Hammock plots: visualizing categorical and numerical variables

Matthias Schonlau*
Department of Statistics and Actuarial Science,
University of Waterloo, Canada
schonlau@uwaterloo.ca

March 8, 2024

Abstract

I discuss the hammock plot for visualizing categorical or mixed categorical/numeric data. Hammock plots can be viewed as a generalization of parallel coordinate plots where the lines are replaced by boxes (or plotting elements) and the width of the boxes is proportional to the number of observations they represent. The paper also introduces a modification to the hammock plot to avoid what Hoffman et al. termed the reverse line width illusion. Further, I give an historical overview over hammock-type plots such as common angle, GPCP, parsets, and alluvial plots and discuss the type of boxes used to connect adjacent variables. Supplemental materials are available online.

Keywords: data visualization, common angle, GPCP, line width illusion, ParSets plots, alluvial plots
Number of words: 5276

*I gratefully acknowledge funding from grant 435-2021-0287 from the Canadian Social Sciences and Humanities Research Council (SSHRC) (PI: Schonlau). The author reports there are no competing interests to declare.

1 Introduction

Categorical variables are pervasive. Social scientists routinely collect data about gender, education, race/ethnicity, all of which are categorical. Even income is often measured as a categorical variable, each category corresponding to an income range. For survey research, multiple choice questions, dichotomous questions, Likert scale questions, choose-all-that-apply questions and matrix questions all lead to categorical variables.

Visualizing data is important during exploratory analysis. Univariate plot types are plentiful and include barchart, histogram, box and whisker plots, pie chart, and many others. Bivariate plots include scatter plots and stacked barcharts. Beyond two dimensions, there are fewer options: for continuous variables there are the scatter plot matrix and parallel coordinate plots (Inselberg, 1985; Wegman, 1990); for categorical variables historically the mosaic plot (Hartigan and Kleiner, 1981) was the main tool. However, even mosaic plots cannot accommodate a mixture of categorical/ continuous variables. A work-around is to bin the continuous variables to make them categorical.

In 2003, the hammock plot was introduced in a conference proceedings paper of the American Statistical Association (Schonlau, 2003). The hammock plot is an alternative plot for categorical data, and it also accommodates a mixture of categorical/ numeric data. Section 2 describes the hammock plot and proposes an improvement that addresses the so-called reverse line width illusion.

Much has happened in the 20 years since the hammock plot was first introduced. A considerable number of closely related plots succeeded the hammock plot including the parallel sets plot, the common angle plot, generalized parallel coordinate plot, and the alluvial plot. All these plots have a common ancestor, the parallel coordinate plot. Section 3 reviews these plots in their historical order of appearance. Section 4 compares the plots in terms of the shape and width of the boxes that connect adjacent variables. Section 5 shows how the hammock plot can be used to gain insight into the Shakespeare data. Section 6 concludes with a discussion.

2 The hammock plot

This section introduces the hammock plot; first a two-way hammock plot for two variables and then a hammock plot for multiple variables. The two sub-sections that follow discuss the so-called line width illusion and a remedy for the so-called reverse line width illusion. The section concludes with a brief survey of implementations of the hammock plot.

Throughout this section, we illustrate the hammock plot with the asthma data (Mangione-Smith et al., 2005). The asthma data were originally collected to determine whether an intervention improved processes and outcomes of asthma care. The asthma data set here (N=696) is used to illustrate the hammock plot; no medical conclusions are intended. Here, we consider the variables number of hospitalizations, the number of comorbidities, gender and group. The group variable specifies whether the person was a child, an adolescent or an adult.

2.1 The two-way hammock plot

Figure 1 gives a hammock plot for two categorical variables, group (child/adolescent/adult) and gender (female/male), in the asthma data. Like a parallel coordinate plot, the axes are aligned parallel to one another. Categories of adjacent variables are connected by boxes. (The boxes shown are parallelograms; I use the word *boxes* for simplicity). The width of boxes is proportional to the number of observations. "Width" refers to the minimal (orthogonal) distance between the two longer parallel lines. Choosing the minimal distance rather than the vertical distance avoids the so-called line width illusion.

The categories within a variable are spread out along a vertical axis. Optionally, hammock plots also display category labels. For example, in Figure 1 the variable "group" has three labels "adult", "adolescent", and "child". Optionally, a category is added for missing values and is displayed at the bottom. We see, for example, that most of the adults are female, whereas the gender ratio is more evenly matched for adolescents and children.

In summary, the hammock plot shows a graphical representation of the bivariate counts between two categories of two (adjacent) variables. Zero counts are represented through missing boxes. (For this plot, non-zero counts below 10 were increased to 10, because very

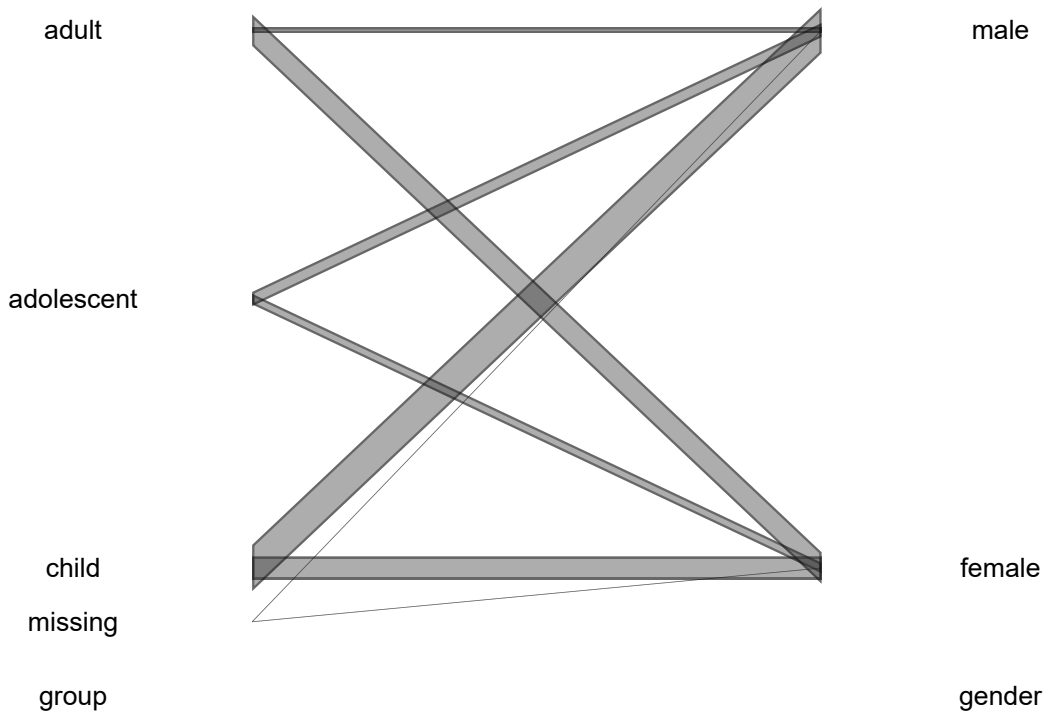


Figure 1: A hammock plot of two categorical variables, group and gender.

thin lines are barely visible, depending on screen resolution). The strength of Hammock plots lies in visualizing more than two variables. The next subsection turns to hammock plots with multiple variables.

2.2 The multi-way hammock plot

Figure 2 shows the hammock plot for the asthma data for four variables. The two additional variables are number of comorbidities and number of hospitalizations. Like a parallel coordinate plot, this plot extends the two-way version by adding axes parallel to existing axes. Figure 2 is highlighted by gender. As before, the hammock plot shows a graphical representation of the bivariate counts between two categories of two adjacent variables. (As before, for this plot, non-zero bivariate counts below 10 were increased to 10).

Figure 2 tells a story about the data: We knew already that most adults are females and that group membership is missing for a small fraction. We learned this because group

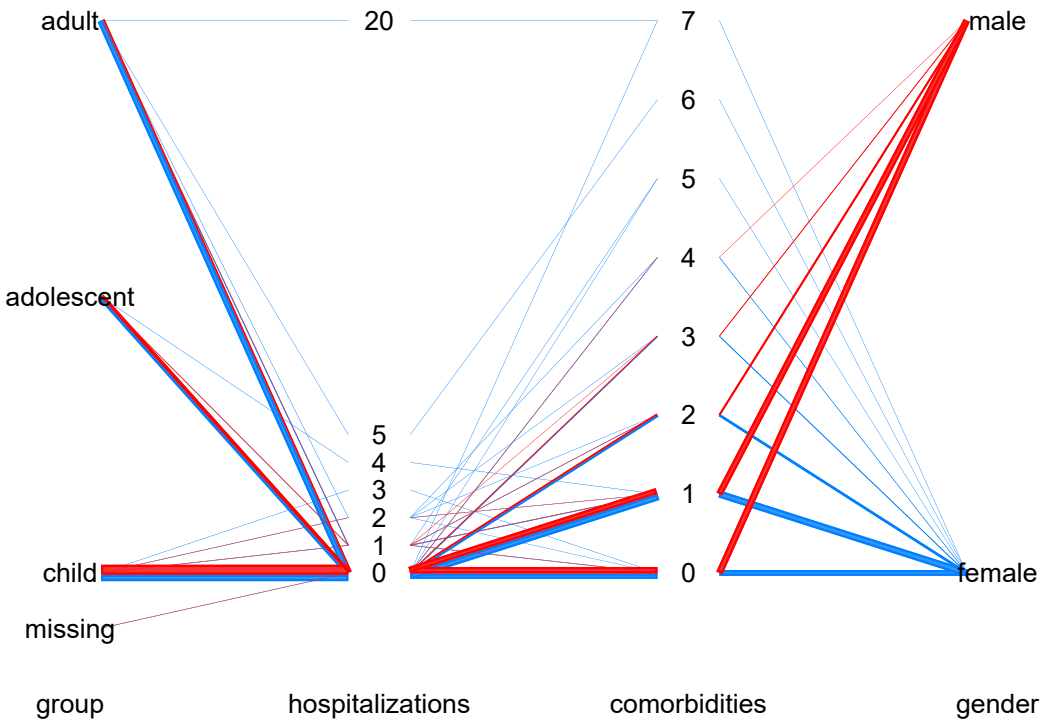


Figure 2: A hammock plot of the asthma data with multiple variables and space reserved for missing values. The plot is highlighted by gender (male in red).

and gender were adjacent in Figure 1. Now, group and gender are no longer adjacent, but highlighting gender allows us to gain the same information. Most males have 0 days of hospitalization. The number of hospitalizations is generally between 0 and 5, though there is one observation (or a small number of observations) with 20 days of hospitalization. This person, a female adult, also has the largest number of comorbidities.

