Stem & Leaf Plots Extended for Text Visualizations

Richard Brath
Computer Science and Informatics
London South Bank University & Uncharted Software
Toronto, Canada
richard.brath@alumni.utoronto.ca

Ebad Banissi
Computer Science and Informatics
London South Bank University
London, U.K.
banisse@lsbu.ac.uk

***Abstract*—Stem and leaf plots are data dense visualizations that organize large amounts of micro-level numeric data to form larger macro-level visual distributions. These plots can be extended with font attributes and different token lengths for new applications such as n-grams analysis, character attributes, set analysis and text repetition.**

Keywords - stem & leaf, text visualization, exploratory data analysis.

I. INTRODUCTION

Stem and leaf plots were described by John Tukey [1] and popularized by Edward Tufte in *The Visual Display of Quantitative Information* [2]. The stem and leaf plot provides a macro-level visual distribution and a micro-level view of the individual data points. Identified benefits [3] of stem and leaf displays include 1) more information is retained than a bar chart; 2) reveals fine structure while showing the distribution and 3) allows easy hand-calculation of measures based on ordered values (e.g. median, quartiles). Cox also points out limitations including: 1) problems with large datasets, 2) whether extra digits are useful to the task and 3) comparison can be awkward.

While the basics of stem and leaf plots may be known to the visualization community (e.g. [4,5]), the contribution of this paper is to:

1) Collect existing extensions;
2) Identify extensions and applications to text visualization.
3) Extend stem and leaf plots of text to operate with tokens of individual characters, words or phrases.

II. BACKGROUND

Stacked alphanumeric values to indicate distributions of data predate their use by Tufte and Tukey. Modern uses of stem and leaf plots in the wild indicate many extensions including the use of colour (fonts, backgrounds) or added indicators such as shading, lines or markers.

A. Statistics

Historic examples of stacked alphanumerics forming distributions occur prior to Tukey. For example, in this 1937 chart (a subset shown in fig. 1, from [6]), individual U.S. states are represented by numbers and stacked according to percent of the population receiving debt relief during the great depression (note that 2-letter state codes did not exist until 25 years later).

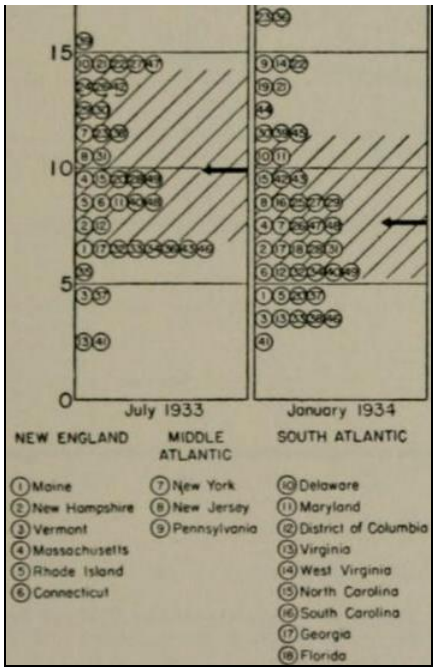

Figure 1. Portion of a 1937 chart showing percent of population per U.S. state receiving relief during the Great Depression.

The numerics denoting each state are surrounded by circles: the circles make both one digit and two digit glyphs a consistent size; facilitate disambiguation in a sequence of numbers; and make the number of items within a stack more visually apparent than a string of digits. The diagonal line texture denotes the inner quartiles (i.e. the centre half of the population) and the arrow indicates the median.

Stem and leaf plots in statistics have evolved with various added features. Fig. 2 (from [7]) shows two parallel stem and leaf plots indicating two subpopulations (e.g. male vs. female). On the left distribution, the inner set of numbers adjacent to vertical line indicate the stem on the left (i.e. the most significant digits) and the next significant digit to the right, with each successive digit indicating an additional observation. For example 15|459 indicates data observations with values of 154, 155 and 159 (e.g. the heights of three individuals observed). Visually scanning the distribution

Figure 2. Paired stem and leaf plots with additional metrics.

shows that the longest bar is beside the stem 18, indicating that observations between 180-189 occur most frequently.

