Visualizing Sets with Linear Diagrams

PETER RODGERS, University of Kent
GEM STAPLETON, University of Brighton
PETER CHAPMAN, University of Brighton

This paper presents the first design principles that optimize the visualization of sets using linear diagrams. These principles are justified through empirical studies that evaluate the impact of graphical features on task performance. Linear diagrams represent sets using straight line segments, with line overlaps corresponding to set intersections. This work builds on recent empirical research which establishes that linear diagrams can be superior to prominent set visualization techniques, namely Euler and Venn diagrams. We address the problem of how to best visualize overlapping sets using linear diagrams. To solve the problem, we investigate which graphical features of linear diagrams significantly impact user task performance. To this end, we conducted seven crowd-sourced empirical studies involving a total of 1760 participants. These studies allowed us to identify the following design principles, which significantly aid task performance: use a minimal number of line segments, use guide-lines where overlaps start and end, and draw lines that are thin as opposed to thick bars. We also evaluated the following graphical properties which did not significantly impact task performance: colour, orientation, and set-order. The results are brought to life through a freely available software implementation that automatically draws linear diagrams with user-controlled graphical choices. An important consequence of our research is that users are now able to create effective visualizations of sets automatically, thus improving human-computer interaction.

General Terms: 500 Human-centered computing Empirical studies in HCI

Additional Key Words and Phrases: Sets, Visualization, Linear Diagrams

1. INTRODUCTION

The volume of data available to society has rapidly risen over recent years and, consequently, so has the level of interest in analyzing it. As a reflection of this interest, research into information visualization is both important and timely [Alsallakh et al. 2014]. In this paper, we are concerned with data whose items lie in overlapping sets, which arises in many situations [Ahn et al. 2010]. For instance, this data arises in criminal investigations where those people under investigation (the data items) may belong to multiple organizations or frequent common locations (the sets). Similar data occur in biological settings [Dinkla et al. 2014] where sets represent shared features of genes. Overlapping sets also arise in social networks where, for example, sets are formed of people with common interests [Wasserman and Faust 1994].

Linear diagrams represent overlapping sets and were introduced by Leibniz in 1686 [Couturat 1903], with parallel bargrams [Wittenburg et al. 2001] double decker plots [Hofmann et al. 2000] and UpSet [Lex et al. 2014] being similar. Each set is represented as one or more line segments, with all sets drawn in parallel. Where lines overlap, the corresponding intersection of sets contains an element that is not in any of the remaining sets. Moreover, between them the overlaps represent all of the non-empty set intersections. As an example, consider the linear diagram in Fig. 1 which displays information about languages spoken by a group of people. This diagram displays information about seven sets: French, German, Hungarian, Italian, Spanish, Turkish, and Welsh. Because there is an overlap in the linear diagram that involves just the lines for Italian and Spanish (on the righthand side of the figure), there is someone that speaks both of these two languages but none of the others. Set complements are represented by the absence of line segments. For instance, the overlap including just

Author's addresses: Peter Rodgers, University of Kent; Gem Stapleton and Peter Chapman, University of Brighton.

the lines for Italian and Spanish formally represents the set

$$\text{Italian} \cap \text{Spanish} \cap \overline{\text{French}} \cap \overline{\text{German}} \cap \overline{\text{Hungarian}} \cap \overline{\text{Turkish}} \cap \overline{\text{Welsh}},$$

where $\overline{\text{French}}$ represents the complement of French, and so forth. In this overlap, the lines for the complemented sets are absent. Linear diagrams do not, however, represent other properties of sets such as their cardinality or provide information about the sets' elements[1].

Fig. 1. Visualizing sets: linear diagrams.

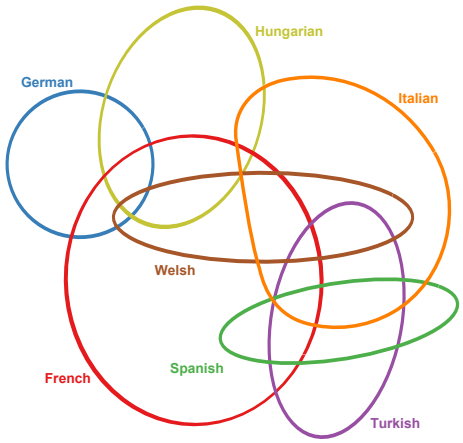

Fig. 2. Visualizing sets: Euler diagrams.

Some other visualizations of sets, such as LineSets [Alper et al. 2011] and Bubble Sets [Collins et al. 2009], place lines or contours over existing data items. Therefore, the existence of embedded items is required for these techniques. As such, members of this class of techniques are not direct competitors for linear diagrams, which do not display any individual items. Node-link techniques, such as PivotPaths [Dörk et al. 2012] also rely on the existence of items, using links between items to indicate shared set membership. Matrix based methods exploiting Karnaugh maps have been developed, such as KMVQL [Huo 2008], but they are typically considered to be restricted to 6 sets or fewer [Alsallakh et al. 2014] meaning that they lack general applicability. Some aggregation techniques cope better with larger number of sets, including Set'o'grams [Freiler et al. 2008] and Radial Sets [Alsallakh et al. 2013]. Set'o'grams are designed to be used in an interactive environment, only revealing set intersections on user input. Radial Sets represent set intersections by hyperedges connecting the sets of interest, which can rapidly increase the clutter in the diagram even with a relatively small number of sets.

The most common methods for visualizing overlapping sets are based on Euler diagrams (or variants, such as Venn diagrams), which employ overlapping closed curves to represent overlapping sets [Collins et al. 2009; Riche and Dwyer 2010; Rodgers et al. 2008; Set Visualiser 2014; Simonetto et al. 2009; Stapleton et al. 2011; Stapleton et al. 2009]. An example Euler diagram can be seen in Fig. 2, which visualizes the same sets as the linear diagram in Fig. 1. Since there is a region inside both Italian and Spanish,

[1]Formally, the expressiveness of linear diagrams is equivalent to a fragment of monadic first-order logic without equality. They are capable of expressing the emptiness or non-emptiness of sets, including sets formed by using the intersection, union and complement operators. Through this, linear diagrams directly express set-theoretic properties such as subset, equality, and disjointness.

but outside the remaining curves, the diagram asserts that somebody speaks these two languages but not the remaining languages.

There are well-known difficulties when representing sets using Euler diagrams. Their automated layout is computationally complex and even for relatively simple data sets (in the sense that few overlapping sets are to be visualized) the resulting layouts can compromise users' ability to perform tasks. This is because the automated techniques produce diagrams that possess properties that are known to hinder cognition [Rodgers et al. 2012]. These properties, sometimes called well-formedness conditions, include concurrency between curves, points that are passed through more than twice by the curves (called *triple points*), points at which curves meet but do not cross, and representing each set by more than one curve; the Euler diagram in Fig. 2 has two triple-points. Indeed, any Euler-diagram-based technique that is capable of visualizing an arbitrary collection of sets necessarily produces some diagrams that break one or more well-formedness condition [Flower and Howse 2002].

It has been established, empirically, that users can perform set-theoretic tasks significantly more accurately and significantly faster using linear diagrams than when using Euler diagrams [Chapman et al. 2014; Gottfried 2015]. In Chapman et al.'s empirical study, the Euler diagrams were drawn to adhere to as many of the well-formedness conditions as possible, whilst accurately visualizing the sets, to ensure task performance was not overly compromised. Moreover, other graphical choices were adopted for the study's Euler diagrams that are known to aid task performance [Blake et al. 2014a; Blake et al. 2014b]. By contrast, there are no known results about how the graphical choices made when drawing linear diagrams affect task performance. As choices necessarily have to be made when drawing linear diagrams, [Chapman et al. 2014] made assumptions about what constituted effective layouts.

In order to take full advantage of the cognitive benefits of using linear diagrams, in this paper we address the problem of how to best visualize overlapping sets using linear diagrams. In particular, we identify graphical choices that should be made when drawing linear diagrams in order to significantly improve task performance, in terms of accuracy and time. The paper makes the following specific contributions:

— We applied Bertin's Semiology of Graphics to linear diagrams. This allowed us to identify six graphical features that could impact on task performance in the context of linear diagrams, namely: size, colour, texture, orientation, relative horizontal positioning, and relative vertical positioning.
— We developed hypotheses about the impact of these graphical features. We performed six studies, one for each graphical feature, and have provided corresponding statistical analysis.
— We derived a set of design principles for linear diagrams, corresponding to the graphical features that led to improved task performance.
— We established the combined effect of the design principles, by conducting a final, seventh, study performed with various combinations of graphical choices. We established that adhering to all or most of the design principles significantly improves task performance.
— We have ensured that our results have practical benefit by implementing an online software tool that automatically draws linear diagrams. The freely available tool allows users to select their preferred graphical choices, including those that meet all of the design principles.

As a consequence of our work, human-computer interaction is improved: linear diagrams can now be automatically drawn, following empirically validated design principles that lead to significantly more accurate and significantly faster task performance.

All of the diagrams used in the studies, along with the questions and details of the real-world data from which the diagrams were derived, can be found in the supplementary material associated with the paper. The anonymised data collected during the studies is also included with the supplementary material. The studies themselves can be taken on-line at `www.eulerdiagrams.com/linear`, where our freely available software can also be found.

The paper is structured as follows: In section 2, we present our analysis of Bertin's Semiology of Graphics in relation to linear diagrams. In section 3 we present hypotheses about the impact of the six graphical features derived from the analysis. Section 4 overviews how the six studies, one for each graphical feature, are executed. Sections 5 to 10 present the statistical analysis of the data collected in the six studies. The results are summarized in section 11, from which a set of design principles for linear diagrams are derived. The seventh study is presented in section 12 which evaluated various combinations of graphical choices. Section 13 discusses the threats to the validity of our studies. The software implementation is discussed in section 14. Finally, section 15 gives our conclusions and discusses further work.

2. PERCEPTUAL THEORIES AND GRAPHICAL CHOICES

There are many graphical features of linear diagrams that can be varied. The purpose of this section is to identify a set of graphical choices for linear diagrams that may impact on user comprehension. In order to make such an identification, it is important to take account of set-theoretic tasks that users will perform when interpreting linear diagrams. Considering the task taxonomy in [Alsallakh et al. 2014], we have identified three major set-theoretic tasks that can be performed using linear diagrams:

(1) **Subset:** establish whether one set is a subset of another, i.e. all members of one set are also members of another set. In Fig. 3, the linear diagram represents information about social media topics and the sets represent the interests of people. In detail, this linear diagram expresses that Games is a subset of Android: everyone with interests in Games is also interested in Android. To extract this information, the user has to identify that each of the overlaps that include the set Games also includes the set Android.
(2) **Intersection:** establish whether two sets intersect, i.e. the two sets share a member. In Fig. 3, the linear diagram expresses that someone has interests in both Android and Hifi. To extract this information, the user has to identify that one of the line segments that represents the set Android overlaps with one of the line segments that represents Hifi.
(3) **Disjoint:** establish whether two sets are disjoint, i.e. the two sets do not share any members. In Fig. 3, the linear diagram expresses that Games and Internet are disjoint: nobody has interests in both Games and Internet. To extract this information, the user has to identify that the line segments that represent the set Games never overlap with the line segments that represent the set Internet.