Alternative one-plot visualizations for these four variables include mosaic plots and scatter plot matrices. A mosaic plot (see Figure 3) converts the number of comorbidities and the number of hospital visits into categorical variables and loses the quantitative information. Because numerical variables can have many categories with few observations per category, overplotting of the labels tends to occur. In Figure 3, some labels are removed to avoid overplotting. For this particular variable order, the mosaic plot brings out the larger percentage of females among adults-with-zero-hospital-visits (larger vertical blue box) as compared to the corresponding percentage for adolescents and children. Overall,

the Mosaic plot is not effective for these data.

A scatter plot matrix of the asthma data (not shown) loses information about frequencies due to overplotting and there is no information about missing values. Jittering, a counter measure to overplotting, is also not effective here.

Figure 2 has two numeric (hospitalizations, comorbidities), one ordered categorical (group), and one unordered categorical variable (gender). The user can choose the order of unordered categories, and can reverse the order of the categories for the ordered categorical variable. Moreover, the variable order can also be changed. Different orderings may facilitate different insights in the data. The order of categories is further discussed in the discussion.

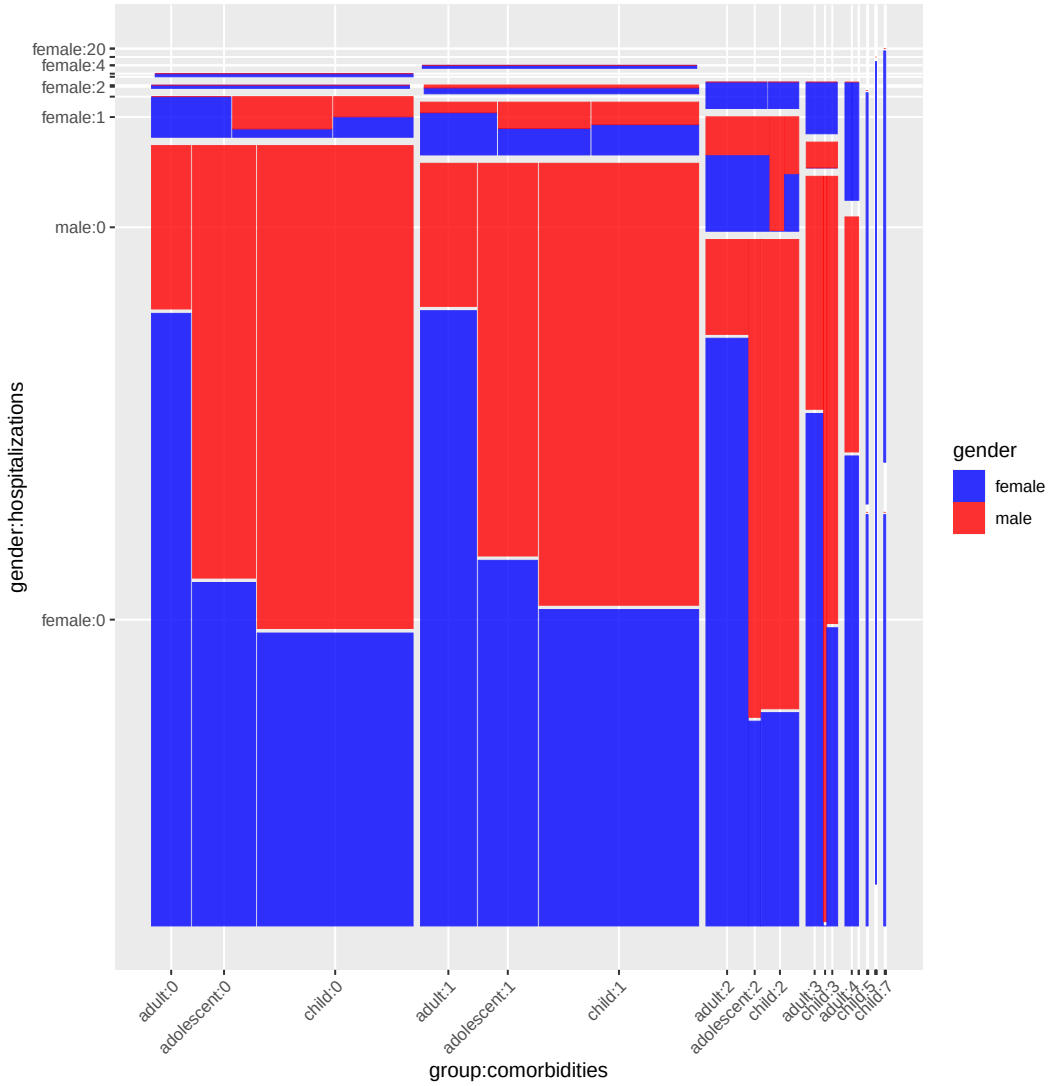


Figure 3: Mosaic plot of the asthma data (Implementation *ggmosaic* in R). Male gender is highlighted in red. Labels have been thinned out to avoid overplotting. The Mosaic plot is not as effective as the hammock plot for these data.

2.3 The line width illusion

The distance between two parallel lines is perceived at a right angle rather than as the vertical distance between the lines (Wallgren et al., 1996; Tufte, 2001). An example of such a line width illusion is shown in Figure 4. The vertical distance between all parallel lines in Figure 4 is the same. However, most readers focus on orthogonal rather than vertical

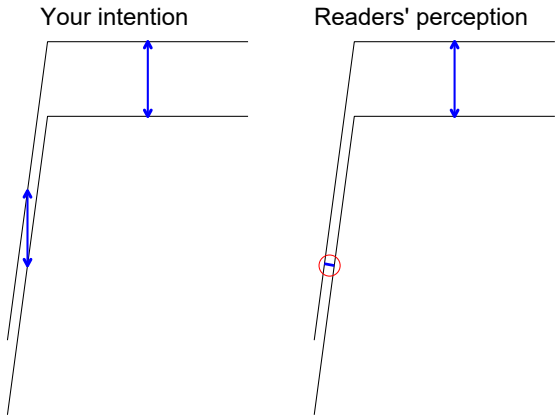


Figure 4: The line width illusion: The distance between two parallel lines is perceived at a right angle (inside the circle) rather than as the vertical distance between the lines. The vertical lines with arrows on both sides all have the same length.

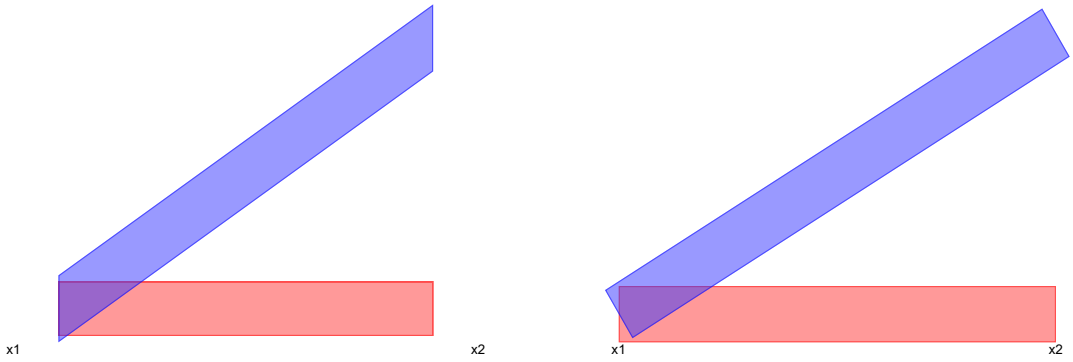
width.

The line width illusion is part of the family of Müller-Lyer illusions where two lines of same length appear to be of different lengths. (Hofmann and Vendettuoli, 2013). Closely related is also the sine illusion (VanderPlas and Hofmann, 2015) where a sine wave composed of equal-length vertical lines is perceived to have lines of unequal lengths (lines at the peak and trough of the curve appear to be longer).

2.4 Rectangles avoid the reverse line width illusion

Hofmann and Vendettuoli (2013) point out "centering of the lines creates a strong contextual cue that encourages an evaluation of line widths using the [vertical] measure, leading to a reverse line width illusion." They demonstrated in an empirical study that this can

sometimes confuse readers. An example is shown in Figure 5a. Both parallelograms have equal (orthogonal) widths. Because the blue parallelogram is angled, the vertical width of the angled blue box is larger. When you just focus on the end point on the left, the blue angled box appears bigger. This is the reverse line width illusion. To remedy the situation, I am now proposing to use rectangles instead of parallelograms (see Figure 5b). Because the vertical length is no longer shown at the end point, the illusion disappears.



(a) Parallelograms (b) Rectangles

Figure 5: Illustrating the reverse angle width illusion. Left: Even though they have the same right-angle width, the endpoints of the parallelograms appear to suggest that the red box is wider than the blue box. Right: The use of rectangles instead of parallelograms prevents this illusion.

Appendix A explains how to compute the coordinates of the rectangle. The Shakespeare example will use rectangles in Figure 14.

2.5 Implementations

Hammock plots are implemented in Stata in the package *hammock* (downloadable as usual from the Boston archive SSC by typing "ssc install hammock") and in R in the package *ggparallel* (Hofmann and Vendettuoli, 2016) with *method="hammock"*. A Python implementation called *hammock_plot* is now available via *pip install*.

Figure 6 shows a hammock plot as implemented in *ggparallel* in R. The plot looks a little different mainly because the numerical variables are shown as categorical variables.

(Hammock plots with numerical variables is not implemented in *ggparallel*.) This implementation has nice stacked bars for the marginal frequencies. The frequencies are separately displayed on the y-axis. If including missing values is desired, in the R implementation, missing values need to be converted into a separate category (labeled "missing" or similar) first.