To the left of the stem is a second set of numbers indicating the cumulative count of members from the extremes to the median denoted in round brackets [8], where the bracket number indicates the number of observations at the median. Note that the second parallel distribution is aligned to the same vertical axis as the first distribution. This facilitates comparison between the two populations. Note that the two stem and leaf plots are aligned to the same vertical axis and thus stem labels are indicated only in the left stem and leaf plot.

B. Finance

In the late 1800's alphanumeric charting techniques began as a matrix of prices over time initially as figure charts [9,10] and evolved into point and figure charts, using X's to indicate rising prices, O's to indicate falling prices. Coloured characters predate the use of computers, e.g. Livermore [11] used colour pencils to record price levels in his alphanumeric charts.

Market profile charts (fig. 3) closely resemble stem and leaf plots with stems representing discrete price levels and leaf characters indicate the time of day that a commodity trades at that level. A common encoding uses A-X, a-x to indicate half hour intervals starting at midnight with uppercase indicating trades in the morning, lowercase for afternoon. Characters are aligned vertically by price and stacked horizontally forming a histogram, enabling a macro-reading (the distribution) and a micro-reading (the individual characters).

Many organizations provide market profile software (e.g. CBOT, CQG, TradeStation, Reuters, ProRealTime, etc) with many variants used to differentiate individual characters,

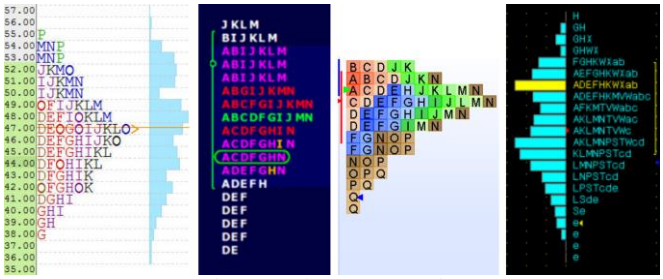

Figure 3. Market profile charts from various software providers. Stem represents discrete prices, while letters indicate times during which a trade occurred at that price level.

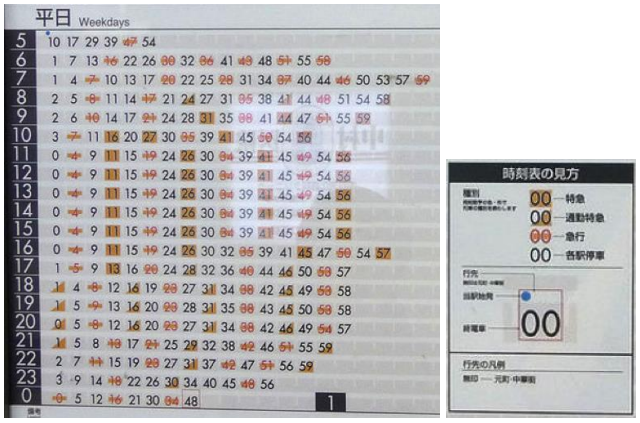

Figure 4. Stem and leaf plot of train times. Note text colour, background colour, background shape, added dots and outline square indicate attributes of trains

rows and ranges of rows. These include, per character: 1) character colour, 2) background colour, 3) upper/lowercase; per row: 4) background line, 5) background outline (e.g. rounded rectangle), 6) added mark (e.g. coloured triangle at start or end of row; or a glyph such as >); per vertical range, such as 7) shaded background, 8) line beside representing extent of range; and separately 9) an added distribution representing a second metric either back-to-back or side-by-side.

C. Timetables

One popular modern use of stem and leaf plots is timetables, such as commuter rail schedules. The stem indicates the hour of departure and each leaf indicates the number of minutes past the hour. These timetables may have additional information encoded for each specific train by providing additional graphical attributes per leaf. In fig. 4, information is added by the font colour; background shading and shape of shading (e.g. box, bar, triangle) underneath the leaf; a blue dot above the leaf; or an outline around the leaf. Note how legibility can be reduced with particular combinations, such as red text over an orange bar: the lack of contrast between the red and orange and the overlapping shapes interfere with the legibility of the text.

