to those for continuous data [2], with relatively few techniques supporting the exploration of more than a handful of categorical variables at the same time [3]. Consequently, there are clear opportunities for advancing the state-of-the-art in categorical data visualisation, including developing novel techniques and improving upon existing ones. This paper adopts the latter approach, contributing a set of extensions for enhancing the readability, functionality and scalability of the *Heatmap Matrix* [4], [5], which represents multiple categorical variables by breaking them down into pairwise relations. The proposed extensions collectively provide more nuanced insights into the association between categorical variables, enabling the viewer to detect patterns at both a local and global level that might otherwise be missed.

Given the focus of this paper, a more detailed description of the heatmap matrix [4] is in order. This technique provides a concise visual summary of all possible two-way contingency tables for a given set of categorical variables. The plot is a matrix of heatmaps whose rows and columns are categories grouped by variable, such that each heatmap 'panel' relates to a distinct pair of variables, and each 'cell' represents the intersection of two categories. To aid readability, the categories are ordered consistently along both axes. In previous work, the heatmap matrix has only been used to show the frequency of occurrence of the corresponding categories; however, as this paper will show, other information can also be fruitfully encoded. The technique facilitates quick identification of salient patterns and values, accentuating outliers, as well as large numbers of cells with very low or high frequencies. Since each heatmap can be taken as an independent unit, patterns can be discovered at both a local (panel) and global (matrix) level. For instance, a user may wish to analyse each panel one at a time, locate the highest and lowest values across the entire matrix, or focus on specific categories or variables of interest by isolating particular rows or columns of the matrix.

While the original heatmap matrix was static, the authors describe several interactive enhancements in later work [5]. These include: four methods for reordering categories according to different seriation algorithms; filtering based on both Spearman's correlation coefficient and association rules; bucketing of continuous variables by producing bins of equal width or frequency [6]; and the choice of a local or global colour mapping to highlight patterns within or across the matrix, respectively. However, without a publicly available prototype or a more detailed description of the user interface, it is not clear how these features are operationalised. This paper focuses predominantly on novel features that are intended to supplement, rather than replace, those mentioned in [5].

In terms of scalability, the size of a heatmap matrix is proportional to the number of categories in the display, with higher-cardinality variables occupying more space. The number of data items (records) has no bearing on the dimensions of the visualisation, as this is conveyed through the colour of the heatmap. Although it is theoretically possible to generate a heatmap matrix for any number of categories or variables, the visualisation is in practice restricted by the screen resolution. When drawing heatmap matrices, datasets comprising multiple high-cardinality variables pose a significant challenge [5], which is a key consideration for this work.

II. Related Work

Most techniques for visualising categorical data—including the heatmap matrix—are based on contingency tables [7]. Alsallakh et al. [8] classify these methods into three main types: frequency representations, deviation representations, and intermediate representations. Frequency representations display the observed frequencies in a contingency table directly, typically

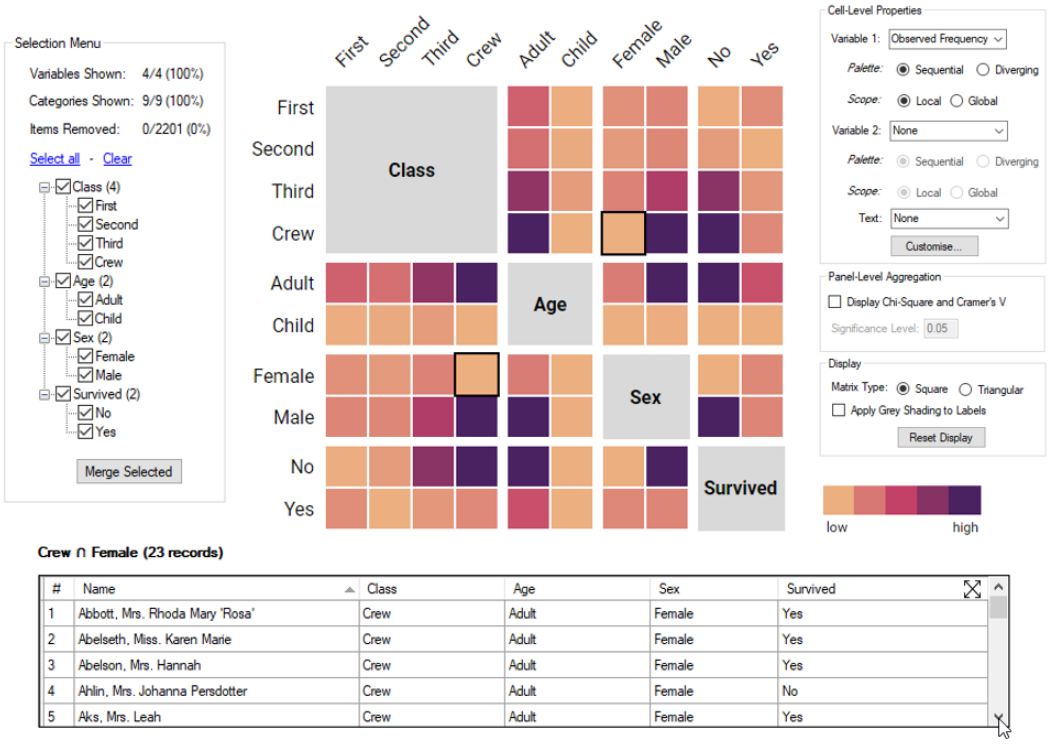

Fig. 1. Design overview of the *Heatmap Matrix Explorer*, which represents the intersection of every pair of categories in a dataset. This example shows the Titanic dataset, consisting of 4 categorical variables and 10 categories.