Fig. 4 shows an alternative linear diagram, where different choices have been made, that represents the same information as Fig. 3. It may be easier to perform the tasks just described with Fig. 4, since the number of line segments has been reduced and, so, fewer comparisons need to be made. This is just one graphical feature that can be varied.

Bertin divides graphical features into elements and their properties, with his Semiology of Graphics [Bertin 1983] being widely regarded as one of the classical works of graphical visualization. In linear diagrams, line segments are graphical elements. Their properties include orientation, length, width and colour. Bertin highlights our sensitivity to properties such as these, calling them retinal variables. In terms of

Fig. 3. Performing tasks with linear diagrams.

Fig. 4. A different layout with fewer line segments.

Bertin's theory, manipulating these retinal variables impacts on the effectiveness of a visualization. Bertin's retinal variables are: size, colour value, colour hue, texture, shape, and orientation. In addition, Bertin recognizes that we are also sensitive to planar variables, which correspond to the relative position of graphical elements.

We exploit Bertin's work as a source for identifying important graphical choices for linear diagrams. Whilst there is not always a direct correspondence between Bertin's graphical properties and those exploited by linear diagrams (for instance, colour value and texture do not readily apply), they provide significant motivation for the graphical choices studied in this paper.

We also make reference to the Gestalt principle of good form [Koffka 1935]. This principle tells us that there is a tendency for people to group together graphical objects that share some feature, like colour or shape.

2.1. Size

In the context of linear diagrams, size corresponds to the length and width of the line segments. There are two clear choices for line width: thin lines, giving a high proportion of white space in the diagram, or thick lines – essentially bars – where adjacent lines are touching. These two choices can be seen in Figs 5 and 6. White space may give the impression of less visual clutter (Fig. 5). Touching bars may aid the comparison of different line segments, an activity required for subset, intersection and disjointness tasks (Fig. 6). We examined the impact of line width on task performance. By contrast, we did not examine different choices of line lengths, since they can be arbitrarily long or short, with no constraints imposed on the lengths given the other graphical choices that are made when laying out the diagram.

We note that size is often recognized as a powerful variable to control when visualizing *quantitative* differences. Size can, therefore, be exploited to convey cardinality information (i.e. to indicate how many members each set contains): line length or width could correspond, proportionally, to set cardinality. However, reflecting the fact that no studies have been undertaken to ascertain how to effectively draw linear diagrams at all, we are focussed on the simpler case where cardinality information is not visualized. As such, one could argue that it is misleading to use a range of line widths to represent the sets or to vary the lengths of the overlaps: this could wrongly imply different set cardinalities or set intersection cardinalities respectively. Therefore, there is a good reason to use a fixed line thickness and fixed overlap length.

We therefore have the following question:

QUESTION 1. *Should linear diagrams be drawn with thin lines or with thick touching lines (bars)?*

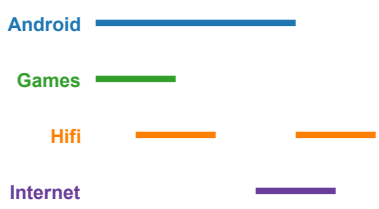

Fig. 5. A linear diagram drawn with thin lines.

Fig. 6. A linear diagram drawn with thick bars.

2.2. Colour Hue and Colour Value

Colour hue (different hues have different colours) is recognized as important when visualizing *qualitative* differences [Card et al. 1999; Leborg 2006; Mazza 2009], backed up by the Gestalt principle of good form. Therefore, varying colour hue could impact on task performance in the case of linear diagrams. The following discussion identifies three colour hue treatments that we investigated.

In our case, each set is represented by a collection of collinear line segments. Assigning unique colours to each set, rather than just using black say, could reinforce the fact that collinear line segments are being used to represent a single set. It is, however, unclear whether such a colour treatment is a secondary irrelevant variable as the collinearity of the line segments already indicates the common property of representing the same set. Thus, we chose to evaluate the following colour treatments: all sets assigned the colour black (i.e. black line segments) versus each set being assigned a unique colour (i.e. coloured line segments).

An alternative way of treating linear diagrams with colour is to assign unique colours to the *overlaps*. In this way, the collinearity of line segments indicates that they represent the same set and colour is used to aid the set-theoretic tasks described above, where it is necessary to determine whether line segments overlap. With this way of treating linear diagrams with colour, two line segments overlap if and only if they share a common colour. Again, this alternative colour treatment is consistent with the principle of good form, at least for the set intersections, and may aid with performing tasks. For instance, if two groups of line segments share a common colour then it is perhaps easier to see that the sets they represent have a non-empty intersection. Likewise, having no common colour may aid the detection of a disjointness relationship. This colour treatment is not, therefore, using colour as a secondary variable but as a visual clue that sets share elements. There is the potential, however, with this third colour treatment that the use of multiple colours assigned to each set actually counters the advantage of exploiting collinearity to indicate line segments represent the same set: it breaks the principle of good form at the set-level.

The three colour treatments can be seen in Figs 7, 8, and 9. There is clear tension between these uses of colour. The set-theoretic tasks require the relevant sets and the relevant intersections (or their absence) to be identified. It is unclear whether the principle of good form, and corresponding colour treatments, applies more strongly to the line segments or the line overlaps.

Colour value (brightness) is recognized as being an important variable when representing *quantitative*, as opposed to qualitative, data. As such, there does not seem to be any motivation to empirically establish the impact of varying colour value when drawing linear diagrams given the tasks we used in our studies. However, this use of

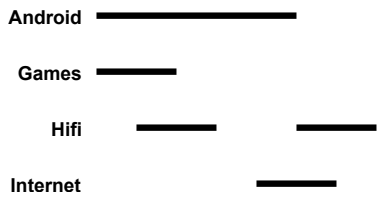

Fig. 7. A linear diagram drawn with black lines.

Fig. 8. A linear diagram drawn with coloured line segments.

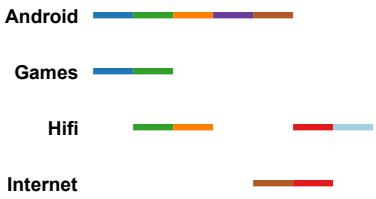

Fig. 9. A linear diagram drawn with coloured overlaps.

colour could be an important extension to the design of linear diagrams if we wish to increase their expressiveness to convey cardinality information.

We therefore posit a further question about colour hue:

QUESTION 2. *How should we treat linear diagrams with colour: assign black to each set, assign a unique colour to each set, or assign a unique colour to each set intersection (i.e. overlap)?*

2.3. Texture and Guide-Lines

It is possible to augment the syntax of linear diagrams to indicate where overlaps begin and end. Figs 10 and 11 show how guide-lines can be used. Guide-lines appear to have an advantage over the suggested use of colour to group overlaps: it still allows colour to be used to promote the identification of line segments representing the same sets (Fig. 8). Using guide-lines allows the principle of good form to be met by the sets (collinear line segments have a unique colour) *and* met by the set intersections (each overlap is in a unique vertical column formed by the guide-lines). However, guide-lines increase the visual complexity of diagrams so their actual benefit remains unclear.

Texture is normally recognized as a useful variable when representing both qualitative and quantitative differences. Linear diagrams could be extended to use texture for encoding additional information, such as indicating relative set cardinality. Using texture in this way, whilst interesting, does not apply to the tasks identified for our studies.

Thus, we have the following question:

QUESTION 3. *Should we use guide-lines to indicate the start and end of overlaps?*

2.4. Shape

Shape is a variable that can be exploited to visualize qualitative data [Card et al. 1999; Leborg 2006; Mazza 2009]. In linear diagrams, each line segment is straight and it may be misleading to assign different shapes to different lines (e.g. using wiggly vs straight lines). Indeed, an important feature of linear diagrams is the fact that the collinear

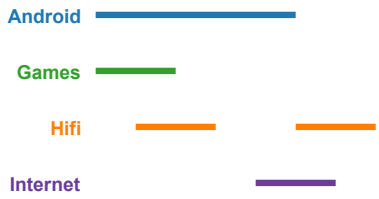

Fig. 10. A linear diagram drawn without guidelines.

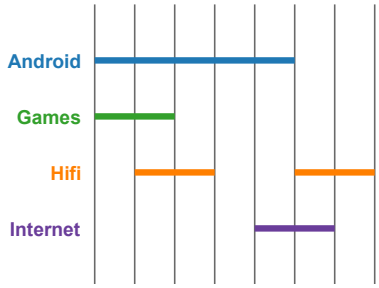

Fig. 11. A linear diagram drawn with guidelines.

line segments for pair of each sets are drawn in parallel (and, thus, have the *same* shape), allowing easy identification of the overlaps. This is evidenced by [Wagemans et al. 2012], who cite [Feldman 2007] as the source of this insight:

> "[the] comparison of features lying on pairs of line segments is significantly faster if the segments are parallel or mirror symmetric suggesting faster grouping of the segments based on these cues."

Thus, there seems to be little value in empirically testing whether using different shapes for the lines impacts user understanding.

2.5. Orientation

Bertin argued that we are also perceptually sensitive to the orientation of visualizations. Varying the orientation of a linear diagram corresponds to choosing the direction in which to draw the lines. For example, the horizontal or vertical orientation of a linear diagram could impact on the ease with which sets and the overlaps are identified. Two contrasting orientations are shown in Figs 12 and 13; notice that both diagrams preserve the alphabetic order of the sets and the order in which colours are assigned. Thus, we ask:

QUESTION 4. *How should we orientate a linear diagram, horizontally or vertically?*

Fig. 12. A linear diagram drawn horizontally.

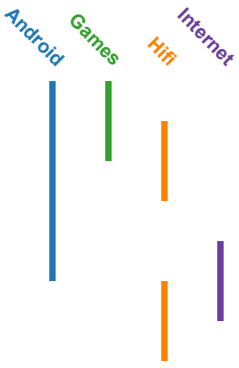

Fig. 13. A linear diagram drawn vertically.

2.6. Planar Variables

As well as the retinal variables just discussed, Bertin also identified that planar variables impact perception. For visualizations drawn in two dimensions, including linear diagrams, the choice of planar variables correspond to the *relative* positioning of syntactic items in the plane. Given a linear diagram with line segments drawn horizontally, the order in which the sets are visualized is a vertical positioning choice. It may be beneficial to represent the sets in alphabetical order, since this could aid the identification of the relative position of sets in the list. However, other orders may be sensible in terms of aiding task performance. In particular, it may be helpful to have no other sets drawn between two sets that have collinear end-points: diagram semantics are, in part, determined by the line end-points.

Taking the distance between two sets to be the number of sets drawn in between them, we ask whether we should minimize the distance between sets represented by line segments that have collinear end-points. A layout with these characteristics, as opposed to alphabetic order, has an *adjacency-driven set-order*. These two different choices of set-orders are shown in Figs 14 and 15; the order of sets has changed, but the order in which the colours are assigned has remained fixed. In Fig. 15, Android is drawn between Hifi and Games: Android has line segments that share end-points with those representing these other two sets. Hence this layout minimizes the distance between Android and Hifi as well as Android and Games.