III. Extensions to Text Visualization

Stem and leaf plots can be extended to text visualization. There are several potential enhancements:

A) Font attributes: In addition to visual attributes such as colour, typographic attributes such as bold, italic, underline and so forth, can be used to add data. The use of font attributes has been discussed earlier in [12]. The contribution of this paper is to extend font attributes to stem and leaf plots.

B) Token - character, word or phrase: Numerically oriented stem and leaf plots typically use one or two characters. However, in the context of text visualization, the scope of the textual unit (i.e. token) can vary depending on the application: e.g. individual characters, words, or phrases.

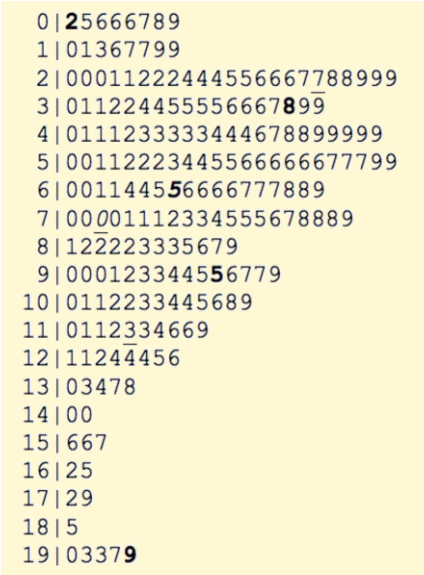

Figure 5. Stem and leaf plot with statistical values indicated via font formats.

A key contribution of this paper is the extension of stem and leaf plots to text visualization where the tokens may be at the level of characters, words or phrases.

A. *Font Attributes for Statistics*

In the earlier stem and leaf plots shown, additional data is encoded into the display using attributes such as foreground colour, background colour, background shape and dots as shown in figure 4. Instead, font attributes can be used. Figure 5 shows the plot of mountain heights (from [2]), with quartiles in bold, median in bold italic; standard deviation with underline, and mean with underline italic.

B. *Character Examples: Letter frequencies and state stats*

Instead of stems and leaves representing numerical values, a simple extension for text analytics is to use stems and leaves to represent alphabetic values. *Bigrams* (more generally *ngrams*) are sequences of adjacent letters used to provide the conditional probability of a token given the preceding token. Frequency of bigrams can be used for statistical language identification, prediction for auto-completion and cryptography.

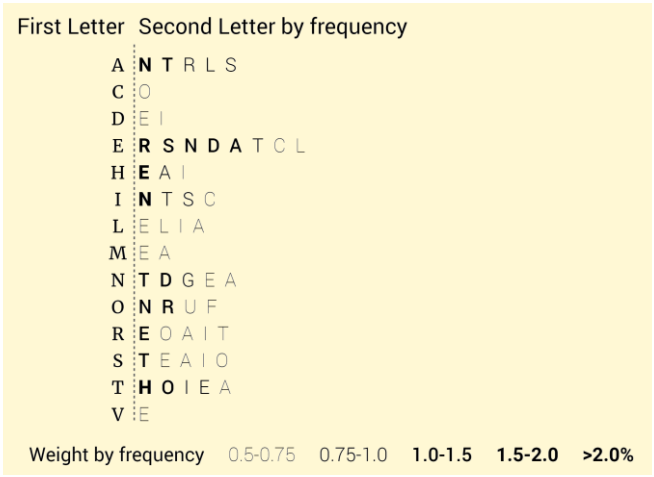

Figure 6. Bigrams in the English language. Font weight indicates frequency of occurence.

Figure 7. U.S. States: stem indicates poverty rate, font weight indicates life expectancy and colour indicates murder rate.

Figure 6 shows English language bigrams that occur more than 0.5% of the time based on bigrams calculated from the Leipzig Corpora Collection (corpora.uni-leipzig.de). The stem indicates the first letter of the bigram, the leaf indicates the second letter. Font weight indicates the bigram frequency, e.g. TH is among the most frequent bigrams in English.

In this case, the stem and leaf approach is used as a layout to organize the first and second token. Leaf stack length indicates the frequency of the initial token, e.g. bigrams starting with E are most common in bigrams occurring more than 0.5%. Additional data has been added by the font weight, in this case, indicating the frequency associated with each specific bigram.