using an area-proportional encoding. Prominent examples include Mosaic Plots (without residual-based shading) [9] and Parallel Sets [10], which are described in further detail below. The heatmap matrix also falls under this category, though it uses colour rather than area to encode frequency, sacrificing precision for greater scalability [5].

Deviation representations visualise differences between observed and expected frequencies. Examples include Association Plots [11], Sieve Diagrams [2] and the dot-based Contingency Wheel [12]. Section V shows how the heatmap matrix can be extended to support deviations as well as frequencies; the two approaches are complementary, not mutually exclusive.

Finally, intermediate representations convert categories into numerical values before visualising them. Correspondence Analysis (CA) identifies associations between the cells in a contingency table by projecting points into a low-dimensional space [13].

Matrix-based visualisations for representing every pair of variables in a dataset constitute another class of relevant techniques. These include the Scatterplot Matrix (SPLOM) for continuous data [14], together with its enhancements [15]–[17]; the Mosaic Matrix for categorical data [18]; and the Generalised Pairs Plot [19] and GPLOM [20] for dealing with mixed data types (first suggested in [18]). These latter representations use different kinds of charts depending on the variable types that are present in each pair. While there is a range of options for representing two categorical variables (including Mosaic Plots [9], Fluctuation Diagrams [21] and Faceted Bar Charts [22], as posited in [19]), GPLOMs use a heatmap for simplicity. However, since GPLOMs use a fixed panel size for all pairings, regardless of variable cardinality, these heatmaps are not always readable.

Mosaic Matrices [18] are specifically designed for visualising pairs of categorical variables. In this representation, variables are crossed among themselves in a matrix, and a Mosaic Plot is shown in each of the resulting panels, with variable names displayed along the main diagonal. The size of each tile in a Mosaic Plot is proportional to the cell frequencies, and the tiles are often also coloured according to Pearson residuals [23], yielding a blended frequency/deviation representation. These residuals provide an indication of the goodness-of-fit of the model of independence, and show which values occur more or less often than expected. Similar to GPLOMS, however, Mosaic Plots become difficult to read when visualising categorical variables with high cardinality [20]. Furthermore, Mosaic Matrices are generally restricted to visualising three or four variables at a time, due to space limitations [18].

Although originally intended for visualising hierarchies of variables, Parallel Sets [10] can also be used to show pairwise relations between categorical variables [24]. However, as the number of variables increases, so too does the number of repeated bands or small multiples needed to explicitly capture all possible relationships.

Reflecting on the various strengths and weaknesses of

these techniques, the heatmap matrix offers a more compact alternative for exploring pairwise associations, on which this paper seeks to build.

III. Empirical Prototype

The following sections describe the fundamentals of the prototype that has been developed to extend the capabilities of the heatmap matrix. An overview of the *Heatmap Matrix Explorer* is given in Fig. 1, showing the familiar Titanic dataset [25]. The design consists of four components: the *Matrix View* (centre) containing the heatmap; the *Selection Menu* (left-hand side) for filtering the data and merging categories; the *Main Menu* (right-hand side) for customising the heatmap; and the *Linked Table View* (bottom) showing underlying data for selected cells. While the Titanic dataset is used as the primary example throughout the paper, a second, more complex dataset is examined in Section VIII to provide a clearer indication of the technique's scalability.

IV. Matrix View

At the heart of the prototype is the *Matrix View* where the main visualisation is displayed. While this view is similar to the original heatmap matrix [4], there are several key differences. In previous work, the design had a black background and variables were separated with grey grid lines; in contrast, the new design uses a white background and replaces these grid lines with white space. This helps to achieve a minimalist aesthetic that is easier on the eye [26]. The variable groupings can be perceived solely through the spacing between panels, in accordance with the Gestalt Law of Proximity [27]. In addition, all cells have been given a thin white border to help distinguish individual values. Category labels for columns are rotated 45 degrees for readability, and a legend has been added to indicate the exact range of values present in the heatmap (for global mappings) or the general direction of the encoding (for local mappings). Like in the original design, all cells are square-shaped, so as not to privilege one axis (variable) over the other.

Another point of difference is the main diagonal of the matrix. In the updated design, intra-variable cell frequencies are replaced with a single grey 'box' showing the name of the corresponding variable, akin to how variables are labelled in the Mosaic Matrix [18]. The motivation for this is two-fold. First, it removes redundant and potentially distracting information. At least half of the cells in diagonal panels represent intersections that are structurally impossible, assuming the categories within each variable are mutually exclusive. If this is the case, the only cells that can occur represent categories' marginal frequencies; however, such data is univariate rather than bivariate, and thus has a different interpretation from the rest of the matrix. Of course, information regarding individual category frequencies may still be useful, but this can be displayed in a less obtrusive manner, by means of a tooltip; see Fig. 3. The second reason for this change is that labelling variables along the main diagonal frees up space, since the variable names are included within the matrix itself and do not need to be added as external row or column labels.