Fig. 14. A linear diagram drawn with alphabetical set-order.

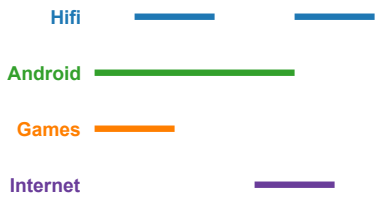

Fig. 15. A linear diagram drawn with adjacency-driven set-order.

So, we ask the following question:

QUESTION 5. *Should we represent the sets in alphabetical order or use adjacency-driven set-order?*

With a horizontal layout, the order in which the overlaps are visualized is a horizontal positioning choice. We are free to choose any order of the overlaps, without altering the semantics of the linear diagram. This order has a profound effect on the number of line segments required. Choosing a layout that minimizes the number of line segments may be beneficial because fewer line segments need to be compared when performing tasks. As illustrated above with Figs 3 and 4, fewer comparisons need to be made to establish set theoretic relationships, possibly improving accuracy and reducing the time taken to perform set-theoretic tasks. This leads to our final question:

QUESTION 6. *Should we minimize the number of line segments?*

Producing linear diagrams with a minimal number of line segments is computationally complex. In our empirical studies, we adopted a heuristic approach to produce diagrams with a reduced number of line segments; further details are given in section 14.

2.7. Summary

By varying the choices made when drawing linear diagrams, we could positively or negatively impact users' ability to perform set-theoretic tasks. Through an examination of Bertin's characterization of graphical features, we have posed six questions concerning choices made when laying out linear diagrams. Whilst Bertin's work, and in some cases the Gestalt principle of good form, may lead us to hypothesize which graphical choices lead to more effective linear diagrams, we have no empirical insight into the impact of these choices on task performance. As such, we proceeded to empirically evaluate these choices, yielding answers to the posed questions.

3. HYPOTHESES AND METHODOLOGY

Our aim was to determine a set of design principles for linear diagrams that allow users to effectively obtain information about sets. We did this by measuring task performance in terms of accuracy and time. Of these two performance measures we viewed accuracy as more important than completion time, consistent with other researchers, such as [Alper et al. 2011]. We judged 'most effective' as follows: one graphical choice is more effective than another if users perform tasks significantly more accurately with it or, when no significant difference in accuracy exists, perform tasks significantly faster with it.

To formulate our hypotheses, we started by considering how one may perform set-based tasks. Users may be attempting to answer one of the following three types of questions, given a set X:

Subset: Which sets are subsets of X? This corresponds to the syntactic problem of identifying all of the sets where each line segment is drawn concurrently with the line segments for X.
Intersection: Which sets intersect with X? This corresponds to the syntactic problem of identifying all of the sets that have a line segment which overlaps with at least one line segment for X.
Disjointness: Which sets are disjoint with X? This corresponds to the syntactic problem of identifying all of the sets that have only line segments which do not overlap with any of the line segments for X.

The actual wording of the tasks that we asked participants to perform in our studies is given in section 3.1.

Answering any one of these questions required two activities to be undertaken:

(i) find the names of the sets specified in the question, written in the diagram, and
(ii) compare the line segments to determine their relationship, as necessary for the question type.

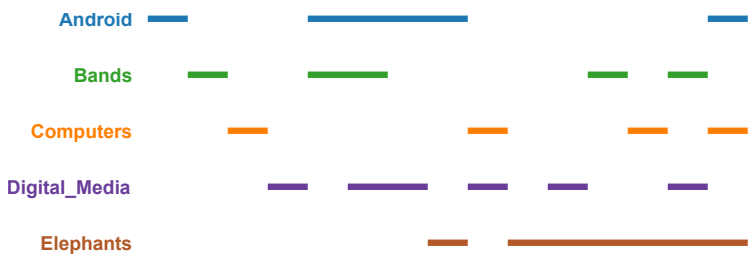

Fig. 16. Performing set-theoretic tasks.

In Fig. 16, suppose we wish to determine whether Bands and Elephants are disjoint. In the diagram we need to find, therefore, the words Bands and Elephants (activity (i)). Scanning the first line segment for Bands – assigning 'primary attention' to Bands – we can see that it does not overlap with any part of a line segment for Elephants (activity (ii)). The same holds for the second line segment. The third line segment for Bands does overlap with a line segment for Elephants, establishing that these two sets are not disjoint. Other reading orders could be followed, such as giving primary attention to Elephants instead of Bands or, after looking at the first two line segments for Bands, switching one's primary attention to Elephants as the next x-coordinate to contain either a Bands or an Elephants line segment arises with Elephants. Other ways of reading the diagram also exist, but the semantics of the diagram always require the two activities as described above to be undertaken in order to perform the task. There is potentially a frequent need to perform activity (ii), leading us to posit that graphical choices that could impact users' ability to do so are more likely to have a significant impact on task performance.

Activities (i) and (ii), along with our discussions in section 2 on perception, led us to the following hypotheses, presented in the same order as their corresponding question in section 2. In each case, we conjecture whether the graphical feature under discussion would have a high, medium, or low impact on task performance, relative to the other graphical features under consideration.

H1 *Linear diagrams should be drawn with bars as opposed to lines.* Whilst we conjecture that white space is important, leading to lower visual clutter, having touching bars may aid activity (ii). When wishing to ascertain whether $Games$ is a subset of $Android$, Fig. 6 may be more helpful than Fig. 5 due to the touching bars. However, when bars are not adjacent, it is unclear whether there is any significant benefit over lines. We, thus, conjecture that using bars may have a *low impact* on assisting with activity (ii).

H2 *Linear diagrams should not be drawn with black lines.* Whilst it is unclear which colour treatment is preferable, we believe that using colour in some way will aid task performance. First, *performing* activity (i) does not appear to be aided by any one of the three proposed colour treatments. Moreover, using black lines does not appear to aid with activity (ii).

If the sets are assigned a unique colour then the relative start and end points of the line segments still need to be compared, activity (ii). However, the colours can be used to 'keep track' of where the sets are relatively positioned, rather than having to refer back to the set names if the relative position is forgotten, reducing the need to repeatedly undertake activity (i) whilst undertaking activity (ii). This motivates our hypothesis that assigning colours to sets, therefore, aids task performance relative to black lines, since it is conjectured to reduce the cognitive load on the user.

If unique colours are assigned to overlaps, performing set-theoretic tasks no longer *requires* the start and end points of line segments to be compared. Activity (ii) can be reduced to seeing whether, for each task type, the following occur in the linear diagram:

(a) subset: the line segments for one set use only colours assigned to the other set,
(b) intersection: line segments share a common colour, and
(c) disjointness: line segments share no common colours.

The user still needs to be aware of the relative position of the sets – related to activity (i) – but no longer has to compare the start and end points of the line segments – activity (ii). This second way of using of colour appears to reduce the cognitive load on users, aiding task performance relative to black lines.

Since the coloured sets and coloured overlaps treatments impact on the two activities, it is unclear which allows users to perform best overall. In summary, it is possible that using colour, as opposed to black lines, will have a *high impact* on assisting with activity (ii).

H3 *Linear diagrams should be drawn with guide-lines assigned to the ends of the overlaps.* The use of guide-lines is thought to assist with activity (ii) relative to the absence of guide-lines. However, we anticipated that guide-lines do not aid activity (ii) as much as using colour, since the guide-lines give rise to visual clutter (i.e. linear diagrams drawn with guide-lines contain more syntax). Moreover, since the guide-lines could help compare sets regardless of whether they are adjacent in the diagram, we expect them to be more useful than using bars instead of lines. In summary, we conjectured that guides-lines have a *medium impact* on task performance.

H4 *The vertical or horizontal orientation of linear diagrams does not impact on user performance.* It is possible that reading the labels, activity (i), is easier with a horizontal orientation, at least for people who read left to right. We conjectured that orientation has a *low impact* on task performance.

H5 *Linear diagrams should be drawn with adjacency-driven set-order.* Placing set names in alphabetical order is likely to aid users when finding them in the diagram, embodied in activity (i). However, using adjacency-driven set-order is thought to aid activity (ii), since the sets are drawn closer to each other when they have line segments with collinear end-points. To justify this, consider the line segments for Android and Hifi in Figs. 14 and 15. The second line segment of Hifi has an end-point that is collinear with an end-point of the line segment for Android. There is more white space between these two end points in Fig. 14 than in Fig. 15, potentially increasing the difficulty of determining that these two line segments do not overlap. However, in general such a layout may not make other comparisons easier than an alphabetic layout. Thus, an adjacency-driven choice may not assist with activity (ii) as much as the use of colour. We therefore conjectured that this graphical choice has a *medium impact* on task performance.

H6 *Linear diagrams should be drawn using a minimal number of line segments.* As indicated above, using a minimal number of line segments reduces the number of comparisons that need to be made when performing activity (ii). As fewer comparisons need to be made, task accuracy and time taken should be assisted. The number of overlaps is a global property of the diagram, so it aids with the comparison of many sets, not just the comparison of sets drawn in close proximity as was the case with bars versus lines and adjacency-driven set-order. Also, minimizing the number of the line segments *reduces* visual clutter, unlike bars which increase visual clutter. These two reasons are why we argue that the number of line segments could have a *high impact* on task performance.

Our experiment design, described below, allowed us to determine which graphical choices significantly impact performance in terms of accuracy and time taken.

3.1. Tasks

We felt that it was important that each study had a diversity of tasks in order to provide a rounded insight into the relative *overall* performance of each graphical choice. With this in mind, we also used ‘simple’ and ‘complex’ versions of each task type:

Subset
(a) *Simple:* identify all of the subsets of X
(b) *Complex:* identify all of the subsets of $X \cup Y$.
Intersection

(a) *Simple:* identify all of the sets that intersect with X.
(b) *Complex:* identify all of the sets that intersect with $X \cup Y$.
Disjointness
(a) *Simple:* identify all of the sets that are disjoint from X.
(b) *Complex:* identify all of the sets that are disjoint from $X \cap Y$.

There is a notable difference between the complex questions for subset and intersection and the complex questions for disjointness: the former use union and the latter uses intersection. This differing choice was made to reduce the number of occurrences for which 'none' was the answer. Moreover, asking participants to identify all of the sets in the diagram for which the particular set-theoretic relationship holds ensures that the question requires participants to look at parts of the diagram that have been affected by the graphical choice.

The participants could select their answers using check boxes in our data collection software. The check boxes for each question included all of the sets in the associated linear diagram, except for those stated in the question, alongside a 'none of the above' option. A complex instance of each task type is as follows, which could be asked of Fig. 16:

Subset: Tick the check boxes where **all** of the people are also interested in **either** Bands **or** Elephants. (Boxes to be ticked: none of the above.)
Intersection: Tick the check boxes where **some** of the people are also interested in **either** Computers **or** Digital_Media. (Boxes to be ticked: Android, Bands, and Elephants.)
Disjointness: Tick the check boxes where **none** of the people are also interested in **both** Computers **and** Elephants. (Boxes to be ticked: Bands.)