Leaves do not need to be restricted to a single alphanumeric character. Figure 7 shows U.S. states with stem indicating poverty rate and additional data via hue and font weight. Multi-attribute correlations are visible, e.g. higher murder rates (red) and lower life expectancy (non-bold) are associated with higher poverty rates (top portion of plot).

C. *Word Examples: Character Traits and Families*

Leaves do not need to be restricted to tokens of the same length, e.g. words can be used instead. Figure 8 illustrates a character trait analysis by identifying adjectives that occur within +/- 3 words from a character in *Grimms' Fairy Tales*, with the stem indicating the character and the leaves indicating descriptors. Adjectives are ordered left to right based on frequency with font weight indicating the level of frequency (note that kings tend to be old while princesses are beautiful).

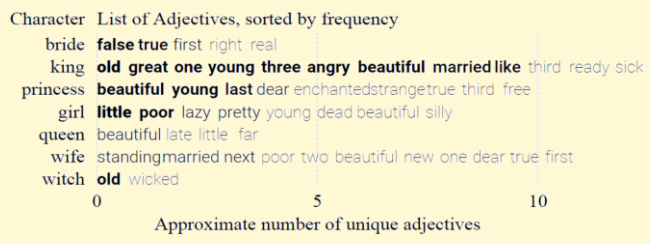

Figure 8. Adjectives associated with chaacters from *Grimms' Fairy Tales*

One challenge with variable length tokens is spacing: if the space provided per word is based on the longest word, then there will be a lot of wasted whitespace. Also, uneven whitespace across a string of adjectives is more difficult to read than word spacing based on typical text spacing (e.g. see [13]). Here, words are placed in sequence with an expected single space between words. A horizontal scale indicates the approximate number of words, based on average word lengths in the plot. Longer word lists should experience reversion to the mean: at the grid line for 10 average words, the number of words for king is approximately 9.3, for princess 9, for wife 10 - i.e. an error rate of only 10% in this example.

Another challenge is finding multiple visual attributes that can be combined together and remain legible in any combination (e.g. orange bars in fig. 4 interfere with text legibility). Font attributes can be combined together while retaining legibility [12]. Figure 9 uses font attributes on a subset of third class Titanic passengers. The stem indicates surname and the leaf indicates given name. In this particular example, 1) font weight represents survival: e.g. bold indicates death, 2) italics represent gender: e.g. italic indicates female, 3) capitalization represents age: e.g. all caps indicates children, and 4) font family represents class: a plain font for third class, a serif font for first class (fig. 10).

Since there are many binary attributes indicating membership for different sets, back-to-back stem-and-leaf plots can be used to more clearly show membership for a specific attribute. For example, fig. 10, shows a subset of first class families, with females on the left and males on the

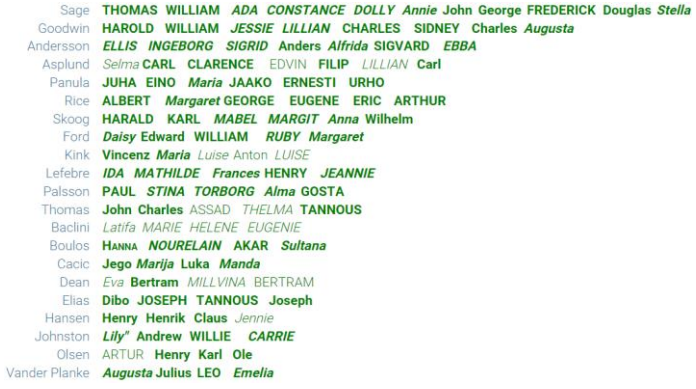

Figure 9. *Titanic* third class families. Sterm indicates surname, leaf for given name, bold indicates death, italics for women, allcaps for children.

Figure 10. *Titanic* first class families, women left, men right.

right, clearly indicating higher survivorship among women (i.e. there is more bold on the right side of the plot). Although there are survivors among the first class men, capitalization reveals that all the dead men are adults ("women and children first").

The greater proportion of heavyweight font for third class passengers (fig. 9) vs. first class passengers (fig. 10) is clearly visible: far fewer first class passengers die. Similarly only one first class child (all caps) death is visible in the subset shown while many of the third class deaths are children (allcaps bold).