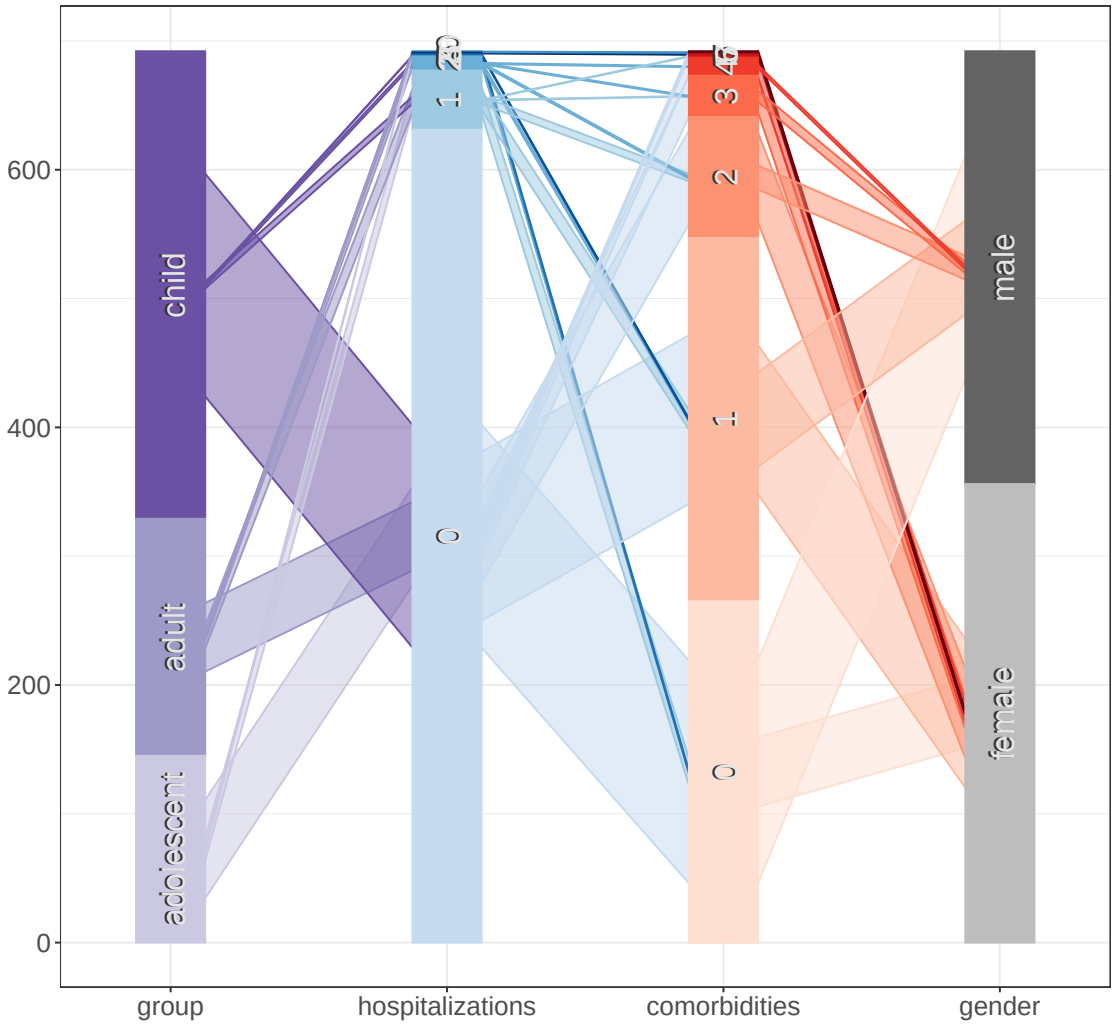


Figure 6: A hammock plot of the asthma data as implemented in *ggparallel* in R. This implementation is limited to categorical variables and turns any numeric variables (hospitalizations, comorbidities) into categorical variables.

3 Historical development

The focus of this section is on plots in the family of parallel coordinate plots as well as the Sankey diagram which has some similarities. Mosaic plots (Friendly, 1994, 2002; Hofmann, 2008) and other plots for categorical data are not further considered. The plots are presented in historical order starting with the earliest plot. Table 1 gives a timeline.

Table 1: Timeline of Graphs with parallel coordinates and the related Sankey diagram

Year	Name	Special Purpose	Categorical vars?	Numeric vars?
1885	Parallel coordinates	coordinate transformation		
1898	Sankey Diagram	chart of flows		
1985	Parallel Coordinate Plot			yes
2002	Clustergram	cluster assignments		
2003	Hammock Plot		yes	yes
2005	Parallel Sets		yes	binned
2010	Alluvial Plot	network vars over time	yes	
2013	Common Angle Plot		yes	
2013	Categorical par. coord. plot		yes	yes
2020	Generalized par. coord. plot		yes	yes

Parallel coordinates. In 1885, Philbert Maurice d'Ocagne wrote a book about parallel coordinates (d'Ocagne, 1885). d'Ocagne's book describes a method of coordinate transformation much like the transformation between polar coordinates and Cartesian coordinates. While d'Ocagne was the first to propose parallel coordinates, his mathematical treatment – full of equations – did not, in my opinion, foreshadow parallel coordinates as a technique to visualize data.

Sankey Diagram. Sankey diagrams (Sankey, 1898) are used to visualize flows of energy and materials. To do so, Captain Sankey used arrows with a width proportional to the flow. In modern implementations the arrow tips are sometimes omitted. Figure 7

gives a simplified example of the flows of materials as they pass through the EU economy (inspired by Eurostat, 2021). The flows can merge and separate again. Some flows end or start sooner than others in the graph, and it is possible to have circular flows and flows that go around corners (not shown in Figure 7 due to software limitations).

Unlike the plots described below, Sankey diagrams do not visualize variables. While Figure 7 might suggest there are 5 parallel axes corresponding to 5 variables, this is not the case. Rather than specifying variables, to construct a Sankey diagram each flow segment has to be specified separately. Schmidt (2008) gives a historical introduction to Sankey diagrams.

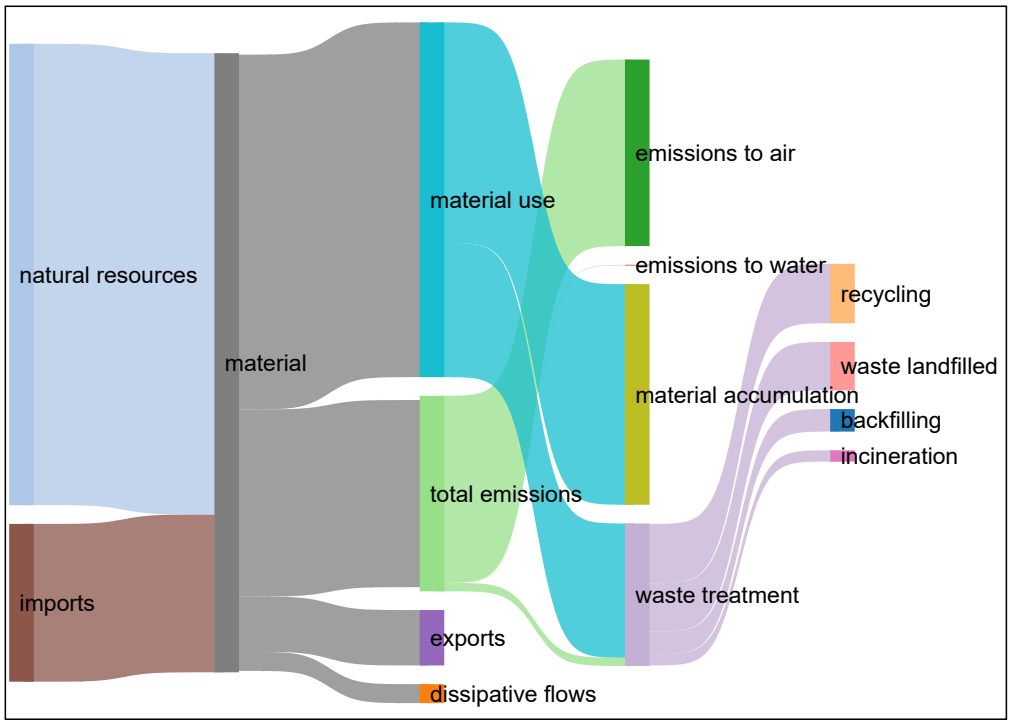


Figure 7: Sankey diagram of flows of materials through the EU economy. Sankey diagrams do not visualize variables. They may have different end points (shown) and circular flows (not shown). (Implemented in *sankey* in Stata)

Parallel coordinate plot. The well-known parallel coordinate plot (Inselberg, 1985; Wegman, 1990) is suitable for visualizing continuous variables. Missing values are not explicitly considered. The parallel coordinate plot can be generated from a hammock plot

by replacing the box elements with lines (or making the box width so small that all boxes look like lines). Figure 8 displays the hammock plot for the asthma data shown in Figure 2 as a parallel coordinate plot.

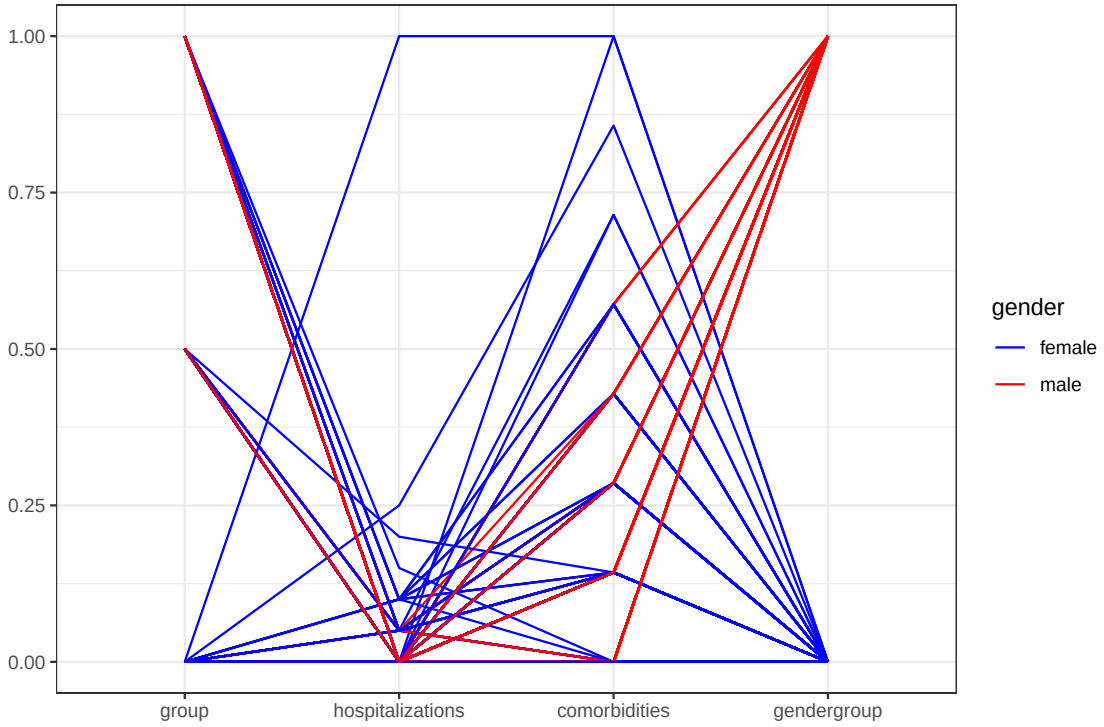


Figure 8: Parallel coordinate plot of the asthma data. (Implemented in *ggparcoord* in R). Categorical variables result in the overplotting of lines. This plot is less effective here.