Fig. 17 shows how questions were displayed to participants.

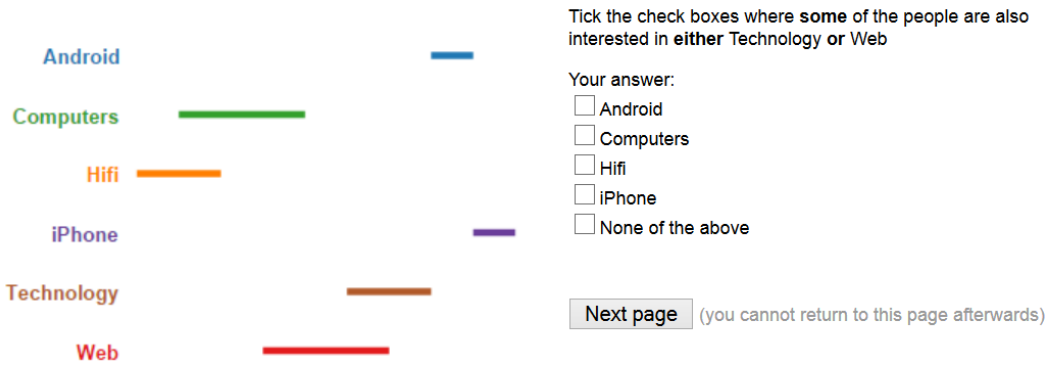

Fig. 17. A screenshot showing how the tasks were displayed.

Each study that we ran included four of each type of task (two simple, two complex), giving 12 questions in total. In addition, the linear diagrams used in our studies included either six or ten sets, to ensure that the data to be visualized were not trivial. Within each task type, one simple and one complex task was assigned to a six-set linear diagram, likewise with respect to the ten-set linear diagrams. As well as these main study questions, a further four questions were used to train the participants (a simple subset task with three sets, a simple intersection task with six sets, a complex disjoint task with ten sets, and a complex intersection task with ten sets). The *same training questions* were used for each study, but *the linear diagrams were tailored* to

the study being undertaken. A further two questions were asked of participants in order to identify whether they were *inattentive*; this is further discussed in section 3.3. So, in total, each participant was asked 18 questions: four training questions, 12 main study questions, and two inattentive participant identifiying questions.

The participants were given the following instructions on starting the study:

> You should maintain concentration on the HIT and answer questions without delay, unless a question explicitly allows you to take a break, in which case you can have a rest before continuing.

The terminology HIT is explained in section 3.3.

3.2. Data for Visualization

To perform our studies, we required diagrams for training participants and for collecting performance data during the main study phase. The questions used in the main phases of the studies were asked of linear diagrams generated from real-world data. In particular, we derived 84 data sets from Twitter, obtained from the SNAP data set collection [Leskovec 2011]. This collection contains ego-networks of 1000 Twitter users. For each Twitter list, a set is formed comprising the users that subscribe to that list. As study participants need not be familiar with Twitter, the questions used in the study made no reference to Twitter. Moreover, to avoid any possibility of previous knowledge of the data impacting the results, all set names were changed, but keeping a real-world scenario: the interests of people. The set names were chosen so that no two sets in any one linear diagram started with the same letter to reduce the potential for misreading set names.

For the main questions in each study, 12 of these 84 data sets were randomly selected; different data sets were used for each study to ensure that participants could not obtain information on the correct answers from those who took part in earlier studies. If the data set included more sets than required (i.e. six or ten) then the appropriate number of sets were removed. This removal process selected sets that reduced the number of overlaps minimally, thus maintaining the complexity of the data set as much as possible. In addition, we required that the six-set and ten-set diagrams had at least seven overlaps and 11 overlaps respectively, and at most 30 overlaps in total. The lower bound on set overlaps ensured the diagrams were not completely trivial, whilst the upper bound was necessary to limit their size from a pragmatic perspective: the diagrams needed to comfortably fit on computer screens without the need for scrolling. This consideration was important for the study, but does not reflect an inherent limitation of linear diagrams. The linear diagrams were then generated from these data sets; later, we present a profile of the data sets for each study, to convey their complexity further.

The linear diagrams used for the training questions and for identifying inattentive participants were generated from randomly produced data sets that were not derived from the Twitter data. This allowed us to ensure that the diagrams were suitable for training and for identifying inattentive participants.

3.3. Data Collection Methods

We adopted a crowdsourcing approach, using Amazon Mechanical Turk (MTurk) [Chen et al. 2011; Paolacci et al. 2010] to automatically out-source tasks to participants. The tasks, called HITs (Human Intelligence Tasks) are completed by anonymous participants who are paid on *successfully* completing the HIT. Crowdsourcing is becoming more popular as a method for conducting research-oriented studies. There is evidence that crowdsourcing is a valid approach for collecting data and it has gained recognition within the scientific community [Heer and Bostock 2010; Paolacci et al. 2010]. The

HITs were based on the templates provided by [Micallef et al. 2012][2]. Every question, in both the training and the main study, was displayed on a separate page of the HIT. Previous pages could not be viewed and subsequent pages were not revealed until the question on the current page was answered. Unlike the training questions, in the main study the questions were randomly sequenced.

There is little control, in MTurk, over who participates in the study and, so, some participants may fail to give questions their full attention [Chen et al. 2011] or have difficulties with the language; we call these participants *inattentive*. To reduce the impact of language issues, a system qualification was used, allowing only participation from people based in the USA with a HIT approval rate of 95%.

Another recognized technique for identifying participants who cannot understand the language used, or who are not giving the tasks sufficient attention, is to include questions that require careful reading, yet are very simple to answer [Oppenheimer et al. 2009]. In our study, we included two such questions which asked participants to simply click on the diagram, whilst still presenting them with redundant check boxes as seen for the 12 main study questions; these catch questions appeared as the third and ninth questions after the training phase. Participants were classified as inattentive if they clicked check boxes on either of the two inattentive participant identifying questions. All data obtained from inattentive participants was removed before analysis.

4. EXECUTION OF THE STUDIES AND DATA COLLECTED

Each of our six hypotheses relate to different graphical choices, so to test all combinations (of which there are $2 \times 3 \times 2 \times 2 \times 2 \times 2 = 96$) of choices is clearly not practical. Our approach, therefore, is to test them in sequence, ordered by those which we conjecture to have the most significant effect first[3].

We identified H2 (on the use of colour) and H6 (on the number of line segments) as potentially having a high impact on task performance. Therefore, we prioritized testing these two hypotheses. We believed that minimizing the number of line segments would bring more performance benefits than using colour because of the substantial impact the number of line segments has on diagram clutter. Therefore, we tested H6 first, then H2. For the first test, we had to choose the graphical features of the linear diagrams. Given that linear diagrams are already used, we adopted the graphical choices already seen in the literature [Chapman et al. 2014] and varied only the number of line segments, as required for H6. As the tests proceeded in sequence, we adopted the graphical choices implied by our studies, in order to reach a final design.

After conducting the first two studies, we then focused on the two hypotheses related to graphical choices that we conjectured to have a medium impact on task performance: H3 (the use of guide-lines), and H5 (the order of sets). We believed that using guide-lines would have more impact than set-order, since they are thought to aid the task performance when comparing any pair of line-segments. By contrast, an adjacency-driven set-order more naturally aids the comparison of line segments drawn close to each other. Therefore, our third study tested H3 and our fourth study tested H5.

Our final two studies focused on graphical choices that we conjectured to have a low impact on task performance: H1 (the use of bars versus lines) and H4 (orientation). We anticipated that there could be an effect of using bars over lines. H4, by contrast,

[2]Available from `http://www.aviz.fr/bayes`.

[3]One could make many different choices about the order in which to determine the effect of graphical choices on linear diagrams. We do not argue that our approach is the only way, just that it is a reasonable method. One could choose alternative orders in which to evaluate the graphical choices, possibly with different results.

predicted that there would be no significant difference between diagram orientations. Thus, our fifth study investigated H1 and our sixth study focused on H4.

The *major goal* of each study was to determine whether there are overall differences in performance between treatments. To this end, for each question asked in the study, we recorded the check boxes selected (accuracy data) and the time taken to submit an answer. A correctly selected check box was counted as one correct answer. A check box that was correctly not selected was also counted as one correct answer. The remaining responses were incorrect answers. This provided us with categorical data which was subjected to chi-square tests. The time data allowed us to compare time performance across treatments. In all of our studies, the time data collected was not normal. In each case, using $\log_{10}(\text{time})$ results in data that, whilst still not normal, has little skewness and, thus, could be used to conduct a robust ANOVA. For both the ANOVA and chi-square tests we take p-values of less than 0.05 to be significant.

Each of the six studies included a pilot phase, with 10 participants per group. Any adjustments that were necessary to the study materials are detailed in the relevant sections below; we do not discuss the pilot studies if no problems came to light. Each main study included 100 participants per group, less any identified as inattentive. Each participant was paid $1 for taking part and randomly allocated to a treatment group. Participants could only take part in a study once and could not take part in more than one study. This was enforced by recording participant identifiers (called worker IDs).

The next six sections correspond to the studies for hypotheses 1 to 6. We present a comprehensive account for the first study, on line segments, as follows: an example question from the study, information on diagram complexity, participant demographics, analysis of errors, analysis of time, and summary of results. For the remaining studies, the information on diagram complexity and participant demographics can be found in the appendix. A full set of diagrams used in the studies, along with the questions asked, can be found in the supplementary material. The studies can be found online at `www.eulerdiagrams.com/linear`.

5. LINE SEGMENTS

Hypothesis H6 indicates that using a minimal number of line segments could significantly improve task performance. For the study, we compared using a minimal number of line segments with a randomly chosen number of line segments. Two linear diagrams used in this study can be seen in Figs 18 and 19. Participants were asked to 'Tick the check boxes where **none** of the people are also interested in Music'. The correct solution was to tick the check boxes for Android, Bands, Design, Games, Journalism, Stars, and Technology. Table I indicates the level of complexity of the linear diagrams used in this study, stating the number of line segments used in both the minimal and random visualizations. It also states which data set was randomly assigned to which task (I=intersection, S=subset, D=disjoint, C=complex). Figs 18 and 19 are for data set 6.

Of the 100 participants recruited to each of the two groups (which we call the 'minimal group' and the 'random group'), three were inattentive and they were all in the random group. The demographics of the participants were as follows:

— **gender:** 93M, 104F, 0 other, 0 not stated
— **age range, in years:** 19 to 73 (mean: 34)
— **qualification level:** 1 not stated, 1 some high school, 18 high school graduate, 49 some college, 12 associates degree, 89 Bachelors degree, 23 Masters degree, 4 doctorate degree.

Fig. 18. A minimal number of line segments.