D. Phrase Examples: Sectors and Companies

Figure 11 shows a stem and leaf plot for the performance of 500 stocks aggregated into 150 different industries. In this example, a different strategy is used for variable label length: instead of placing each word in sequence, a fixed width is provided for each label. Long labels are compressed using a narrow version of the particular font with narrow inter-character spacing (i.e. tracking), while short labels use a wide version of the same font with a wide inter-character spacing.

Another approach to scalability with long names and many items is to rotate the stem and leaf plot 90 degrees as shown fully in fig. 15 and closeup in fig. 12. This plot shows a horizontal distribution of earnings performance of 500 companies. In this orientation, the plot more closely resembles a bar chart. There is no layout error in counts as the height is consistent for each item. Phrase length is irrelevant – some phrases can be short (e.g. The Gap) and some can be long (e.g. Molson Coors Brewing Company). In

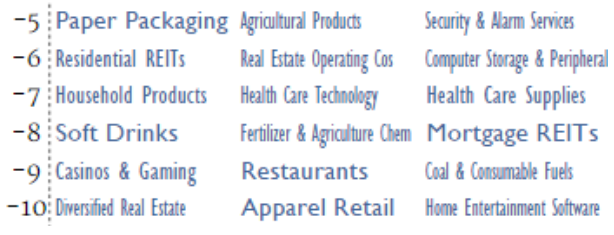

Figure 11. Closeup of stock performance for 150 industries: see fig. 14 for full view.

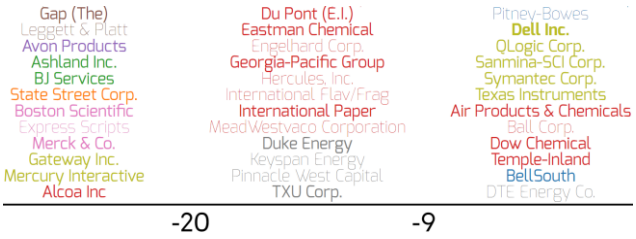

Figure 12. Distribution of companies by performance. Stack height indicates number of companies within a range of performance.

this example, font color indicates sector (e.g. Tech, Financial, Retail), font weight indicates stock trading volume.

The horizontal approach can be extended to longer phrases. Figure 13 and 16 are subsets of a visualization of common phrases repeated in the *Book of Psalms* from *The King James Version of the Bible* (www.gutenberg.org/ebooks/10). The source text is split on punctuation marks into phrases. Commonly repeated phrases are shown on the centerline. Phrases immediately prior the common phrase are above the centerline, phrases immediately following are below. Font weight indicates frequency.

For example, the phrase "O give thanks unto the Lord", is very commonly preceded by "Praise ye the Lord", although on one occasion is preceded by "I will exalt thee". Notice the common phrase "I will praise thee" is frequently followed by "O Lord", while on one occasion it is followed by "O Lord my God" – one may wonder if the latter is a transcription anomaly and then investigate the original document.

IV. Conclusions

The contribution of this paper includes many novel variants of stem and leaf plots, including 1) use of font attributes to indicate data; 2) text markers to indicate either categoric or numeric values for either stems or leaves; 3) text markers which range from single character, to words to phrases; 4) horizontal and vertical orientations.

Applications include text analytics such as n-gram analysis, character analysis and set analysis.

One issue identified is variable length leaves, and multiple strategies were identified: 1) tokens packed together with a horizontal axis based on average token size; 2) consistent token size and use of multiple font widths to accommodate for variance in token width; and 3) a vertical orientation so token width is not relevant.

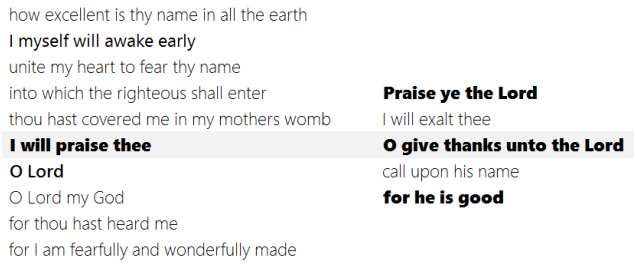

Figure 13. Common phrases in the *Book of Psalms* (centerline) with preceding phrases above, following phrases below.