Clustergram. The hammock plot was motivated by the clustergram (Schonlau, 2002, 2004), a plot for visualizing the result of clustering algorithms. Figure 9 shows an example visualizing the assignments resulting from a kmeans clustering algorithm. We see how observations are assigned to clusters as the number of clusters on the x-axis increases. As before, the width of the box is proportional to the number of observations in it. Of course, there are variations on how width can be defined. On the y-axis the clustergram displays either the mean values or the PCA weighted mean values as suggested in the R implementation by Tal Galili. We can see that there are hierarchical splits until we have three clusters. Then non-hierarchical splits start as we increase the number clusters to four and some observations rejoin other branches. The clustergram is implemented in R (Galili,

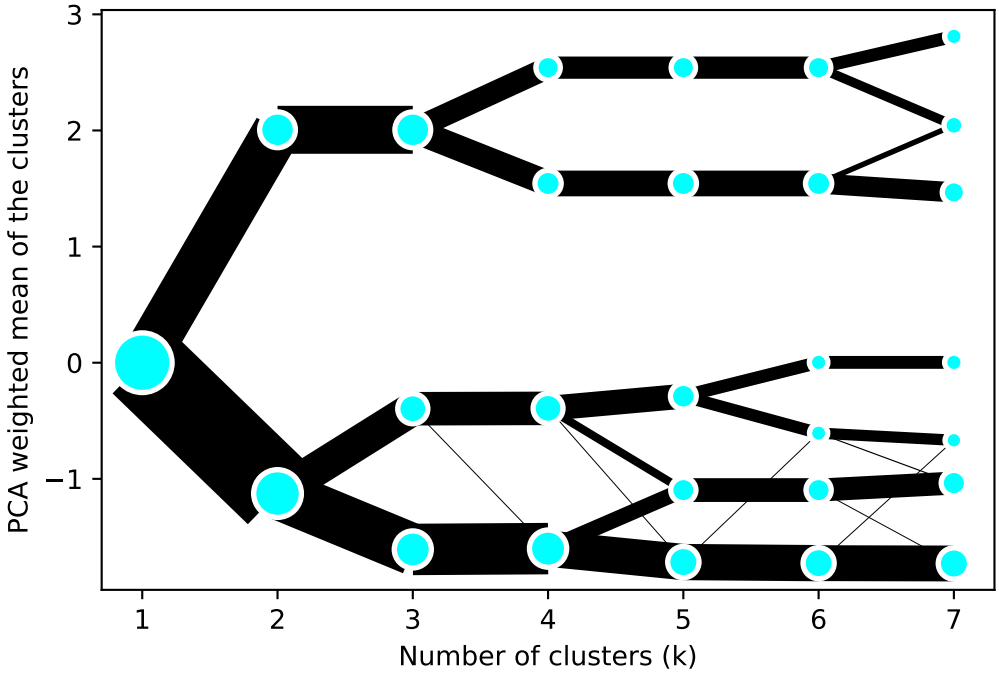


Figure 9: Example of a clustergram (implemented in *clustergram* in Python). Clustergrams visualize changing cluster assignments. They do not visualize variables, but are the direct precursor of the hammock plot.

2010), Python (Fleischmann, nd) and Stata (Schonlau, 2002).

Hammock plots. Among plots with parallel axes, hammock plots (Schonlau, 2003) were the first plots to accommodate categorical variables. Unlike some successor plots, hammock plots also accommodate numerical variables (see Table 1).

Parallel sets plot. The parallel sets plot (Kosara et al., 2006) is a variation on the earlier hammock plot. The main difference is as follows: for the parallel set plot the number of observations is proportional to the vertical width of the box instead of the minimal (orthogonal) distance. Since the vertical distance is meaningful, different boxes that lead from a single category to multiple categories on the next axis can easily be arranged on top of each other. This gives nice univariate views of how single levels on a given axis are subdivided into boxes. A disadvantage is that the parallel sets plot suffers from the line width illusion that the hammock plot avoids.

Parallel sets plots incorporate numeric variables by dividing the axis into bins, effectively turning the numeric variable into a categorical variable. The Kosara et al. (2006) implementation is very strong and includes interactive elements.

Alluvial plot. In 2010, Rosvall proposed alluvial plots (Rosvall and Bergstrom, 2010) to visualize network variables over time. Rather than using bars to connect axes, alluvial plots use rounded curves. Alluvial plots are now also used to visualize categorical variables. Figure 10 shows an alluvial plot for the asthma data. This implementation displays a category for missing values (at the top of the *group* variable), but only for variables that have missing values. Alluvial plots are not suitable for numeric variables.

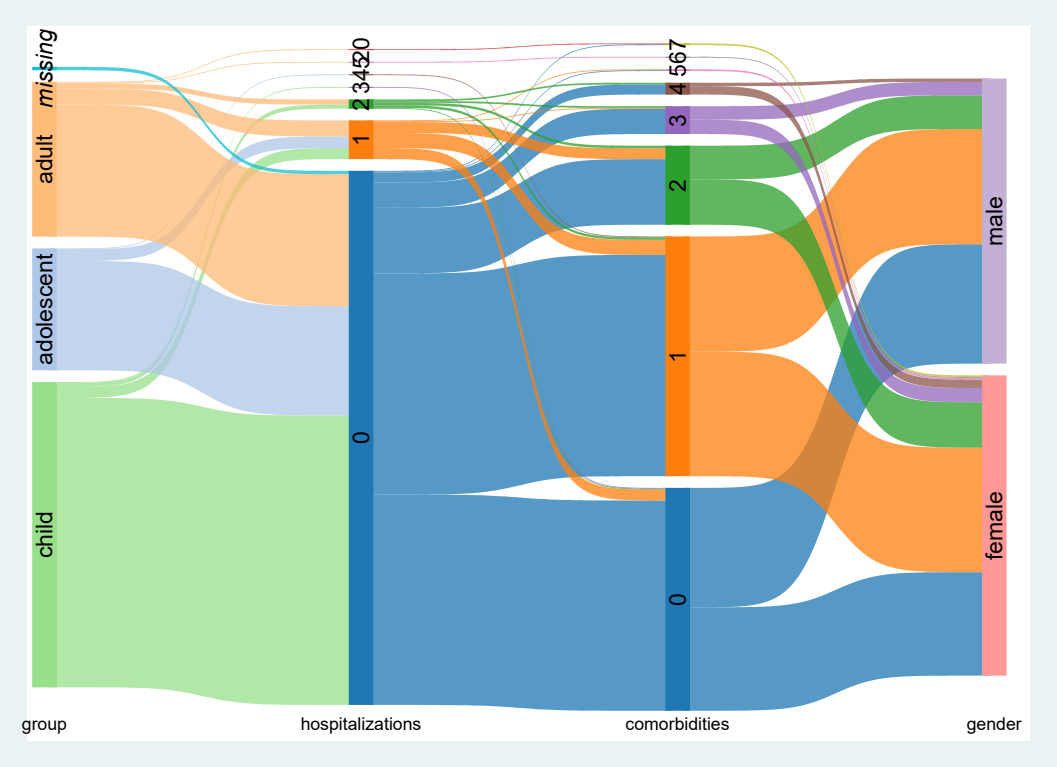


Figure 10: Alluvial plot of the asthma data (Implementation *alluvial* in Stata). The default color arrangement is shown. Alluvial plots use rounded curves to connect neighboring axes. They are not suitable for numeric variables.

Common angle plot. In order to avoid the line width illusion, the orthogonal distance in hammock plots is proportional to the number of observations. This increases the vertical width of individual blocks. The increase factor depends on the angle of the boxes. Because

each box protruding from a category has different angles, the increase factor is different for different categories. Therefore, it is not possible for the vertical and orthogonal distance to both be proportional to the number of observations. Hofmann and Vendettuoli (2013) point out that the line width illusion arises when comparing line segments drawn at different angles. They proposed a novel box element consisting of two rectangles (or box stumps) at either end and a angled parallelogram in the middle. An example is shown in Figure 11. Importantly, they use the *same* angle for all angled segments leading from one category

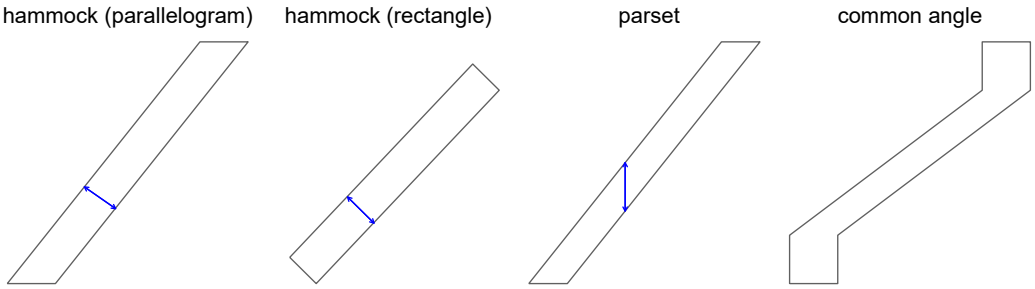


Figure 11: Box elements used by hammock plots, parset and common angle plots. The double-headed arrow inside of the boxes defines the width, the quantity that is proportional to the number of observations.

to the next variable. The farthest away segment determines the angle; the remaining angles can be kept constant by increasing the length of the rectangles at either end. The width of the angled segments is not explicitly specified; it arises as a function of the angle. This ingenious solution allows comparison between multiple segments protruding from one category. It does not allow comparison among line segments from multiple categories (because the angles will not be the same). A common angle plot of the asthma data is shown in Figure 12. By default, the package assigns a different color to every category of every variable displayed. Here, we chose a different color palette for each axis.