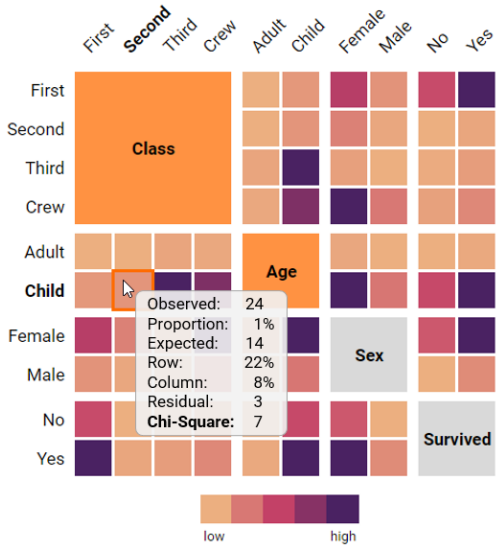

Fig. 2. Example of a cell-level tooltip and associative highlighting. The user is currently hovering over the cell representing children in second class. The heatmap itself shows cell Chi-square values (sequential palette, local scope) for the Titanic dataset.

Two additional features supported by the prototype are tooltips and associative highlighting[1], which work together to provide "details-on-demand" [28]. There are two different kinds of tooltips, depending on whether the user hovers over a cell (Fig. 2) or one of the variable boxes along the main diagonal (Fig. 3). In the former case, the tooltip displays rounded values for all seven supported metrics (Section V-A), including observed frequency. Bold text is used to indicate the metric(s) that are currently encoded in the heatmap. If the user hovers over one of the variable boxes, the tooltip instead shows the distribution of category frequencies in the form of a bar chart. Categories are sorted in descending order of frequency, regardless of their position in the heatmap.

Associative highlighting helps the user to see which categories have been selected, and which variables they relate to. When the user hovers over a cell, not only does a tooltip appear, but the cell is given an orange outline, and the two related variables are highlighted orange. The corresponding row and column labels are emphasised in bold, enabling the user to accurately pinpoint their position within the matrix, which is not trivial for more complex datasets.

V. Main Menu

The *Main Menu* allows the user to customise the heatmap matrix in simple yet powerful ways. There are three sub-menus that control different aspects of the visualisation: cell-level properties, panel-level aggregation and general appearance of the display. The first two sub-menus cannot be used at the same time (they provide different modes for exploring the

[1]This term is used in a different sense from [20].

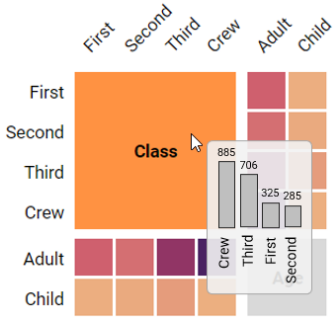

Fig. 3. Tooltip showing a bar chart of category frequencies for the "Class" variable.

matrix), whereas the third sub-menu is compatible with both of the others, and thus always available. All features supported by these menus are novel, except for the scope setting, which was proposed in [5]. Together, these controls provide a diverse range of complementary views that encourage users to examine the data from fresh and varied perspectives.

A. Cell-Level Properties

In the same way that contingency tables can display different measures of association, the heatmap matrix is not confined to visualising only observed frequency. The cell-level menu supports five additional metrics, from row percentages to Pearson residuals, which can be used in combination to provide additional insights and supporting evidence about the nature of association between categorical variables. Within the *Heatmap Matrix Explorer*, users must select either one or two metrics to control the colour of the heatmap, and can optionally specify a third metric (or one of the same metrics) to annotate the cells with the corresponding numerical values.

For metrics defining the colour of the heatmap, the user must specify a colour palette (either sequential or diverging) and a scope (either local or global). Sequential colour palettes accentuate high values (or low values if the scale is reversed), whereas diverging colour palettes emphasise values at both ends of the spectrum. When the user chooses a local scope, the colour of each panel is scaled according to its local minimum and maximum values, rather than the extremities across the entire matrix. In general, a local mapping seems more appropriate than a global one, since, unless all variables happen to have the same cardinality, the panels in the heatmap will contain different numbers of cells, and individual cells within smaller panels are more likely to draw higher counts.

The metrics that appear in each drop-down menu are detailed below.