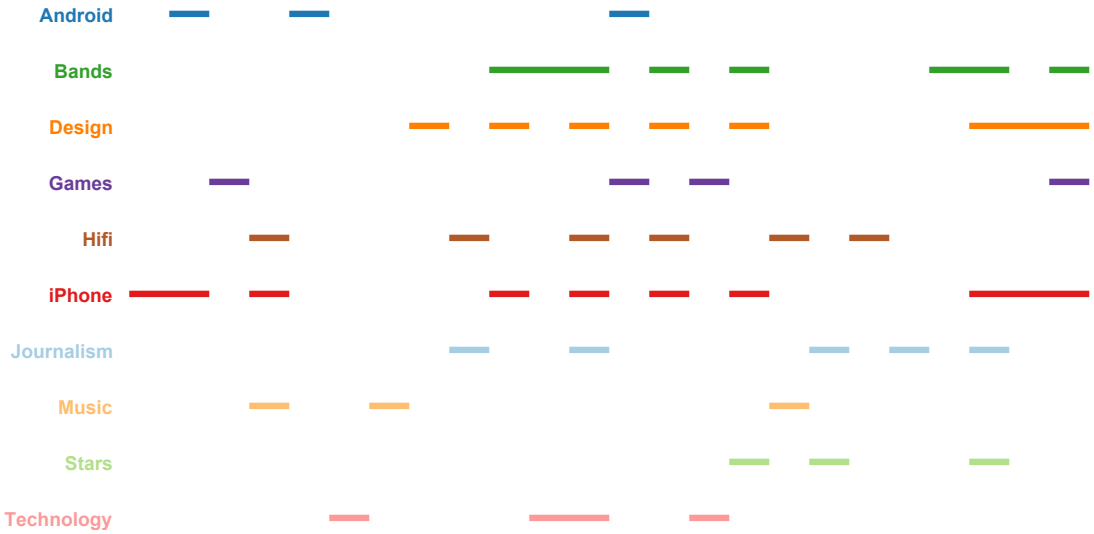

Fig. 19. A random number of line segments.

5.1. Analysis

The major goal of this first study was to establish whether significant differences exist in task performance when using a minimal number of line segments as compared to a random number of line segments. As there were three inattentive participants, the following results are based on $200 - 3 = 197$ participants each answering twelve questions, thus the total number of questions answered was $197 \times 12 = 2364$.

5.1.1. Accuracy. Each of the 12 questions had either five, six, nine or ten check boxes. For instance, a complex six-set question involves two sets in the question, leaving four sets corresponding to check boxes along with a 'none of the above' check box: such a question has five check boxes. Likewise, a simple six-set question has six check boxes

Table I. Diagram Complexity for Study 1: Minimal versus Random Numbers of Line Segments

Data Set	1	2	3	4	5	6	7	8	9	10	11	12
Sets	6	6	6	10	10	10	6	6	6	10	10	10
1-set overlaps	2	3	4	8	7	9	6	6	5	4	3	4
2-set overlaps	4	1	2	7	7	8	3	10	4	10	8	3
3-set overlaps	4	2	1	1	2	2	0	7	0	5	5	1
4-set overlaps	3	1	0	0	0	3	0	0	0	1	0	1
5-set overlaps	1	0	0	0	0	1	0	1	0	3	0	2
6-set overlaps	0	0	0	0	0	1	0	0	0	1	0	1
>6-set overlaps	0	0	0	0	0	0	0	0	0	0	0	$1\times 9, 1\times 10$
Total number of overlaps	14	7	7	16	16	24	6	24	9	24	16	13
Minimal Line Segments	10	8	6	12	12	18	6	15	6	20	12	17
Random Line Segments	21	11	11	19	25	45	10	38	9	40	20	32
Task Type	I	S	D	I	S	D	CI	CS	CD	CI	CS	CD

and so forth. Twenty five percent of the questions correspond to each number of check boxes. Therefore, since there were 2364 questions answered, each number of check boxes occurred $\frac{2364}{4} = 591$ times. This gives a total of $591 \times (5+6+9+10) = 17730$ check boxes. Some of these check boxes were correctly selected, and some were not, which allowed us to compute the accuracy rate.

In this study, the total number of correct responses was 15474 and there were 2256 incorrect responses. Table II shows the breakdown of responses by treatment, with proportionally fewer errors being accrued by the minimal group. Participants from the minimal group had error rate of 11.4% whereas this increased to 13.5% for the random group. Thus, we saw approximately two more errors for every 100 answers from the random group. Performing a chi-square test reveals significant differences between the two graphical choices ($p = 0.003$). *Therefore, using a minimal number of line segments allows participants to perform significantly more accurately than those using a random number of line segments.*

Table II. Accuracy for Line Segments: Minimal versus Random

Treatment	Correct	Incorrect	Total	Error Rate
Minimal group	7920	1080	9000	11.4%
Random group	7554	1176	8730	13.5%
Total	15474	2256	17730	12.7%

5.1.2. Time. The grand mean for the time taken to answer questions in this study was 35.81 seconds per question (standard deviation: 33.50). The average completion time for questions for the minimal group was 33.23 seconds (SD: 34.61) and for the random group the mean was 38.47 seconds (SD: 32.12). Overall, on average each question took over 6 seconds longer to complete when using a random number of line segments (an increase of 19.4%). Performing an ANOVA with $\log_{10}(\text{time})$ data (skewness: 0.24), revealed a significant effect ($p = 0.002$). *Therefore, using a minimal number of line segments allows participants to perform significantly faster than those using a random number of line segments.*

5.1.3. Summary of Analysis. Our analysis of both the accuracy data and time data has revealed that participants perform significantly better when using a minimal number of line segments as opposed to a random number of line segments. We note that the relative benefit of using a minimal number of line segments in terms of errors accrued is, in our opinion, relatively small from a practical perspective (respective error rates of 11.4% and 13.5%). However, increasing time taken by nearly 20% (over six seconds per question) we believe represents a substantial increase when using a random number of

line segments. Overall, the data support our conjecture that task performance is aided when using fewer line segments, particularly with regard to time taken. *In conclusion, our statistical analysis allows us to accept hypothesis 6: using a minimal number of line segments significantly improves task performance.* Thus, the next five studies were all performed using diagrams drawn with a minimal number of line segments.

6. COLOUR

Hypothesis H2 conjectured that using black lines significantly decreases task performance. Three linear diagrams used to study the effect of colour treatment can be seen in Figs 20, 21, and 22. Participants were asked to 'Tick the check boxes where **none** of the people are also interested in **both** Stars **and** Travel'. The correct solution was to tick the check boxes for College, Internet, Music, and Web.

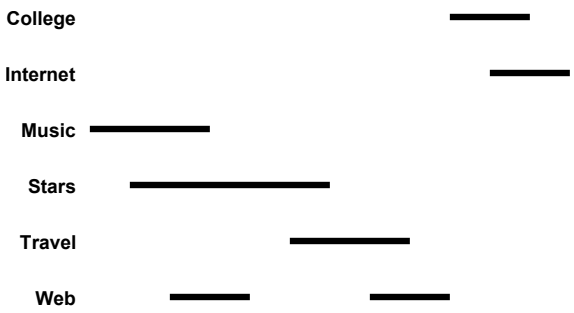

Fig. 20. Black lines.

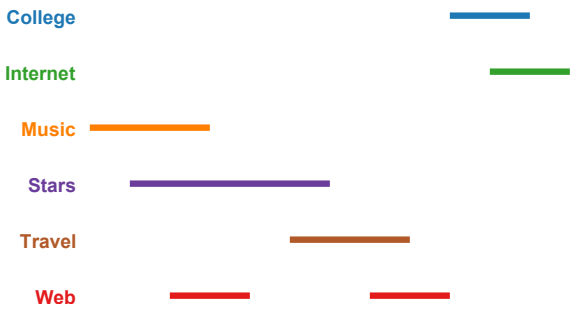

Fig. 21. Coloured lines.

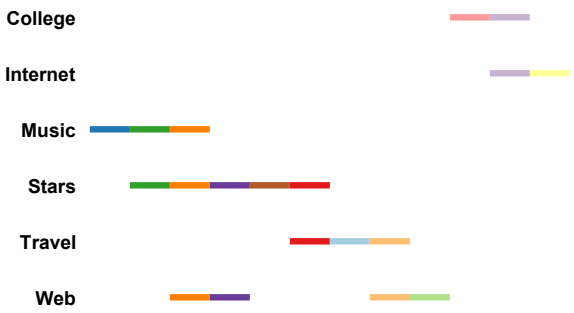

Fig. 22. Coloured overlaps.

Of the 300 participants recruited (100 to each of the three groups – which we call the 'monochrome group', the 'coloured lines group' and the 'coloured overlaps' group – seven were inattentive. The remaining number of participants per group was: monochrome group 99, coloured lines group 95, and coloured overlaps group 99.

6.1. Analysis

The major goal of this second study was to establish whether significant differences exist in task performance when using black lines, coloured lines or coloured overlaps. There were seven inattentive participants, so the following results are based on $300 - 7 = 293$ participants each answering twelve questions, thus the total number of questions answered was $293 \times 12 = 3516$.

6.1.1. Accuracy. The total number of correct responses was 22389 and there were 3981 incorrect responses, giving an overall error rate of 15.10%. Table III shows the breakdown of responses by treatment, with proportionally fewer errors being accrued by the coloured lines group. Approximately one fewer error for every 100 answers was accrued by the coloured lines group over the coloured overlaps group and 1 fewer error versus the monochrome group for every 200 answers. Performing a chi-square test reveals no significant differences between the three choices ($p = 0.129$). Similarly, conducting pairwise chi-square tests also reveals no significant differences, with the smallest p-value being 0.065, between the coloured lines group and the coloured overlaps group. Performing the same analysis with the six colour-blind participants removed does not alter the results. *Therefore, the three colour treatments do not lead to significant differences in task performance in terms of accuracy.*

Table III. Accuracy for Colour: Black Lines vs Coloured Lines vs Coloured Overlaps

Treatment	Correct	Incorrect	Total	Error Rate
Monochrome group	7573	1337	8910	15.00%
Coloured lines group	7299	1251	8550	14.63%
Coloured overlaps group	7517	1393	8910	15.63%
Total	22389	3981	26370	15.10%

6.1.2. Time. The grand mean for the time taken to answer questions in this study was 36.27 seconds per question (standard deviation: 42.01). The average completion time for questions for each group was as follows: monochrome group, 36.88 seconds (SD: 51.94); coloured lines group, 34.58 seconds (SD: 28.03); and coloured overlaps group, 37.27 seconds (SD: 42.08). Overall, on average each question took 2.69 seconds less to complete when using coloured lines as opposed to coloured overlaps (an increase of 7.8%). Performing an ANOVA using $\log_{10}(\text{time})$ data (skewness: 0.33) found no significant effect of visualization technique ($p = 0.829$). Similarly, conducting pairwise comparisons reveals no significant differences between treatments, with the smallest p-value being 0.544. Performing the same analysis with the colour-blind participants removed does not alter the results. *Therefore, the three colour treatments do not lead to significant differences in task performance in terms of time taken.*

6.1.3. Summary of Analysis. Our analysis of the accuracy and time data revealed no significant differences in performance between the three colour treatments. *In conclusion, our statistical analysis does not allow us to reject H2: we have no evidence that using black lines significantly decreases task performance.*

In order to perform the remaining studies, we need to assign one of these colour treatments to linear diagrams. Absent of any other evidence, we picked the coloured lines treatment since this had the lowest point estimate for the error rate in this

experiment. In any case, colour can be used without harming task performance and those who prefer a non-colour option can, in practice, display the diagram using monochrome.