Categorical parallel coordinate plots (CPCP). Starting from parallel set plots, categorical parallel coordinate plots (Pilhoefer and Unwin, 2013) propose reordering boxes to reduce visual clutter. Specifically, for a given category in each axis, boxes can be reordered to minimize the crossing of boxes. The paper contains a plot with one continuous variable. Like in hammock plots, the numeric variable is not binned. (However, the current

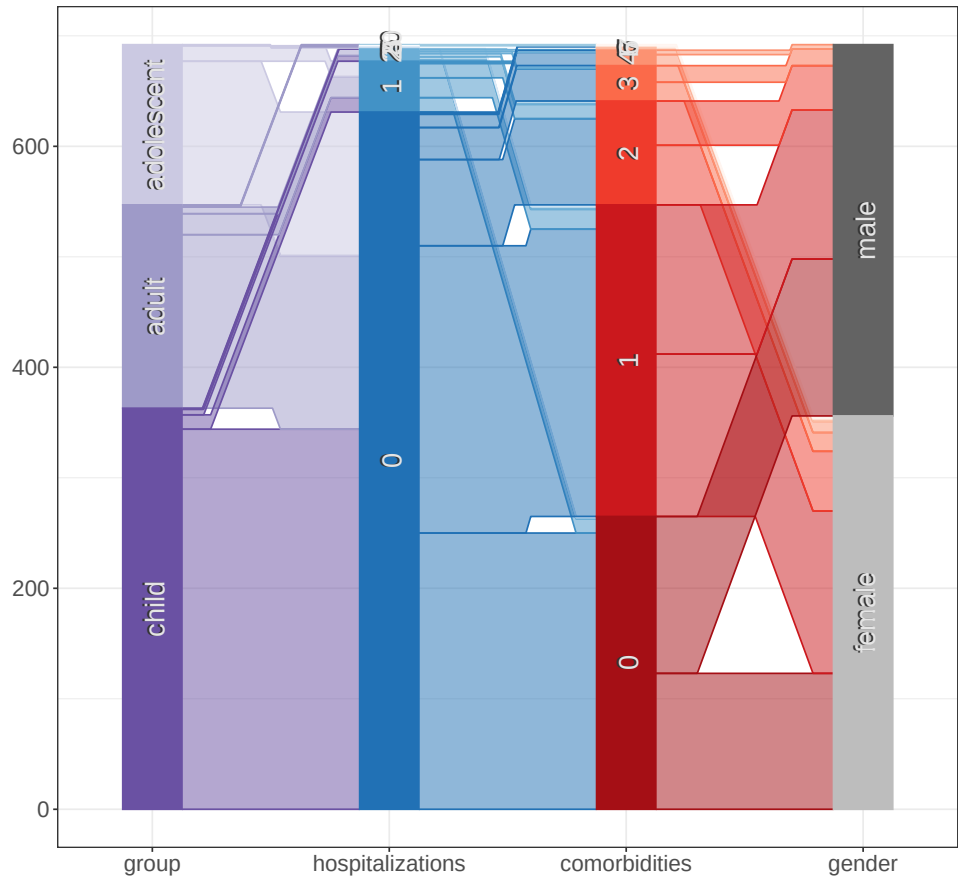


Figure 12: A common angle plot of the asthma data. (Implemented in ggparallel in R.) Such plots use the same angle for all angled segments leading away from one category.

implementation in R, *extracat::scpcp*, treats numerical variables as categorical.)

Generalized parallel coordinate plot (GPCP). Generalized parallel coordinate plots (VanderPlas et al., 2023; Ge and Hofmann, 2020) are designed to follow individual observations. A GPCP plot for the asthma data is shown in Figure 13. When observations start and end in the same two categories of two neighboring categorical variables, their lines are parallel and effectively form a box that looks like a parallelogram. When the number of observations is large, one cannot easily trace individual observations and the parallelogram may appear dense.

GPCP plots are similar to hammock plots in that they both allow for numeric variables and that the width of the boxes between two categorical variables is proportional to the

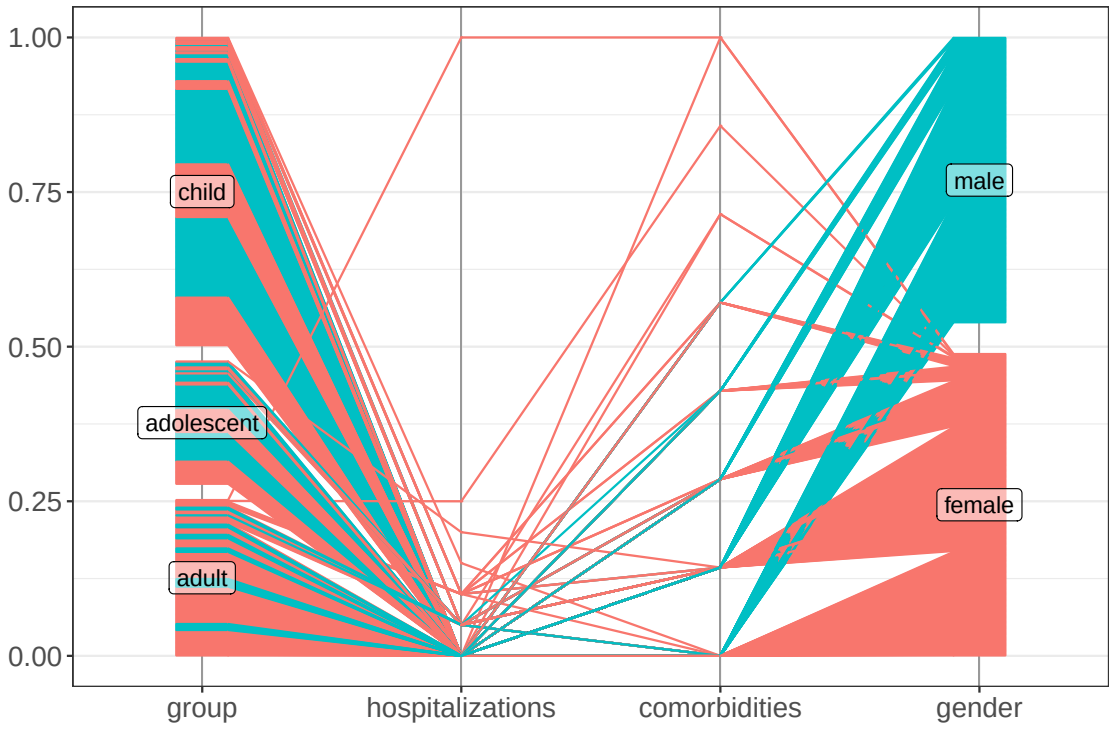


Figure 13: A generalized parallel coordinate plot of the asthma data. (Implemented in ggpcp in R.) For larger data sets, the individual lines for each observations appear as areas. When many observations have the same value for a categorical and an adjacent numerical variable, the corresponding area looks like a triangle.

number of observations (where the GPCP boxes are formed by parallel lines). They look different for numerical variables: When multiple observations terminate in the same value, triangles (between a categorical and a numerical variable) and lines (between two numerical variables) appear. The hammock plot uses constant-width boxes. This is further explained in Section 4. Notice the lines/boxes between the variables hospitalizations and comorbidities in the GPCP (Figure 13) and hammock plots (Figure 2). Most of the observations are in the boxes leading from *hospitalizations=0* to either *comorbidities=0* or *comorbidities=1*. This is far more obvious in the hammock plot than in the GPCP plot.

By contrast, hammock plots can trace some individual observations using highlighting by assigning some individual observations different colors.

4 Boxes that connect adjacent variables

Here, we consider different shapes and widths of boxes in hammock-type plots, that is, plots that use parallel axes and accommodate categorical variables. Table 2 lists characteristics of the boxes that connect adjacent variables. Different box shapes between two categorical

Table 2: Boxes that connect two adjacent variables in hammock-type plots

	box shape between...				
	2 categorical	1 cat + 1 num	2 numerical	what defines box width?	box width
		... variables			
hammock	parallelogram/ rectangle	parallelogram/ rectangle	parallelogram/ rectangle	right-angle	frugal
parset	parallelogram	[a]	[a]	vertical	wide
CPCP	parallelogram	[a]	[a]	vertical	wide
GPCP	like parallelogram	narrows to single point	line	like vertical	wide
common angle	angled	[a]	[a]	box stump[b]	wide[c]
alluvial	curved	[a]	[a]	box stump[b]	wide

[a] Plot does not accommodate numerical variables.

[b] For a horizontal box stump right-angle width and vertical width coincide.

[c] An "adjusted" version of the common angle plot allows reducing the box width.

variables were shown in Figure 11. The box shape between two categorical variables for the generalized parallel coordinate plot (GPCP) is listed as "like parallelogram" for the following reason: The GPCP traces individual observations. To the extent that multiple observations have the same start and end category, their lines are parallel. The shape of the parallel lines resembles a parallelogram, and the implied width of the parallelogram is the vertical width.

The hammock plot does not change the box shape when connecting to a numerical variable. By contrast, the GPCP does not maintain the width of the boxes but narrows to a single point. Narrowing boxes between categorical and numerical variables appear as triangles (see, for example, Figure 13 between *group* and *hospitalization*), and, between

two numerical variables, the box shape is an (overplotted) line (see, for example, Figure 13 between *hospitalizations* and *comorbidities*). When connecting two numerical variables, the GPCP is identical to parallel coordinate plots.

Parallel Sets, common angle, and alluvial plots do not accommodate numerical variables. The implementation of categorical parallel coordinate plots (R implementation *extracat::scpcp*) also does not currently support continuous variables. Of course, numerical variables can be binned.