- *Observed Frequency* shows the frequency of occurrence in each cell, which is the same information encoded in the original heatmap matrix [4]. If all categories are shown, the cells in each panel sum to the total number of data items, and each row or column sums to the category frequency. This is the default setting, useful for obtaining a preliminary overview of the data but limited in terms of measuring associations.
- *Expected Frequency* displays the quantities that would be expected in each panel if there were no association between the two variables. This is calculated by multiplying each cell's row total by its column total, then dividing by the total number of observations.
- *Row Percentages* and *Column Percentages* display the relative contribution of the observed frequency of each cell to the *local* row or column total, respectively. These metrics reveal how the categories belonging to one variable are distributed with respect to the other. For *Row Percentages*, each cell shows $P(X \mid Y)$, where X is the category on the x-axis and Y is the category on the y-axis. The correct interpretation is *what percentage of* Y *is* X? *Column Percentages* shows the inverse, i.e., $P(Y \mid X)$. The matrix generated by either metric is the transpose of the other.
- *Pearson Residuals* measure, for each cell, the magnitude and direction of the deviation from independence, adjusted for the expected variability. They are calculated by dividing the difference between the observed and expected frequencies by the square root of the expected frequency. This provides a quick visual summary of over- and under-represented pairs of categories. Cells with large residuals (in either direction) may be indicative of patterns or relationships between two variables that warrant further investigation. Pearson Residuals are best suited to a diverging colour palette (preferably one that is continuous [29]), since this emphasises both positive and negative residuals.
- *Cell Chi-Square Values* (shown in Fig. 2) denote the individual contribution of each cell to the overall Chi-square (χ^2) test statistic (see Section V-C). The cell values are calculated by taking the squared difference between the observed and expected frequencies, and dividing by the expected frequency. A cell Chi-square value less than one means that the observed and expected frequencies are reasonably close to each other, whereas values much larger than one indicate a disparity between the two. While Pearson residuals show similar information, examining the individual cell values can help to identify outliers or unusual patterns that may not be apparent from the residuals alone, and vice versa.

One salient design consideration for any heatmap is the colour palette, which affects the range of values that can be seen [26], [30], [31]. The *Heatmap Matrix Explorer* uses sensible default colours, including Seaborn's [17] perceptually linear "flare" colourmap for a single sequential metric, a blue-white-red palette for a single diverging metric, and Cynthia Brewer's nine-class bivariate maps[2] when two metrics are selected. Since the bivariate heatmap only has nine distinct values, it sacrifices precision for general readability. Nevertheless, exact values are still accessible via interactive tooltips. An example is shown in Fig. 4, which simultaneously encodes observed frequency and Pearson residuals, such that darker

[2]http://www.personal.psu.edu/cab38/ColorSch/Schemes.html

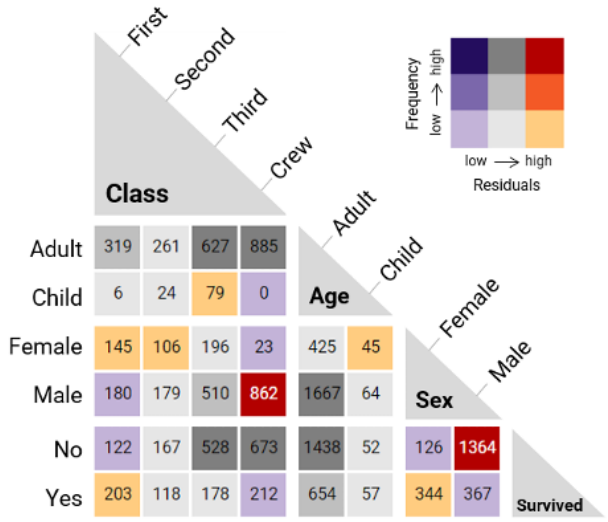

Fig. 4. Triangular heatmap matrix showing both observed frequency (sequential palette, local scope) and Pearson residuals (diverging palette, global scope) for the Titanic dataset. The text labels also show observed frequency.

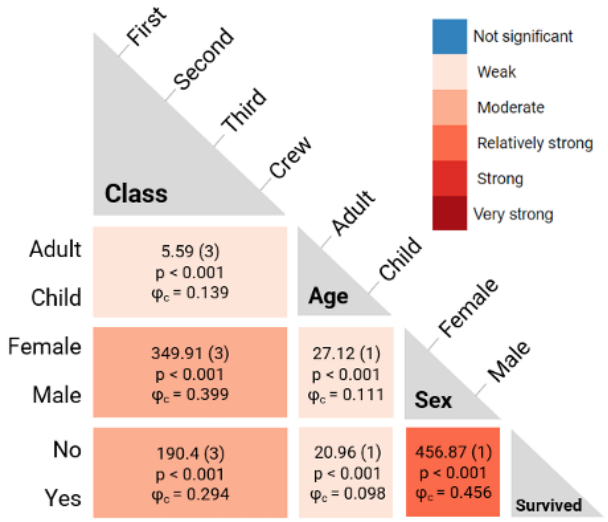

Fig. 5. Panel-level test results for the Titanic dataset, including the Chi-square statistic, degrees of freedom, p-value and Cramér's V. Cell colour is proportional to the strength of the association.

cells represent more frequent values, and blue, grey and orange are used for negative, neutral and positive residuals, respectively. This encodes similar information to a Mosaic Matrix, minus the alignment of tiles, in a more scalable form.