None of the examples here discuss interaction. The interaction of stem and leaf plots with other well-known visualization techniques (e.g. linked interaction [14]) or emerging visualization techniques (e.g. object constancy [15] or sedimentation [16]) should be considered.

Scalability is not addressed, although fig. 15 shows hundreds of phrases in a stem & leaf plot while retaining legibility, and implies thousands of characters can be depicted legibly. Interactive techniques such as zooming and tooltips could allow for much larger stem and leaf plots showing macro patterns zoomed out and details on interaction.

Finally stem and leaf plots are less common than other techniques such as histograms and scatterplots and more prone to errors in interpretation (e.g. [17]). Some of the techniques shown here could be utilized to improve interpretation for novice users, for example, redundant encoding of the primary measure of the distribution using color or font-weight could be evaluated to determine potential improvement in performance.

References

[1] J. Tukey. "Some graphic and semigraphic displays." Statistical Papers in Honor of George W. Snedecor 5. 1966.

[2] E. Tufte The Visual Display of Quantitative Information. Graphics Press, 1983.

[3] N. Cox. "Speaking stata: Turning over a new leaf." The Stata Journal. 2013.

[4] J. Heer, M. Bostock, V. Ogievetsky. "A tour through the visualization zoo." Communications of the ACM 53, 6. 2010, 59–67.

[5] L. Wilkinson. The Grammar of Graphics, Second Edition. Springer, 2005.

[6] W. C. Brinton. Graphic Presentation. Brinton Associates, 1939.

[7] J. D. Emerson, D. C. Hoaglin. "Stem-and-leaf displays." Understanding robust and exploratory data analysis. 1983. 7–32.

[8] I. Salgado Ugarte. "Exploratory analysis of some measures of the asymmetric otoliths of stone flounder kareius bicoloratus (pisces: Pleuronectidae) in Tokyo bay." Anales del Instituto de Ciencias del mar y limnología. 18 (2). 1991, 261-278.

[9] Hoyle: The Game in Wall Street. Ogilvie Publishing, New York, NY, 1898.

[10] Wyckoff R.: Studies in Tape Reading. Noble Offset Printers, New York, NY, 1910.

[11] J. Livermore. How to Trade in Stocks. Quinn & Boden Company, Rahway, NJ, 1940.

[12] R. Brath, E. Banissi. "The design space of typeface." In Visualization and Computer Graphics, IEEE Transactions on. 2013.

[13] W. Tracy. Letters of Credit. Letters of Credit: A View of Type Design. Jaffrey, NH: David R. Godine Publisher, 2003.

[14] R. A. Becker and W. S. Cleveland. "Brushing scatterplots." Technometrics 29, 2. 1987. 127–142.

[15] J. Heer and G. Robertson. "Animated transitions in statistical data graphics." Visualization and Computer Graphics, IEEE Transactions on 13, 6. 2007. 1240–1247.

[16] S. Huron, R. Vuillemot, J.D. Fekete. "Visual Sedimentation." IEEE Transactions on Visualization and ComputerGraphics. Nov. 2013.

[17] R. S. Baker, A. T. Corbett, K. R. Koedinger. "Toward a model of learning data representations." Carnegie Mellon University Research Showcase, 2001.

Figure 14. Distribution of stock market returns across 150 different industries.