The boxes in hammock plots are less wide compared to other plots listed in Table 2. This tends to leave more white space in hammock plots (see Figure 2). The smaller box width is a consequence of maintaining the width of the boxes when connecting to numerical variables (rather than narrowing to a single point). Consider the number of hospitalizations in Figure 2 with values 0, 1, 2, 3, 4, 5 and 20. When treating this variable as categorical with 7 categories, each equally spaced category is assigned (up to) one seventh of the space. (Box width is "wide" in Table 2). When treating this variable as numerical, the range from 0 to 20 leaves $1/21^{th}$ of the space for each unit length. Consequently, the widths of the boxes have to be more frugal.

The boxes of the common angle and alluvial plots are not straight, so no single width applies. They both have in common that they start out and end with a horizontal box stump. The common angle plot connects the stumps at an angle; the alluvial plot connects the stump with curvature. If the box stump is horizontal (that is, not angled), the vertical width and the right-angle width coincide.

5 Example: Shakespeare Data

Social class has traditionally played an important role in England. For example, during Shakespeare's time, Sir Thomas Smith divided the British population into four classes: Gentlemen, Citizens, yeoman artificers and labourers (using lower capitalization for the latter two) (Smith, 1583). Such social class structures are reflected in Shakespeare's plays.

The Shakespeare data compiled by Lee Wilkinson (Wilkinson, 1999b) contain one observation for each play Shakespeare has written. The data set was released as part of

Wilkinson's book (Wilkinson, 1999a).

The data used here contain six variables in addition to the name of the play: *type* (of play), *speaker1*, *speaker2*, *sex1*, *sex2* and the number of *characters* who appear in the play. The variable *type* refers to a play's classification as either a tragedy, history, or a comedy. Variables *sex1* and *sex2* correspond to the gender of the first two persons speaking in the play. The first two persons speaking are also classified by their social station: royalty, nobility, gentry, citizen, yeomanry, beggars. Table 3 gives an overview over how Wilkinson classified characters into social classes. The class "yeomanry" generally refers to individuals

Table 3: Social class membership of the first two speakers of Shakespeare's plays in descending order of class

Class	Membership
Royalty	king, queen
Nobility	prince, governor, duke, lord, count
Gentry	gentleman, senior military, bishop, Tribune (Roman official)
Citizen	merchant, carpenter, poet, painter, shipmaster
Yeomanry	servant, porter, messenger, common soldier, boatswain, guard, hostess
Beggars	beggar
NA	witch

of low status. The class "beggars" as a separate class may be unusual, but is useful here as it extends the social class hierarchy of the first two speakers at the bottom end. The witches in the opening scene of Macbeth are the only characters not classified into one of the social classes. Alongside fairies and ghosts, such supernatural characters defy the strict class hierarchy.

Wilkinson's social status classification is one of several similar classifications. Archer and Culpeper (2003)'s classification differs from Wilkinson's as follows. At the top end, Wilkinson's classification distinguishes between royalty and nobility whereas Archer combines them. At the bottom end, Wilkinson adds "beggars" as a separate class, whereas

Archer combines them with the low status class (yeomanry). Wilkinson's class "citizen" is split further into three classes (in descending order of status): professions (doctor, lawyers, etc.), other middling groups, ordinary commoners. It should be noted that none of the first or second speakers are doctors, lawyers, or similar professionals. Neither Wilkinson nor Archer makes provisions for supernatural characters like witches. Murphy (reported in Gillings, 2017) classified all of Shakespeare's characters by social class. In addition to Archer's classification, Murphy adds a class for Monarchy (exactly like Wilkinson's "Royalty" class), a class for supernatural characters, and a class for problematic characters (actors, poets, and characters whose social status changes).

Figure 14 shows a hammock plot of the Shakespeare data (using rectangles rather than parallelograms as plotting elements). The following statements refer to the data presented only, and no inferences to characters other than the first two speakers are implied. With this in mind, Figure 14 reveals quite a bit about this data set: Historical Shakespeare plays never open with women and most speakers in historical plays are royalty or nobility. Historical Shakespeare plays open more often with nobility than royalty, possibly because there were simply more noble men and women than royalty in society and also in the fictional worlds presenting a selection from society. Shakespeare's comedies tend to have a smaller number of characters than histories or tragedies. Comedies feature women and beggars in the opening of plays (but not often). There is one tragedy that opens with two women, but their social status is missing. We recall these are the witches in Macbeth. Except for the witches in Macbeth, no woman ever talks to another woman in the opening lines of Shakespeare's plays. Except for the witches in Macbeth, female characters in the opening lines only appear in comedies. The first two speakers usually do not speak much below their social station: royalty does not speak with anybody lower than nobility, nobility mostly talk to nobility and royalty, gentry mostly talk to gentry, and so forth.

If there is a dependent variable, we may want to highlight different categories of that variable. If the dependent variable is numeric, we might choose to highlight different quartiles, or something similar. Highlighting the dependent variable allows us to gauge how individual x-variables relate to it. Here, type of play may be considered a dependent

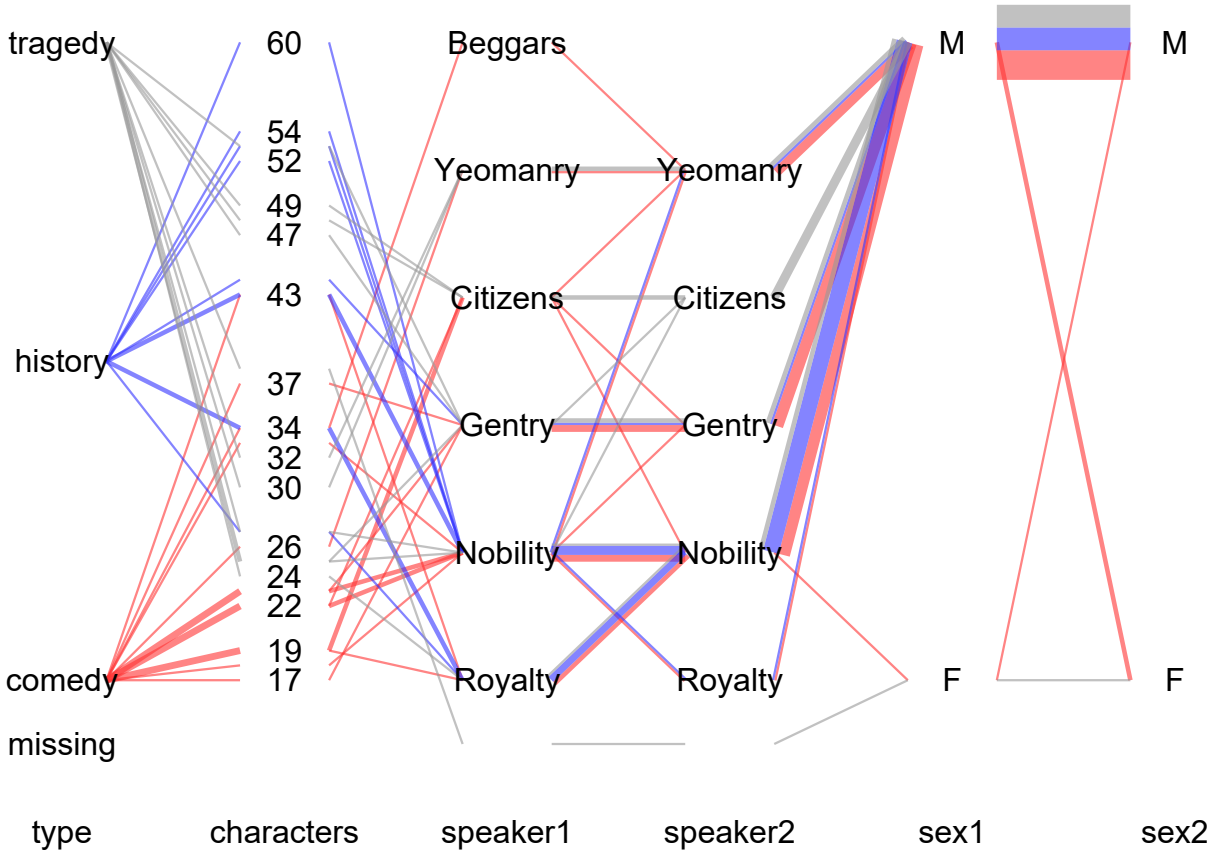


Figure 14: Hammock plot of the Shakespeare data. Each observation represents one of Shakespeare's plays. The indices 1 and 2 refer to the first two characters to speak in each play. "M" refers to "male" and "F" to "female". Play type is highlighted (tragedy=grey, history=blue, comedy=red).

variable, and the plot is already highlighted by type of play. As mentioned above, a smaller number of characters makes a comedy more likely. A play opening with women is not a historical play. A play opening with beggars, yeomanry, citizens, or gentry is likely not a historical play.

Figure 15 gives a GPCP plot of the Shakespeare data. The small number of observations plays to strength of GPCP: we can trace each individual play easily. In the hammock plot, tracing an individual play would require highlighting. In the GPCP, the grey shaded bars nicely give univariate information about each variable. Such bars are not currently