Unlike its predecessor [4], by default, the *Heatmap Matrix Explorer* does not display numerical values in each cell. Removing the text labels makes it easier to glean general patterns, while still allowing users to access tooltips. However, text labels may still be useful in some scenarios, such as in static (e.g., print) environments. The text drop-down menu supports the same metrics outlined above, except observed frequency is split into "Counts" and "Proportions", for which the colours are the same. Proportions display the joint probability of the corresponding categories, $P(X \cap Y)$, by dividing the counts by the total number of observations.

B. Display Settings

Among the general display settings is a "Reset" button that restores the default settings and reverts to the original dataset. There are also inputs for changing the type of matrix and the appearance of the row and column labels, which are explained below.

The prototype allows the user to switch between a square or triangular matrix (see Fig. 2 and Fig. 4 for examples of each). In a square matrix, all pairs of categories are represented twice, once on each axis, whereas a triangular matrix removes this redundancy. As with SPLOMs, a square matrix can be helpful for identifying patterns and trends related to particular variables of interest. This is because the user can focus on a single horizontal or vertical band of the matrix, rather than having to divide their attention between a mixture of rows and columns, while simultaneously transposing parts of the display.

A triangular matrix, on the other hand, makes the display less cluttered and safeguards against novice users misreading the visualisation, e.g., by thinking that each value occurs twice. The outer variables in a triangular matrix are special cases as all panels appear in a straight column or row, like they would in a square matrix. As a result, the left-most variable only needs column labels and the right-most variable only needs row labels.

Regarding the appearance of labels more generally, a tick-box enables the heatmap to be drawn with or without shaded row and column labels. If applied, alternating shades of grey are used to distinguish categories belonging to adjacent variables, as shown in Fig. 6.

C. Panel-Level Aggregation

Two widely used statistical techniques for analysing categorical data are the Chi-square (χ^2) test of independence and Cramér's V (φ_c). These techniques are often paired together to establish 1) whether there is a significant association between two variables and 2) how strong it is. Fig. 5 shows how this information can be integrated into the heatmap itself, providing a concise visual summary of the extent to which different variables are associated. This feature is inspired by correlation matrices for quantitative data [32], especially those which visualise statistical significance [33]. Note that, unlike other views, the cells within each panel are merged, because both measures apply to the variable level rather than the category level.

Importantly, test results are only calculated and displayed if the four basic Chi-square test conditions are thought to be satisfied. The first requirement is that both variables must be categorical. Since only categorical variables are featured in the heatmap matrix (with the possibility of binning continuous variables before the visualisation is generated), this requirement is automatically fulfilled. Secondly, observations must be independent. When the user first ticks the box to "Display Chi-Square and Cramér's V", a dialog box appears asking them

to verify whether this is the case. This is the only check that cannot be automated, since it is highly context-dependent. If the user selects "No" (i.e., observations are not independent), a further message appears informing them that, unfortunately, the test cannot be applied. Otherwise, test results are shown for panels that satisfy the two remaining conditions, namely that categories within each variable are mutually exclusive, and that expected frequencies exceed one in all cells and are at least five in 80% of cells.

If the user confirms independence of observations and the remaining test conditions are satisfied, the corresponding panel is coloured either red or blue, depending on the test result. Red indicates a significant result, whereas blue does not. The shade of red is proportional to the strength of the association, as measured by Cramér's V: a number between 0 and 1, with larger/darker values indicating a stronger association. For completeness, the Chi-square statistic, degrees of freedom, p-value and Cramér's V are all reported in the corresponding panel. Users can also change the significance level in the text box from its default value of 0.05; this updates the test results accordingly. The legend is interactive, such that hovering over one of the values isolates all variable pairs with the corresponding effect size.

If any of the test conditions for a pair of variables is violated, the panel is coloured grey. An error message explaining the reason why the test result was not valid is shown in the tooltip. This is helpful even for datasets where majority of the panels are grey, because it shows the user that the Chi-square test is not an appropriate technique for such data, while perhaps still revealing a handful of associations that are significant.

Embedding Chi-square test results into the plot in this way has a number of benefits: it removes the burden of manual computation (which is particularly onerous for datasets with many variables), visually reinforces correct interpretations, and enables all relevant data to be conveniently displayed in one place. Furthermore, the results in this view can be effectively coupled with the cell-level Chi-square values and Pearson residuals discussed in Section V-A [34]. For instance, Fig. 5 shows that there is a relatively strong association between the variables "Sex" and "Survived", and the cell-level metrics, including Fig. 2, suggest that this is due to more females surviving than would be expected by chance, and more men dying. This aligns with the societal norm of prioritising the rescue of "women [and children] first".

VI. Linked Table View

The linked table view, shown in Fig. 1, connects the heatmap with the underlying data. The user can click on a cell to see the corresponding data items in the table beneath the matrix. Upon being clicked, the cell is given a black border to show that it has been selected. By default, all variables are displayed in the table, with any unique, 'ID-like' variables being shown to the immediate left of all others. For example, in Fig. 1, the user has clicked on a cell representing female crew members and the corresponding records, including people's names, are displayed in the table. The user can navigate with the scroll bar or expand the table to view records that are not currently visible. Additionally, the table columns can be hidden or reordered. The number of rows in the table matches the cell's observed frequency; there are 23 female crew members and thus 23 records in the table. This feature is most useful if the dataset contains one or more 'ID-like' columns, and if a large proportion of cells have relatively low counts, so that the information presented in the table can be readily absorbed.