7. GUIDE-LINES

Hypothesis H3 conjectured that using guide-lines significantly improves task performance. We compared using guide-lines with no guide-lines. For this study, the guide-lines were drawn at the beginning and end of each overlap. Moreover, they were drawn underneath the lines used to represent sets. To further differentiate them from the lines representing sets, they were drawn in grey, as this was not used for any of the sets. Two linear diagrams used in this study can be seen in Figs 23 and 24. Participants were asked to 'Tick the check boxes where **all** of the people are also interested in Economics'. The correct solution was to tick the check box for Bands.

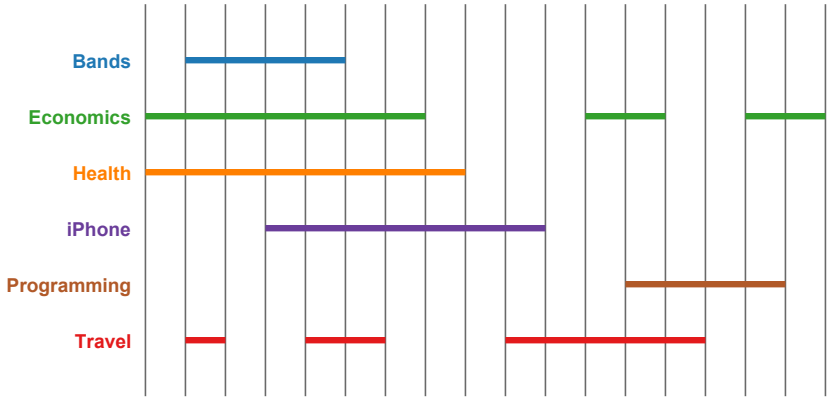

Fig. 23. Guide-lines.

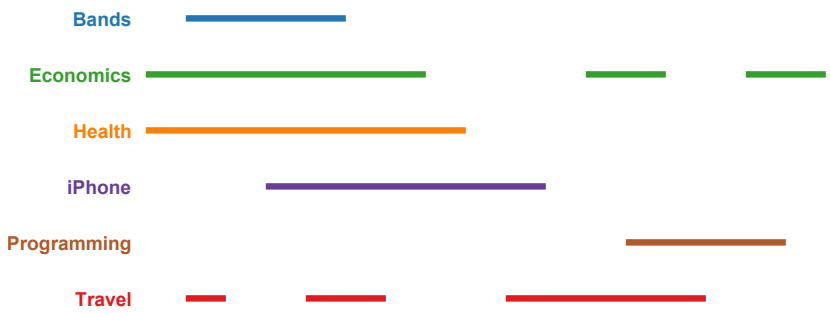

Fig. 24. No guide-lines.

Of the 100 participants recruited to each of the two groups, which we call the 'guide-lines group' and the 'no guide-lines group', there was one inattentive participant in the latter group. The number of participants per group was, thus: guide-lines group 100, and the no guide-lines group 99.

7.1. Analysis

The major goal of this third study was to establish whether task performance is significantly improved using guide-lines. There was one inattentive participant, so the

following results are based on $200 - 1 = 199$ participants each answering twelve questions, thus the total number of questions answered was $199 \times 12 = 2388$.

7.1.1. Accuracy. The total number of correct responses was 14681 and there were 3229 incorrect responses, giving an overall error rate of 18.03%. Table IV shows the breakdown of responses by treatment, with proportionally fewer errors being accrued by the guide-lines group. Almost three fewer errors for every 100 answers were accrued by the guide-lines group over the no guide-lines group. Performing a chi-square test reveals significant differences between the two groups ($p = 0.000$). *Therefore, using guide-lines allows participants to perform significantly more accurately than using no guide-lines.*

Table IV. Accuracy for Guides: Guide-Lines versus No Guide-Lines

Treatment	Correct	Incorrect	Total	Error Rate
Guide-lines group	7500	1500	9000	16.67%
No guide-lines group	7181	1729	8910	19.41%
Total	14681	3229	17910	18.03%

7.1.2. Time. The grand mean for the time taken to answer questions in this study was 36.90 seconds per question (standard deviation: 33.345). The average completion time for questions for each group was as follows: guide-lines group, 36.90 seconds (SD: 30.39); and no guide-lines group, 37.52 seconds (SD: 36.09). Overall, on average each question took 0.62 seconds less to complete when using guide-lines (a decrease of under 2%). The ANOVA performed using $\log_{10}(\text{time})$ data (skewness: 0.38) revealed no significant effect of treatment ($p = 0.713$). *Therefore, using guide-lines leads to no significant differences in performance with respect to time taken.*

7.1.3. Summary of Analysis. Our analysis of the accuracy data revealed that participants perform significantly better when using guide-lines. No significant differences were observed in the time data. Overall, the data support our conjecture that activity (ii) is aided when using guide-lines. *In conclusion, our statistical analysis allows us to accept H3: linear diagrams should be drawn with guides-lines assigned to the ends of the overlaps.* Thus, the next three studies were all performed using diagrams drawn with guide-lines.

8. SET-ORDER

Our next study had the goal of determining whether set-order has a significant impact on task performance. This study compares alphabetic ordering against an adjacency-driven set-order, which places two sets near to each other when their line segments have collinear end-points. H5 conjectures that the adjacency-driven set-order outperforms alphabetic order. An example can be seen in Figs 25 and 26. Participants were asked to 'Tick the check boxes where **all** of the people are also interested in **either** Hifi **or** Journalism' with the correct answer being to tick 'none of the above'.

Of the 100 participants recruited to each of the two groups, which we call the 'alphabetic group' and the 'adjacency-driven group', there were seven inattentive participants in the latter group. The number of participants per group is, thus: alphabetic group 100, adjacency-driven group 93.

8.1. Analysis

This fourth study's goal was to establish whether significant differences exist in task performance when altering the order of the sets. There were seven inattentive participants, so the following results are based on $200 - 7 = 193$ participants each answering 12 questions, thus the total number of questions answered was 2316.

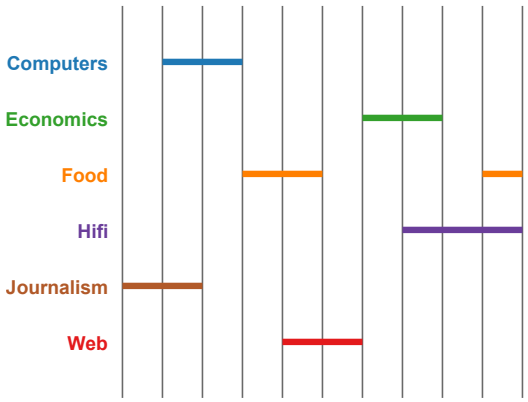

Fig. 25. Alphabetic set-order.

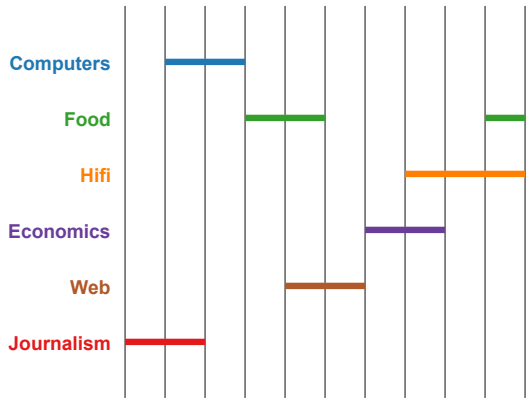

Fig. 26. Adjacency-driven set-order.

8.1.1. Accuracy. In this study into the effect of set-order, the total number of correct responses was 15490 and there were 1880 incorrect responses, giving an overall error rate of 10.8%. Table V shows the breakdown of responses by treatment, with proportionally fewer errors being accrued by the adjacency-driven group, although the difference is very marginal. Performing a chi-square test reveals no significant differences between the two groups ($p = 0.965$). *Therefore, the two set-order treatments do not lead to significant differences in task performance in terms of accuracy.*

Table V. Accuracy for Set-Order: Alphabetic versus Adjacency-Driven

Treatment	Correct	Incorrect	Total	Error Rate
Alphabetic group	8025	975	9000	10.83%
Adjacency-driven group	7465	905	8370	10.81%
Total	15490	1880	17370	10.82%

8.1.2. Time. The grand mean for the time taken to answer questions in this study was 33.41 seconds per question (standard deviation: 37.56). The average completion time for questions for each group was as follows: alphabetic group, 33.87 seconds (SD: 37.29); and adjacency-driven group, 32.91 seconds (SD: 37.84). Overall, on average each question took 0.96 seconds less to complete when adjacency-driven set-order (a decrease of under 3.5%). To establish whether significant differences existed between the two treatments we conducted an ANOVA. The analysis is performed using $\log_{10}(\text{time})$ data (skewness: 0.52). We found no significant effect of treatment ($p = 0.391$). *Therefore, there is no significant difference between the two set-order treatments in terms of time taken to perform tasks.*

8.1.3. Summary of Analysis. Our analysis of both the accuracy and time data revealed that no significant differences in performance exist between the two set-orders. *To conclude, we cannot reject H5: set-order does not significantly impact on task performance.* In order to perform the remaining studies, we need to assign one of these set-order treatments to linear diagrams. In the absence of any other evidence, we opt for using adjacency-driven set-order in the next two studies, since this treatment had the lowest point estimates for the error rate and mean time.

9. LINE WIDTH

Our next study had the goal of determining whether lines or bars should be used to represent the sets. H1 conjectured that bars outperform lines. An example can be seen

in Figs 27 and 28. Participants were asked to 'Tick the check boxes where **some** of the people are also interested in Web' with the correct answer being to tick Design, Economics, Hifi, Internet, and Travel.

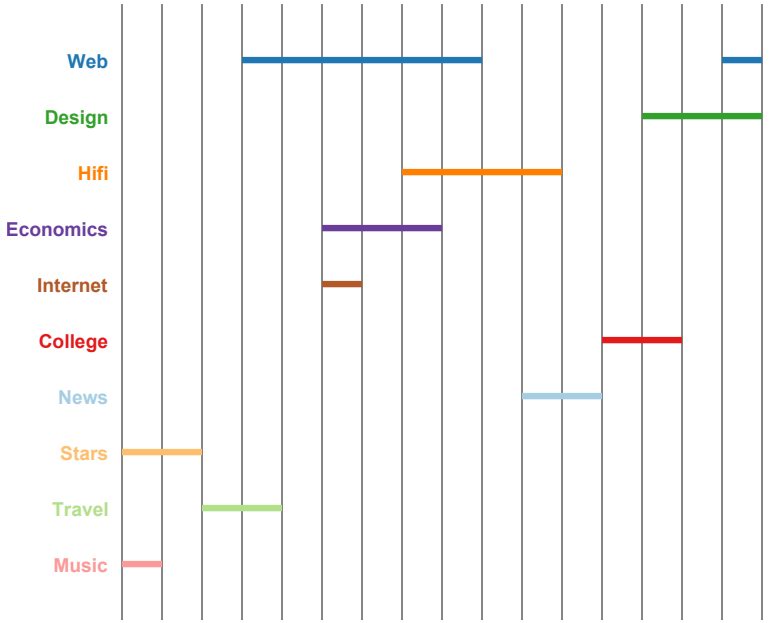

Fig. 27. Lines.