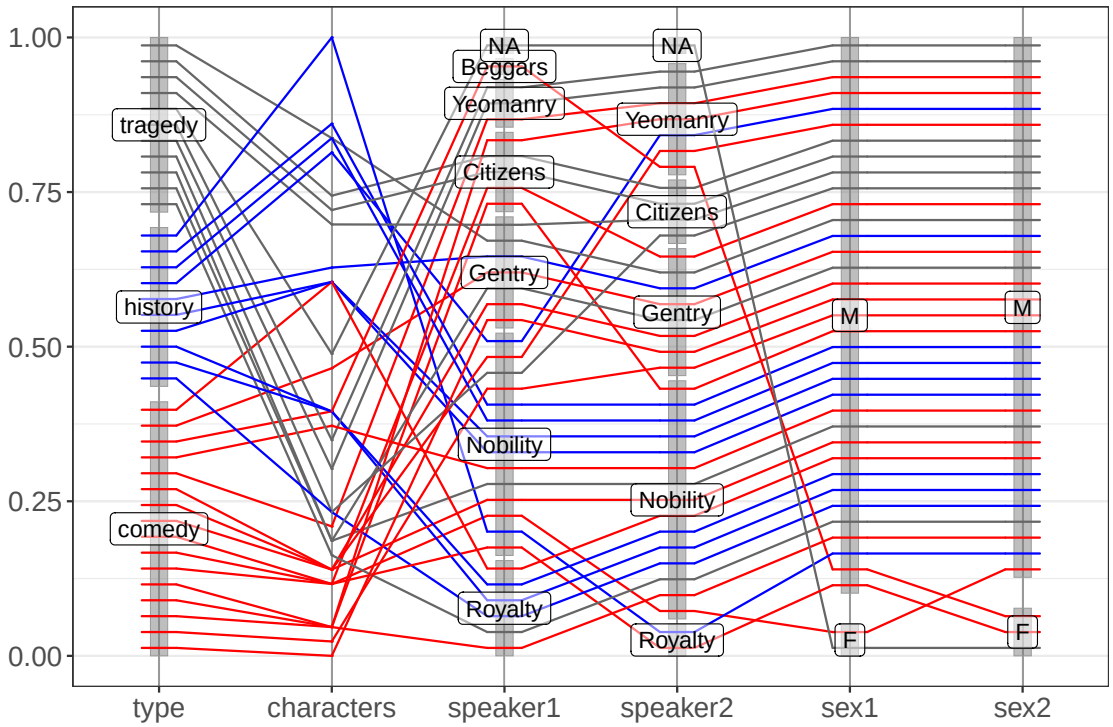


Figure 15: GPCP plot of the Shakespeare data as implemented in *ggpcp* in R. "M" refers to "male" and "F" to "female". Play type is highlighted (tragedy=grey, history=blue, comedy=red). In the black and white version tragedy is the darkest, comedy the lightest color.

implemented in the hammock plot. Similar to Figure 13, several "triangles" are starting to form between the first two variables, specifically for comedic plays that have the same number of characters. Because of the small number of observations, the triangles are not filled solid and they do not look odd here.

In the hammock plot the social class of *speaker1* lines up with that of *speaker2*. This is not the case in GPCP: observations are equally spaced vertically, and so the categories generally do not line up.

Arguably, some aspects of the bivariate relationships maybe a little easier to see in the hammock plots than in the GPCPs. For example, 1st speaker citizens talk to 2nd speaker citizens only in tragedies (because they are connected by a single grey box). Arguably, the color composition of the boxes is also easier to read in the hammock plot. In the GPCP plot

there is little space between two observations that belong to different categories making this harder to see.

Tracing individual plays is possible in the hammock plot also. For example, Figure 16 highlights the play with the fewest number of characters, "Taming of the Shrew". We can

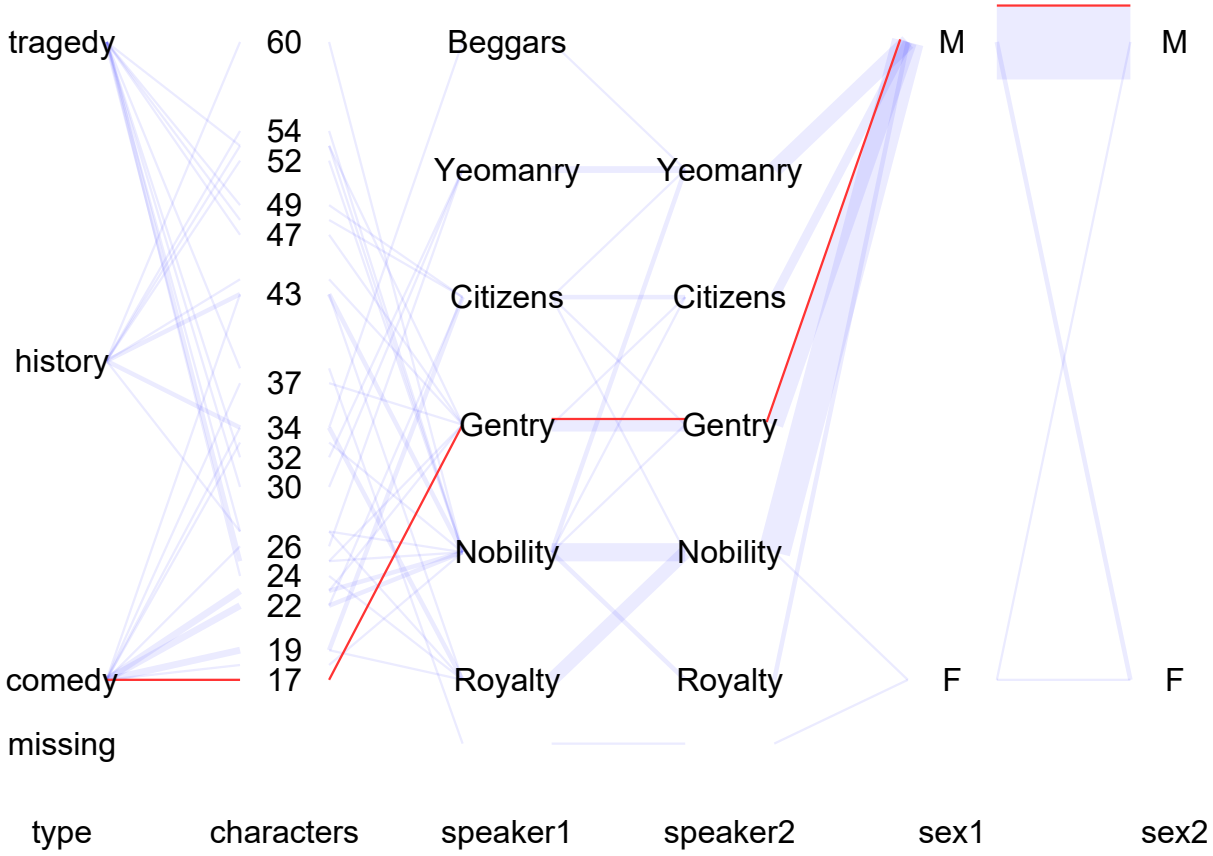


Figure 16: Hammock plot of the Shakespeare data. Each observation represents one of Shakespeare's plays. The indices 1 and 2 refer to the first two characters to speak in each play. "M" refers to "male" and "F" to "female". The play with the fewest number of characters ("Taming of the Shrew") is highlighted.

clearly trace this play through the six variables shown.

To understand what plots look like for larger data sets, we now replicate each observation in the data 50 times and create the hammock and GPCP plots again. The hammock plot in Figure 14 is unchanged. The boxes represent a greater number of observations but the plot

is unchanged. The revised GPCP plots is shown in Figure 17. The individual lines have

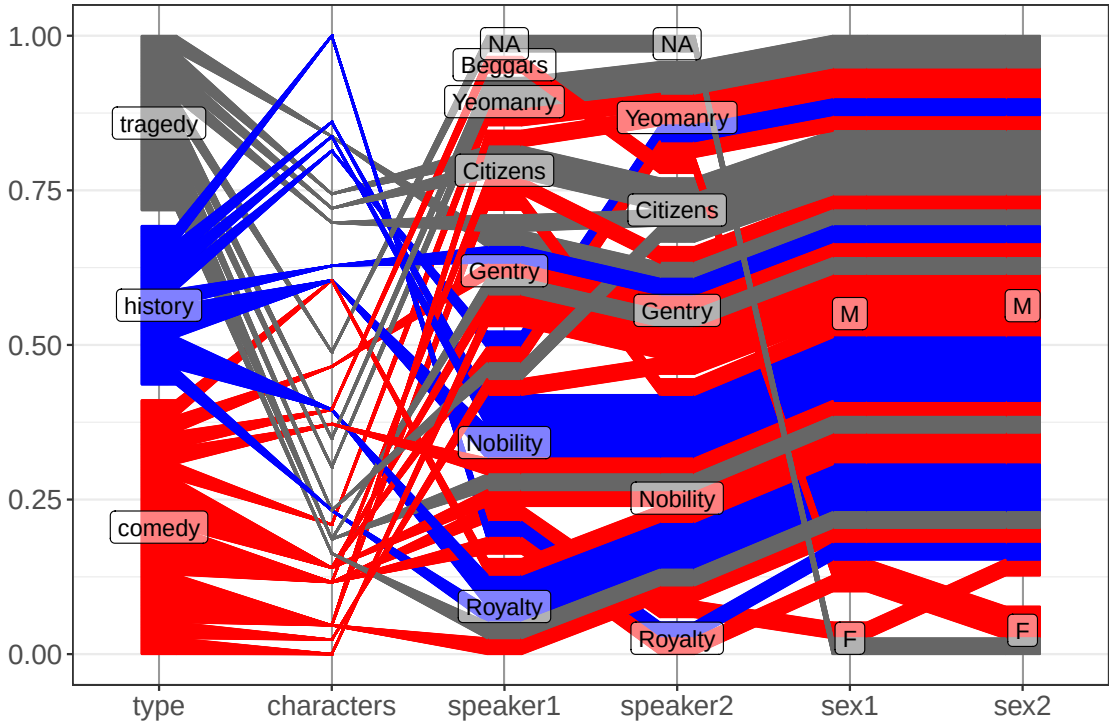


Figure 17: GPCP plot of the Shakespeare data as implemented in *ggpcp* in R. Compared to Figure 15, each observation has been replicated 50 times. In the black and white version tragedy is the darkest, comedy the lightest color.

melted together to form solid areas. Triangles leading to the numerical variable are now clearly visible. Individual observations are no longer visible. The labels and basic layout are unchanged. Arguably, The GPCP plot in Figure 17 looks busier than the hammock plot in Figure 14, but this is perhaps a matter of individual preference.

The Shakespeare data set is available as supplementary material. Some changes from the original data set and the addition of the *character* variable are explained in the section on supplementary material.

6 Discussion

The hammock plot is best suited for visualizing a handful of variables at a time. For a much larger number of variables, the graph may get difficult to read. Hammock plots accommodate a mixture of categorical/numeric variables with ease, and therein lies a key strength. For categorical variables, the hammock plot scales well to a large number of observations: the boxes simply represent more observations. For continuous and count variables, large numbers of observations will lead to more clutter.

Hofmann and Vendettuoli (2013) have pointed out that hammock plots suffer from the reverse line width illusion. Here, I have proposed a modification, using rectangles rather than boxes, that avoids the reverse line width illusion.