While not currently supported, it would be possible to generate supplementary visualisations from the conditional table data. One could imagine hovering over one of the column headings to display a bar chart of the number of occurrences in the table of each category from the corresponding variable. For example, hovering over the "Sex" column would show how many female crew members survived, $P(Yes \mid Female \cap Crew)$, and how many died, $P(No \mid Female \cap Crew)$.

VII. Selection Menu

For complex datasets with a large number of categories and variables, it may not be feasible to view everything at once. A better approach might be to break down the dataset into smaller units of interest, and rotate among these. As shown in Fig. 1 and Fig. 6, the *Selection Menu* consists of an expandable list of checkboxes, with variables at the top level (whose cardinalities are indicated in parentheses) and categories nested inside them. The user can click on the checkboxes to show or hide variables in the matrix. Variables that are currently visible are shown with a black tick, and those filtered out with a blue box. It is also possible to show, hide *or exclude* individual categories, with three clicks required to return to each state. Excluded categories are shown with a red cross icon. The distinction between hiding and excluding a category is that the former simply removes it from the display (without affecting the rest of the matrix), whereas the latter removes all data items associated with that category, likely resulting in changes to other panels. For instance, removing children from the Titanic dataset would update *all* panels to only include information about adults, effectively resulting in a conditional query: $P(X \cap Y \mid Adult)$, where X and Y are the variables on either axis. The advantage of having the checkboxes is that users can re-select categories that they previously hid or excluded. "Select all" and "Clear" links are also available to expedite variable and category selections.

Filtering via the *Selection Menu* is primarily intended for analysing datasets with dozens of variables and/or categories (e.g., census data). This feature is not necessary for relatively simple datasets like the Titanic dataset, where all of the variables and categories can be visualised at once. However, even in such cases, the ability to exclude variables is helpful for visualising conditional queries. Overall, the addition of this menu increases the scalability of the heatmap matrix technique, albeit by requiring the user to work with manually defined subsets. Similar functionality could be added to other pairwise techniques for categorical data, such as the Mosaic Matrix [18].

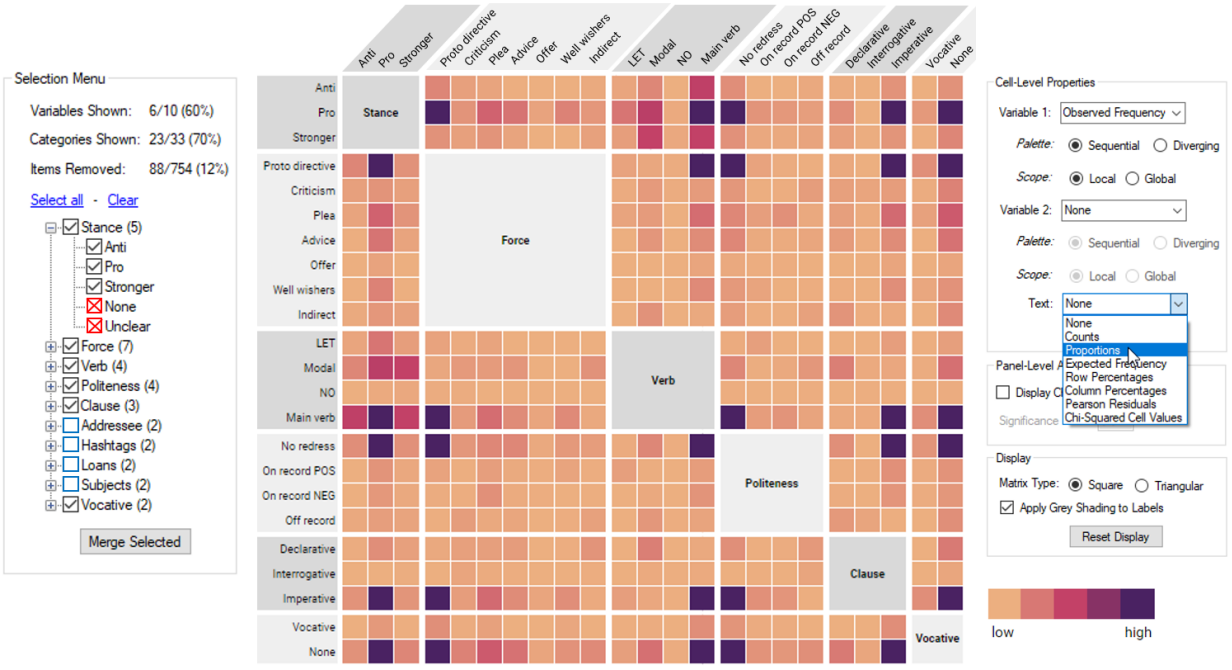

Fig. 6. A more complex example of a heatmap matrix, showing information about Covid directives on Twitter. Some variables have been hidden, and two categories have been excluded from the data, as indicated in the *Selection Menu* on the left-hand side.