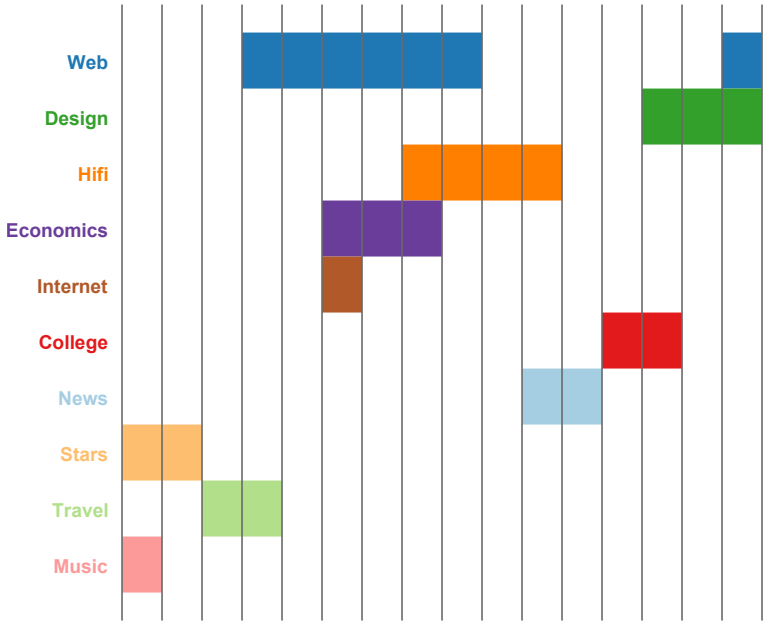

Fig. 28. Bars.

In the pilot study, two questions were identified as too easy (with error rates of 0% and 3% and mean times both under 20 seconds). For the corresponding diagrams, new questions were generated to increase the difficulty. One of the questions was too hard (error rate 42.8% and mean time 56.17 seconds). On inspection, it was felt that this was due to the diagram and it was not possible to create a simpler question of the correct task type. As a result a new diagram was generated along with a new question. In all three cases we noted that the ease or difficulty of the question did not appear to alter across the two treatments.

Of the 100 participants recruited to each of the two groups, which we call the 'bars group' and the 'lines group', there were eight inattentive participants (four in each group). The number of participants per group was, thus: bars group 96, and lines group 96.

9.1. Analysis

This fifth study aimed to establish whether significant differences exist in task performance when altering line width. There were eight inattentive participants, so the following results are based on $200 - 8 = 192$ participants each answering 12 questions, thus the total number of questions answered was 2304.

9.1.1. Accuracy. In this study into the effect of line width, the total number of correct responses was 15380 and there were 1900 incorrect responses, giving an overall error rate of 10.99%. Table VI shows the breakdown of responses by graphial choice, with proportionally fewer errors being accrued by the lines group. Approximately 1.5 more errors were made out of every 100 answers by the bars group. Performing a chi-square test reveals significant differences between the two groups ($p = 0.003$). *Therefore, using lines significantly reduces the error rate as compared to using bars.*

Table VI. Accuracy for Set-Order: Lines versus Bars

Treatment	Correct	Incorrect	Total	Error Rate
Bars group	7628	1012	8640	11.71%
Lines group	7752	888	8640	10.26%
Total	15380	1900	17280	10.99%

9.1.2. Time. The grand mean for the time taken to answer questions in this study was 35.76 seconds per question (standard deviation: 26.66). The average completion time for questions for each group was as follows: lines group, 35.94 seconds (SD: 26.21); and bars group, 35.59 seconds (SD: 27.12). The ANOVA performed using $\log_{10}(\text{time})$ data (skewness 0.03) revealed no significant effect of treatment ($p = 0.584$). Therefore, there is no significant difference between the two treatments in terms of time taken to perform tasks.

9.1.3. Summary of Analysis. Our analysis of the accuracy data revealed that using lines is preferable to using bars. *We cannot accept H1, which conjectured that bars were preferable to lines and instead deduce that lines are significantly better than bars.* This suggests that white space, and therefore less visual clutter, is more helpful in terms of accuracy when performing tasks than using touching bars. However, in this case the benefits of using lines over bars is relatively small. In terms of time performance, there was no significant difference between the bars group and the lines group. Our next study uses lines instead of bars due to the significantly lower error rate.

10. ORIENTATION

This study had the goal of determining whether diagram orientation had a significant impact on task performance, comparing vertical and horizontal orientations. H4 conjectured that there would be no significant difference in performance[4]. An example can be seen in Figs 29 and 30. Participants were asked to 'Tick the check boxes where **some** of the people are also interested in **either** Relaxation **OR** Stars' with the correct solution being to tick Food and Web.

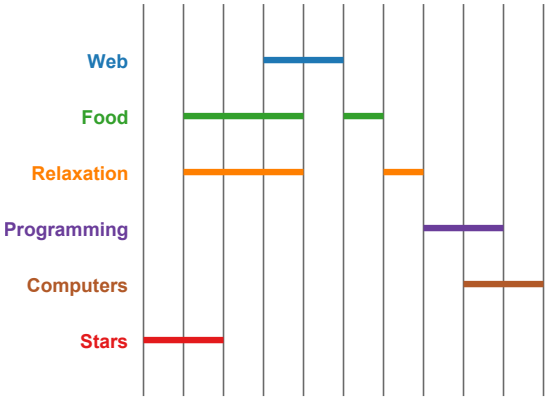

Fig. 29. Horizontal orientation.

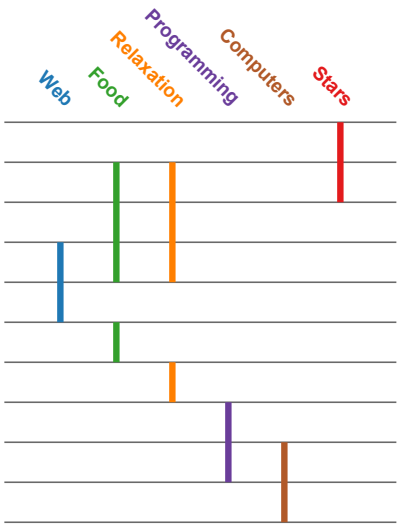

Fig. 30. Vertical orientation.

In the pilot study, the two simple intersection questions were noted as being easy (error rates of around 1%, mean times of 21.07 seconds and 17.65 seconds). There was no obvious reason why one of these questions (with the mean time of 21.07 seconds) was easy so it was not changed. For the other question, participants were asked to identify sets that intersected with Health. The only correct answer was a line drawn immediately next to Health, which we felt made the question easy. The question was changed to make it more challenging.

One question was noted as being particularly hard (error rate of over 50%). This question asked participants to identify subsets of Games. The line for Games comprised four segments, and it was felt that this contributed to the difficulty of the question. An alternative set was chosen for the question. Finally, there was one question which nobody got completely correct, so this was also deemed difficult. In this case, participants were asked to identify the sets that were disjoint with iPhone. In this case, all of the sets were disjoint with iPhone, so all check boxes except 'none of the above' should have been selected. This was the potential cause of difficulty, so the question was changed to reduce the difficulty.

Of the 100 participants recruited to each of the two groups, which we call the 'horizontal group' and the 'vertical group', there were four inattentive participants. The

[4]We acknowledged earlier that the order in which people naturally read (e.g. left to right) may impact which orientation is most effective. We have used participants from the US, which would imply that most participants are speakers of languages written left to right, even if that is not their native language. In any case, the study was performed in English.

remaining number of participants per group was: horizontal group 99, and vertical group 97.

10.1. Analysis

This sixth study aimed to establish whether significant differences exist in task performance when altering diagram orientation. There were four inattentive participants, so the following results are based on $200 - 4 = 196$ participants each answering 12 questions, thus the total number of questions answered was 2352.

10.1.1. Accuracy. In this study into the effect of orientation, the total number of correct responses was 15331 and there were 2309 incorrect responses, giving an overall error rate of 13.09%. Table VII shows the breakdown of responses by treatment, with proportionally fewer errors being accrued by the horizontal group. Performing a chi-square test revealed no significant differences between the two groups ($p = 0.223$). *Therefore, the two orientations do not lead to significant differences in task performance in terms of accuracy.*

Table VII. Accuracy for Orientation: Horizontal versus Vertical

Treatment	Correct	Incorrect	Total	Error Rate
Horizontal group	7771	1139	8910	12.78%
Vertical group	7560	1170	8730	13.40%
Total	15331	2309	17883	13.09%

10.1.2. Time. The grand mean for the time taken to answer questions in this study was 35.56 seconds per question (standard deviation: 37.14). The average completion time for questions for each group was as follows: horizontal group, 34.71 seconds (SD: 20.56); and vertical group, 36.44 seconds (SD: 42.18). Overall, on average each question took 1.73 seconds less to complete with the horizontal orientation (a decrease of 4.75%). Conducting an ANOVA using $\log_{10}(\text{time})$ data (skewness 0.31) revealed no significant effect of treatment ($p = 0.162$). *Therefore, there is no significant difference between the two treatments in terms of time taken to perform tasks.*

10.1.3. Summary of Analysis. Our analysis revealed no significant performance differences between horizontal and vertical orientation. *Thus, we cannot reject H4: orientation does not impact on task performance.* As we need to choose one of the two treatments for our final study, in the absence of any other evidence, we pick the horizontal treatment because it had the lowest point estimates of the error rate and mean time.

11. SUMMARY OF INDIVIDUAL STUDIES

Having performed six empirical studies focusing on individual graphical features, the results are summarized as a set of design principles in table VIII, presented in the order in which the studies were undertaken. As such, we are now in a position to answer the six question posed in section 2:

— *Question 1* asked whether we should use bars or lines. In our study, lines significantly outperform bars in terms of accuracy. Therefore, the answer to question 1 is that we should draw linear diagrams with thin lines.
— *Question 2* asked how we should treat linear diagrams with colour. No significant effect was found in terms of accuracy or time performance. Therefore, the answer to question 2 is that any of the considered colour treatments is suitable.
— *Question 3* asked whether guide-lines would assist task performance. Using guide-lines led to significantly improved accuracy. Therefore, the answer to question 3 is that we should use guide-lines to indicate the start and end points of overlaps.

— *Question 4* focused on diagram orientation. Our study found there to be no significant difference between horizontal and vertical orientation. Therefore, the answer to question 4 is that either a horizontal or vertical orientation can be chosen without impacting task performance.
— *Question 5* was concerned with set-order. Again, we found no significant difference between alphabetic set-order and adjacency-driven set-order. Therefore, the answer to question 5 is that either an alphabetic or an adjacency-driven set-order can be chosen without impacting task performance.
— *Question 6* focused on the number of line segments used. We found that using a minimal number of line segments significantly improved accuracy and completion time. Therefore, the answer to question 6 is that we should minimize the number of line segments.