Missing values form an extra category for visualization. For categorical variables, this increases the number of categories to visualize by one; for continuous variables, the hammock plot adds a category below the smallest value. Because all variables may have missing values, it is natural to visualize missing values in the same position (e.g., at the bottom) on each axis. While the literature has occasionally addressed visualizing missing values (Unwin et al., 1996; Swayne and Buja, 1998; Cheng et al., 2015; VanderPlas et al., 2023), in my experience many analysts often ignore missing values when creating routine plots. One reason may be that implementations usually do not make this easy by not offering a "missing" option that includes missing values as part of the plot.

If the categories of a categorical variable are unordered, then the user is free to choose whatever order they like. In their general approach to "effect ordering for data displays", Friendly and Kwan (2003) propose orderings based on eigenvalue or singular-value decomposition. There has been some work on guiding the user in choosing category order in hammock-type plots (VanderPlas et al., 2023), but there is room for more research. Options for choosing category order include: 1) Arrange categories to minimize crossings of boxes. 2) Arrange categories to minimize the weighted number of box crossings where the weights are proportional to the width of the box, or equivalently, the number of observations the box represents. 3) Arrange categories to maximize similarities of adjacent categories where similarities need to be defined in some way.

The ordering of axes affects the visual presentation. In some cases, the axes may have a natural order, such as when each axis corresponds to a different point in time. More often, the axes do not have a natural order. In practice, trial and error reveals the most effective visualization. Different orderings may emphasize different patterns in the data and some orderings may reduce clutter more than others. Likely, a single plot does not serve all purposes.

The implementation can make a difference to the appearance of the graph. The R implementation of the hammock plot has very nice marginal stacked bar charts but currently treats numerical variables as categorical variables; information about distance between the categories is lost. Of course, it is possible to update the R implementation to accommodate numerical variables also. I am very grateful for the R implementation by Hofmann and Vendettuoli (2016).

The choice among hammock-type plots may boil down to whether the visualization involves only categorical variables and what analysts are focused on. Parsets and alluvial plots focus on univariate descriptors and link the categories with boxes. For parsets, the boxes suffer from the line width illusion. Alluvial plots just connect the univariate descriptors, and the curved connectors may look pretty. Hammock plots focus on the boxes, and thereby the bivariate relationships the boxes represent. Common angle plots are able to focus on both univariate and bivariate relationships at the cost of using a more complicated box element.

If the visualization involves both categorical and numerical variables, the choice narrows to hammock plots and GPCP plots. For small data sets, GPCP plots beautifully show all individual observations whereas hammock plots require highlighting to feature individual observations. For larger data, arguably similarities of hammock and GPCP plots outweigh dissimilarities. To some, hammock plots may appear less cluttered, but others may appreciate the larger bar widths of GPCP plots. When connecting two numerical variables GPCP plots show lines regardless of many observations the line represent. In this particular case the hammock plot may be more helpful. When connecting a categorical to a numerical variable, the GPCP plots show visual triangles whereas hammock plots do not. Triangles

no longer have a constant "width", but they have the advantage of taking up less space.

SUPPLEMENTARY MATERIAL

Appendix A Computing coordinates for the rectangle

Overview Spreadsheet that describes which code file corresponds to which plot.

Code files Stata/R/Python code reproducing the plots shown in this manuscript.

Asthma data set Data set used for illustration. (both csv and Stata12 versions)

Shakespeare data set Data set used for illustration (Wilkinson, 1999a). (both csv and Stata12 versions). The number-of-characters variable (Open Source Shakespeare, 2019) has been added to Wilkinson's data. The following changes have been made to Wilkinson's data:

- *Hamlet* and *King Lear* are missing from the original data and have been added.
- *Macbeth*: The type of play was corrected to be a tragedy. The first two characters are witches. Supernatural characters do not fit into the ordered social class structure. I set their social class to missing. (The original classification mistakenly had two male characters.)
- *Antony and Cleopatra*: The first speaker, Philo, a soldier, talks to Demetrius, another soldier. Before Demetrius has time to respond, Cleopatra and Antony appear on the other side of the stage and are the second and third speakers. Philo never speaks to Antony or Cleopatra. I treat the two soldiers as first speakers for the purpose of this classification. Following Wilkinson's original classification, I assume they are senior soldiers and classify them as gentry.

The following classifications are unchanged but merit additional explanation:

- *Cymbeline*. Wilkinson classified *Cymbeline* as a comedy. Depending on the critic, it is sometimes classified as a tragedy.

- *Edward III* and *Noble Kingsmen*. Wilkinson did not include *Edward III* and *Noble Kingsmen* because they are not in the Yale Shakespeare collection. They are not in Shakespeare's first folio and authorship is disputed. *Pericles* is also not in the first folio but authorship is more certain.

References

Archer, D. and J. Culpeper (2003). Sociopragmatic annotation: New directions and possibilities in historical corpus linguistics. In P. R. A. Wilson and A. M. McEnery (Eds.), *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, pp. 37–58. Peter Lang.

Cheng, X., D. Cook, and H. Hofmann (2015). Visually exploring missing values in multivariable data using a graphical user interface. *Journal of Statistical Software 68*(6), 1–23.

d'Ocagne, M. (1885). *Coordonnées parallèles & axiales: méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèles*. Gauthier-Villars.

Eurostat (2021). Material flow diagram. `https://ec.europa.eu/eurostat/web/circular-economy/material-flow-diagram`.

Fleischmann, M. (n.d.). Clustergram - visualization and diagnostics for cluster analysis in python. Last accessed on June 10, 2022, `https://github.com/martinfleis/clustergram`.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association 89*(425), 190–200.

Friendly, M. (2002). A brief history of the mosaic display. *Journal of Computational and Graphical Statistics 11*(1), 89–107.

Friendly, M. and E. Kwan (2003). Effect ordering for data displays. *Computational Statistics & Data Analysis 43*(4), 509–539.

Galili, T. (2010). Clustergram: visualization and diagnostics for cluster analysis (R code). Blog. https://www.r-statistics.com/2010/06/clustergram-visualization-and-diagnostics-for-cluster-analysis-r-code/.

Ge, Y. and H. Hofmann (2020). A grammar of graphics framework for generalized parallel coordinate plots. arXiv preprint. arXiv:2009.12933.

Gillings, M. (2017). Shakespeare and social status. https://wp.lancs.ac.uk/shakespearelang/2017/06/05/shakespeare-and-social-status/.

Hartigan, J. A. and B. Kleiner (1981). Mosaics for contingency tables. In *Computer science and statistics: Proceedings of the 13th symposium on the interface*, pp. 268–273. Springer.

Hofmann, H. (2008). Mosaic plots and their variants. In *Handbook of Data Visualization*, pp. 617–642. Springer.

Hofmann, H. and M. Vendettuoli (2013). Common angle plots as perception-true visualizations of categorical associations. *IEEE Transactions on Visualization and Computer Graphics 19*(12), 2297–2305.

Hofmann, H. and M. Vendettuoli (2016). ggparallel: Variations of parallel coordinate plots for categorical data. R package version 0.2. 0. URL: https://cran. r-project. org/package= ggparallel.

Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer 1*(2), 69–91.

Kosara, R., F. Bendix, and H. Hauser (2006). Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics 12*(4), 558–568.

Mangione-Smith, R., M. Schonlau, K. S. Chan, J. Keesey, M. Rosen, T. A. Louis, and E. Keeler (2005). Measuring the effectiveness of a collaborative for quality improvement in pediatric asthma care: does implementing the chronic care model improve processes and outcomes of care? *Ambulatory Pediatrics 5*(2), 75–82.

Open Source Shakespeare (2019). Total number of characters in each William Shakespeare play. https://www.opensourceshakespeare.org. Accessed from https://www.statista.com/statistics/1061409/character-count-shakespeare-plays/.

Pilhoefer, A. and A. Unwin (2013). New approaches in visualization of categorical data: R package extracat. *Journal of Statistical Software 53*(7), 1–25.

Rosvall, M. and C. T. Bergstrom (2010). Mapping change in large networks. *PloS one 5*(1), e8694.

Sankey, H. (1898). Introductory note on the thermal efficiency of steam-engines. report of the committee appointed on the 31st march, 1896, to consider and report to the council upon the subject of the definition of a standard or standards of thermal efficiency for steam-engines: With an introductory note. In *Minutes of proceedings of the institution of civil engineers*, Volume 134, pp. 278–283.

Schmidt, M. (2008). The sankey diagram in energy and material flow management: part I: history. *Journal of Industrial Ecology 12*(1), 82–94. https://doi.org/10.1111/j.1530-9290.2008.00004.x.

Schonlau, M. (2002). The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses. *The Stata Journal 2*(4), 391–402.

Schonlau, M. (2003). Visualizing categorical data arising in the health sciences using hammock plots. In *Proceedings of the Section on Statistical Graphics*. American Statistical Association. CD-ROM.

Schonlau, M. (2004). Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. *Computational Statistics 19*(1), 95–111.

Smith, S. T. (1583). *De Repvblica Anglorvm: The Maner of Gouernement or Policie of the Realme of England*. London: Printed by Henrie Midleton for Gregorie Seton.

Swayne, D. F. and A. Buja (1998). Missing data in interactive high-dimensional data visualization. *Computational Statistics 13*(1), 15–26.

Tufte, E. (2001). *The Visual Display of Quantitative Information*. Cheshire.

Unwin, A., G. Hawkins, H. Hofmann, and B. Siegl (1996). Interactive graphics for data sets with missing values—manet. *Journal of Computational and Graphical Statistics 5*(2), 113–122.

VanderPlas, S., Y. Ge, A. Unwin, and H. Hofmann (2023). Penguins go parallel: a grammar of graphics framework for generalized parallel coordinate plots. *Journal of Computational and Graphical Statistics*, 1–16. DOI:10.1080/10618600.2023.2195462, published online first.

VanderPlas, S. and H. Hofmann (2015). Signs of the sine illusion—why we need to care. *Journal of Computational and Graphical Statistics 24*(4), 1170–1190.

Wallgren, A., B. Wallgren, R. Persson, U. Jorner, and J.-A. Haaland (1996). *Graphing Statistics & Data: Creating Better Charts*. Sage.

Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association 85*(411), 664–675.

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

Wilkinson, L. (1999b). Shakespeare data set. https://www.cs.uic.edu/~wilkinson/TheGrammarOfGraphics/shakespeare.txt.