Summary statistics at the top of the menu indicate the proportion of variables and categories visible at any given time, as well as how many data items have been excluded. It might also be helpful to report statistics such as "Categories shown for *selected* variables" (to exclude variables that the user deems irrelevant) or the "(Average) number of categories per variable". These could all be visualised as stacked bar charts or histograms, rather than being given as text labels.

The *Selection Menu* also provides a mechanism for manually re-ordering the data in the heatmap, by dragging-and-dropping the labels. The user can move variables, as well as the categories within them. The order of rows in the matrix mirrors the top-to-bottom ordering in the list of checkboxes, and columns are ordered in the same way from left to right. For instance, moving "Class" beneath "Age" in Fig. 1 would make "Age" the top/left-most box, and dragging the category "Crew" above "First" would make "Crew" the top/left-most category within the "Class" portion of the matrix.

A final operation supported by the menu is merging existing categories. The user can use Shift-click and Ctrl-click to select multiple categories for a particular variable, then right click on one of them to merge those categories. They are then given the option to rename the newly formed category. The number of categories is automatically reduced to reflect the number that remain, with undo and redo functionality supported in case the user makes a mistake or wishes to revisit a previous state.

VIII. Covid Directives Dataset

All the examples given so far concern the Titanic dataset. Fig. 6 shows a more complex example of a heatmap matrix, illustrating a linguistics dataset comprising ten categorical variables [35]. The data consist of directives used in tweets featuring the hashtag "#covid19nz" (e.g., "Stay home!"). This dataset was compiled to examine pragmatic and syntactic variables in relation to the stance of directives towards COVID-19 government measures in New Zealand, during the first nation-wide lockdown. The user has hidden four variables from the display, and removed two categories from "Stance": "None" and "Unclear". This has resulted in 88 of the 754 directives being filtered out of the heatmap. The matrix view shows that there is one dominant pair of categories (or 'flavour' of directive) in each and every panel. For instance, the panels for "Stance" and "Politeness" show that those in most agreement with the status quo ("pro") were also least concerned to mitigate their directive with polite markers ("no redress"). "Stance" and "Verb" exhibit greater variation than any other pair of variables, with main verbs and modal verbs being relatively common across all three stance categories.

IX. Limitations

The *Heatmap Matrix Explorer* has some limitations that need to be acknowledged. First, it does not incorporate several of the useful interactive features described in [5], such as automated methods for sorting the matrix, which would be useful for revealing structural patterns, or the ability to bin continuous values. Second, a lot of the design decisions are based on the authors' subjective preferences and require more comprehensive user testing. For instance, the thin white borders around cells might actually be a distraction for perceiving patterns within and between different panels. Third, displaying cell Chi-square values from the drop-down menu for an invalid Chi-square test may be problematic, and there is currently

nothing to safeguard against this. Furthermore, the Chi-square test and Cramér's V are not well suited to ordinal data, as they do not consider ordering information. The datasets in this paper contain mostly nominal variables, but a Spearman correlation or Kendall's Tau would be more appropriate for panels involving strictly ordinal data. While there are, in fact, several alternative methods for analysing categorical data, the bigger picture is that such tests can be effectively embedded into visualisations to aid the viewer's understanding.

X. Conclusions and Future Work

This paper has proposed a structured set of extensions for augmenting the heatmap matrix, which are realised in an empirical prototype called the *Heatmap Matrix Explorer*. These extensions improve the readability, versatility and scalability of the heatmap matrix technique. The revised design removes non-bivariate cells, re-positions variable labels, removes dense grid lines and has a white background. Interactive drop-down menus allow the user to colour and label cells according to several metrics, including row percentages and expected frequencies. The high-level overview for the Chi-square test helps the viewer to quickly detect patterns and establish which variables have the strongest associations. Examining these findings in relation to cell-level metrics like Pearson residuals and individual Chi-square values can then provide more detailed information about specific cells driving the association. The *Linked Table View* provides a direct and convenient link to individual records, and the *Selection Menu* enables exploration of more complex datasets than was previously possible, by allowing controlled yet flexible filtering. Overall, these extensions provide greater insight into the relationships between categorical variables, by encouraging the user to explore the data from a range of perspectives, and empowering them to uncover more complex patterns in the process.

Future work could centre around turning the empirical prototype into a web-based tool that allows users to visualise their own categorical datasets, and conducting in-depth user testing. Two further avenues of inquiry are dealing with missing values, which may differ across variables, and supporting nested heatmaps for hierarchical categorical data.

References

[1] A. Agresti, *Categorical Data Analysis*, 3rd ed. John Wiley & Sons, 2013.

[2] M. Friendly, "Graphical methods for categorical data," *Proceedings of SAS SUGI*, vol. 17, pp. 190–200, 1992.

[3] R. M. Reza and B. A. Watson, "Hi-d maps: An interactive visualization technique for multi-dimensional categorical data," in *2019 IEEE visualization conference (VIS)*. IEEE, 2019, pp. 216–220.

[4] M. M. N. Rocha and C. G. da Silva, "Heatmap matrix: a multidimensional data visualization technique," in *Proceedings of the 31st Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2018.