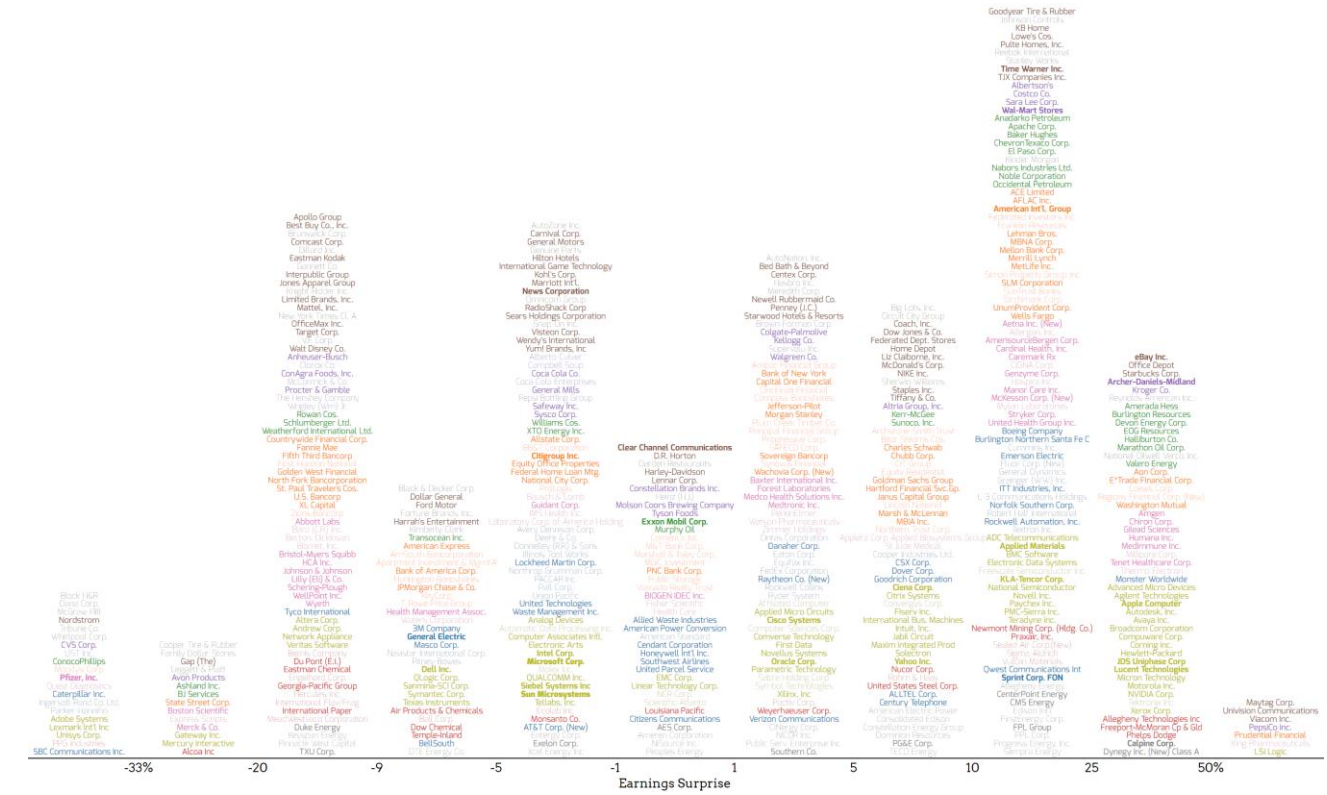

Figure 15. Distribution of earnings surprise for 500 stocks.

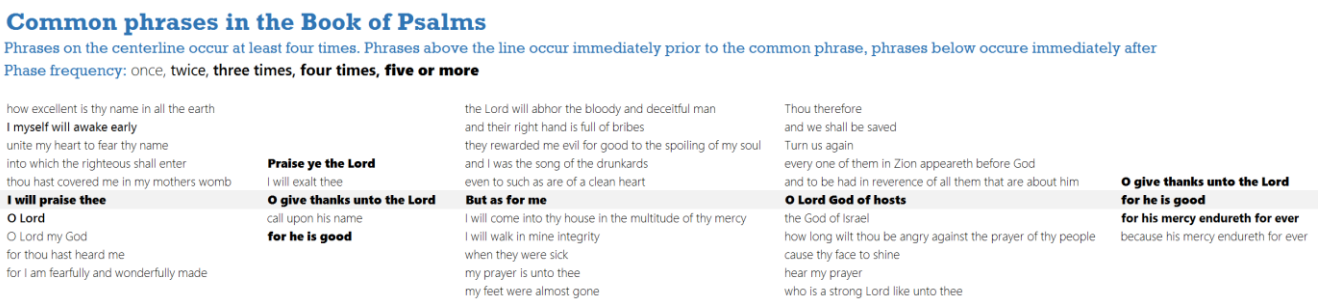

Figure 16. Subset of common phrases in the Book of Psalms (centerline), with preceding phrases (above) and following phrases (below). Phrase frequency indicated by font weight.