[5] M. Rocha and C. G. da Silva, "Heatmap matrix: Using reordering, discretization and filtering resources to assist multidimensional data analysis," 2022. [Online]. Available: https://doi.org/10.13140/RG.2.2.36619.57126

[6] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *Machine learning proceedings 1995*. Elsevier, 1995, pp. 194–202.

[7] S. J. Fernstad and J. Johansson, "A task based performance evaluation of visualization approaches for categorical data analysis," in *2011 15th International Conference on Information Visualisation*. IEEE, 2011, pp. 80–89.

[8] B. Alsallakh, W. Aigner, S. Miksch, and M. E. Gröller, "Reinventing the contingency wheel: Scalable visual analytics of large categorical data," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2849–2858, 2012.

[9] J. A. Hartigan and B. Kleiner, "A mosaic of television ratings," *Am. Stat.*, vol. 38, no. 1, pp. 32–35, 1984.

[10] R. Kosara, F. Bendix, and H. Hauser, "Parallel sets: Interactive exploration and visual analysis of categorical data," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, no. 4, pp. 558–568, 2006.

[11] D. Meyer, A. Zeileis, K. Hornik, and F. Leisch, "Visualizing independence using extended association plots," *Proceedings of DSC 2003*, 2003.

[12] B. Alsallakh, M. E. Gröller, S. Miksch, and M. Suntinger, "Contingency wheel: Visual analysis of large contingency tables." in *EuroVA@EuroVis*, 2011.

[13] M. Greenacre, *Correspondence analysis in practice*. CRC press, 2017.

[14] D. B. Carr, R. J. Littlefield, W. Nicholson, and J. Littlefield, "Scatterplot matrix techniques for large n," *J. Am. Stat. Assoc.*, vol. 82, no. 398, pp. 424–436, 1987.

[15] L. Wilkinson, A. Anand, and R. Grossman, "Graph-theoretic scagnostics," in *2005 IEEE Symposium on Information Visualization*. IEEE, 2005, pp. 157–164.

[16] N. Elmqvist, P. Dragicevic, and J.-D. Fekete, "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, no. 6, pp. 1539–1148, 2008.

[17] M. L. Waskom, "seaborn: statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, pp. 3021–3024, 2021.

[18] M. Friendly, "Extending mosaic displays: Marginal, conditional, and partial views of categorical data," *J. Comput. Graph. Stat.*, vol. 8, no. 3, pp. 373–395, 1999.

[19] J. W. Emerson, W. A. Green, B. Schloerke, J. Crowley, D. Cook, H. Hofmann, and H. Wickham, "The generalized pairs plot," *J. Comput. Graph. Stat.*, vol. 22, no. 1, pp. 79–91, 2013.

[20] J.-F. Im, M. J. McGuffin, and R. Leung, "GPLOM: the generalized plot matrix for visualizing multidimensional multivariate data," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2606–2614, 2013.

[21] H. Hofmann, "Exploring categorical data: interactive mosaic plots," *Metrika*, vol. 51, pp. 11–26, 2000.

[22] R. A. Becker, W. S. Cleveland, and M.-J. Shyu, "The visual design and control of trellis display," *J. Comput. Graph. Stat.*, vol. 5, no. 2, pp. 123–155, 1996.

[23] M. Friendly, "Mosaic displays for multi-way contingency tables," *J. Am. Stat. Assoc.*, vol. 89, no. 425, pp. 190–200, 1994.

[24] H. Hofmann and M. Vendettuoli, "Common angle plots as perception-true visualizations of categorical associations," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2297–2305, 2013.

[25] R. J. M. Dawson, "The "unusual episode" data revisited," *J. Stat. Educ.*, vol. 3, no. 3, 1995.

[26] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman, "The science of visual data communication: What works," *Psychol. Sci. Public Interest*, vol. 22, no. 3, pp. 110–161, 2021.

[27] K. Koffka, *Principles of Gestalt Psychology*. Lund Humphries, 1935.

[28] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings 1996 IEEE Symposium on Visual Languages*. IEEE, 1996, pp. 336–343.

[29] A. Zeileis, D. Meyer, and K. Hornik, "Residual-based shadings for visualizing (conditional) independence," *J. Comput. Graph. Stat.*, vol. 16, no. 3, pp. 507–525, 2007.

[30] N. Gehlenborg and B. Wong, "Heat maps," *Nat. Methods*, vol. 9, no. 3, p. 213, 2012.

[31] T. Munzner, *Visualization analysis and design*. CRC press, 2014.

[32] M. Friendly, "Corrgrams: Exploratory displays for correlation matrices," *Am. Stat.*, vol. 56, no. 4, pp. 316–324, 2002.

[33] T. Wei, V. Simko, M. Levy, Y. Xie, Y. Jin, J. Zemla *et al.*, "Package 'corrplot'," *Statistician*, vol. 56, pp. 316–324, 2017.

[34] M. L. McHugh, "The chi-square test of independence," *Biochemia medica*, vol. 23, no. 2, pp. 143–149, 2013.

[35] J. Burnette and A. S. Calude, "Wake up New Zealand! Directives, politeness and stance in Twitter #covid19nz posts," *J. Pragmat.*, vol. 196, pp. 6–23, 2022.