Table VIII. Summary of Individual Studies

Graphical Prop.	**Lowest Error Rate**	**Lowest Mean Time**	**Design Principle**
line segments	minimal	minimal	use a minimal number of line segments
colour	no sig. diff.	no sig. diff.	–
guide-lines	guide-lines	no sig. diff.	use perpendicular guide-lines
set-order	no sig. diff.	no sig. diff.	–
line width	lines	no sig. diff.	use thin lines
orientation	no sig. diff.	no sig. diff.	–

We now present a final study aimed at determining whether applying the design principles identified through our first six studies leads to significantly improved performance over other combinations of graphical choices. In particular we now compare the following three combinations, that we call 'guided', 'original' and 'anti-guided':

(1) *Guided*: these graphical choices correspond with those found either significantly better than others or those with the lowest error rates and completion times. These choices are:
   (a) line segments: minimal,
   (b) colour: lines,
   (c) guide-lines: yes,
   (d) set-order: adjacency-driven,
   (e) width: lines,
   (f) orientation: horizontal.
(2) *Original*: these graphical choices correspond with those that were made in a previous empirical study comparing linear diagrams to Euler diagrams [Chapman et al. 2014]. These choices are:
   (a) line segments: minimal,
   (b) colour: lines,
   (c) guide-lines: no,
   (d) set-order: alphabetic,
   (e) width: lines,
   (f) orientation: horizontal.

   Here, there are only two graphical choices that are different from 'guided', namely the use of guide-lines (which we found to significantly improve performance) and set-order.
(3) *Anti-guided*: these graphical choices correspond with those found either significantly worse than others or those with the highest error rates and completion times. These choices are:
   (a) line segments: random,

(b) colour: overlaps,
(c) guide-lines: no,
(d) set-order: alphabetic,
(e) width: bars,
(f) orientation: vertical.
By design, all graphical choices differ from 'guided'. Four choices are different from 'original'.

The next section provides the details of this final study, which used the same approach as the first six studies. Thus, we only include the same details as seen in sections 6 to 10.

12. OVERALL STUDY

Our next study had the goal of determining whether applying the guidance derived from the first six studies aids comprehension, using guided, original and anti-guided layouts. An example can be seen in Figs 31, 32 and 33. Participants were asked to 'Tick the check boxes where **some** of the people are also interested in Books ' with the correct solution being to tick Android, Cars, Media, News, and Stars.

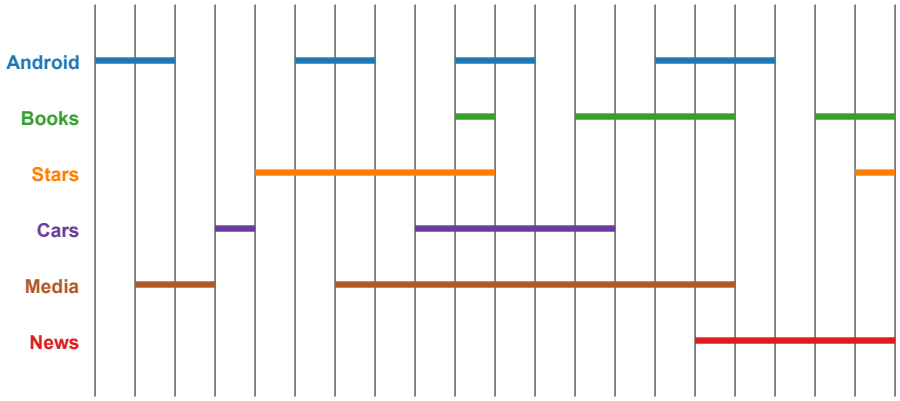

Fig. 31. Guided.

Fig. 32. Original.

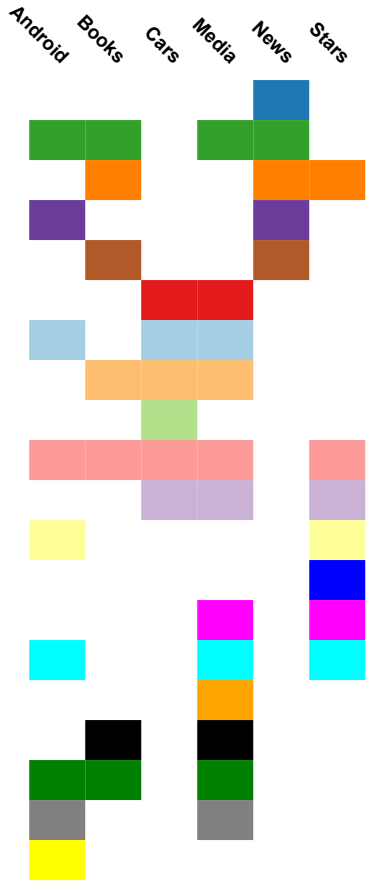

Fig. 33. Anti-guided.

Of the 100 participants recruited to each of the three groups, which we call the 'guided group', the 'original group' and the 'anti-guided group', there were 13 inattentive participants. The remaining number of participants per group was: guided group 95, original 100 group, and anti-guided group 92.

12.1. Analysis

This final study aimed to establish whether significant differences exist in task performance when applying different combinations of the guides. There were 13 inattentive participants, so the following results are based on $300 - 13 = 287$ participants each answering 12 questions, thus the total number of questions answered was 3444.

12.1.1. Accuracy. The total number of correct responses was 21920 and there were 3910 incorrect responses, giving an overall error rate of 15.14%. Table IX shows the breakdown of responses by treatment, with proportionally fewer errors being accrued by the guided group. Performing a chi-square test revealed significant differences between the three groups ($p = 0.000$). To distinguish the three groups, pairwise chi-square tests were performed, establishing the following significant differences: guided versus anti-guided ($p = 0.000$), original versus anti-guided ($p = 0.000$). Therefore both guided and original layouts significantly improve accuracy over using anti-guided layouts. How-

ever, there was no significant difference in accuracy between the guided and original layouts ($p = 0.100$). Therefore, both the guided and original layouts lead to significantly improved accuracy over using the anti-guided layout.

Table IX. Accuracy for Final Study: Guided versus Original versus Not Guided

Treatment	Correct	Incorrect	Total	Error Rate
Guided group	7544	1006	8550	11.77%
Original group	7868	1132	9000	12.58%
Anti-guided group	6508	1772	8280	21.40%
Total	21920	3910	25830	15.14%

12.1.2. Time. The grand mean for the time taken to answer questions in this study was 41.34 seconds per question (standard deviation: 36.84)[5]. The average completion time for questions for each group was as follows: guided group, 38.34 seconds (SD: 38.63); original group, 37.52 seconds (SD: 29.48), and anti-guided group, 48.59 seconds (SD: 41.47). Conducting an ANOVA using $\log_{10}(\text{time})$ data (skewness 0.20) revealed significant effect of treatment ($p = 0.001$). Therefore, there is a significant difference between the three treatments in terms of time taken to perform tasks. Performing a Tukey test yields the following ranking of treatments: guided and original are both significantly faster than anti-guided. However, the mean times for guided and original are not significantly different.

12.1.3. Summary of Analysis. Our analysis of both the accuracy and time data established that adhering to all of our design principles is significantly better than adhering to none of them. Moreover, using the original set of guides also significantly aids performance as compared to the anti-guided treatment. We conclude that the graphical features of linear diagrams significantly impact on task performance.

13. THREATS TO VALIDITY

As we have already acknowledged, the order in which the first six studies were conducted could impact on the results. Different orders may give a different set of design principles. For example, in Fig. 28 it could be argued that the use of guide-lines on top of bars may be considered problematic as they appear to divide up the bars. By contrast, this effect is not noticeable on the lines (see Fig. 27). As a result, evaluating line thickness before guide-lines could have yielded different design principles.

The following discusses further threats to validity, focusing primarily on those arising from using a crowd-sourcing approach, that were considered and addressed to ensure the study is robust and fit-for-purpose. The following factors were considered in our study design:

Laboratory: ideally, all participants would be exposed to the same environment when the study took place. This would ensure that each participant was exposed to the same hardware, in an environment that was free from noise and interruption. By adopting a crowd-sourcing approach, we had no control over the environment in which each participant took part. To reduce the effect of this compromise, a large number of participants were recruited to each treatment group. In total, our seven studies involved 1760 participants.

[5]We observe that this mean is somewhat higher than the grand means of the six earlier studies. However, this is likely to be due to the inclusion of the anti-guided treatment, which has a relatively high mean time of 48.59 seconds.

Time: to ensure the rigour of time measurements, consideration was paid to the precise duration elapsed interpreting a diagram as well as the units employed to measure time. As we used a crowd-sourcing approach, there was little control over any distractions impacting the time taken by each participant on each question. To manage this, a large sample size was used.

Question: it was considered a threat if participants did not spend time reading and understanding the questions and diagrams. To manage this threat, non-trivial real-world data were used to generate the diagrams, resulting in diagrams that participants had to read and understand before being able to answer the posed question. It was also considered a threat if the diagrams were regarded as trivial; having only a few sets was deemed insufficient to yield noticeable differences in response times, should they exist. To manage this, diagrams represented six or ten sets in order to demand cognitive effort. In addition, our questions required participants to look at all of the sets visualized in the diagrams. This was to ensure that differences in diagrams arising from the graphical properties being varied would be reflected in the task being undertaken. Lastly, the study included questions to allow inattentive participants to be identified, catching those who did not read questions carefully.

Participant: participants were representative of the wider population as indicated by the demographic information given (roughly even splits between male and female, large age ranges and diverse qualification levels). They were all from the USA and had a high HIT approval rate.

Thus, the results should be taken to be valid within these constraints.

14. AUTOMATED LAYOUT

To enable linear diagrams to be used effectively, we have implemented a tool that supports their automated layout. The top lefthand panel in the screen shot, Fig. 34, shows how users describe the diagram they require. Each line in the text entry box contains a description of the required overlap, so the first entry indicates that there should be an overlap containing just the set Android. Likewise, the second entry asserts that an overlap should contain Android and Games, and so forth. The selected options determine the graphical features to be possessed by the diagram. By default, the software is set to draw diagrams following the design principles identified in this paper.

Automatically generating linear diagrams is relatively straightforward as the diagrams simply consist of labels and straight line segments. Tailoring their layout, when choosing graphical properties that correspond to Bertin's retinal variables (which correspond to four of the graphical properties), is unproblematic: only minor layout changes are required. This includes switching between bars and lines, the various colour variants, adding guide-lines, and changing orientation.

However, more detailed algorithmic processes are required when choosing an adjacency-driven set-order and for reducing line breaks. In our implementation, the corresponding algorithms are approximate, as optimal solutions which minimize the distance between sets (for choosing a set-order) and minimize the number of line breaks (thus ordering the overlaps) are required to consider all possible orderings. Thus, such minimizations cannot be achieved with a polynomial algorithm.

With regard to adjacency-driven set-order, a greedy method has been implemented. The task is to order the sets so that pairs of sets which have more collinear end-points are neighbours. First, given the required set of overlaps in some order, our algorithm draws the line segments for a set with the greatest number of segments; such a set is represented by line segments with the most end-points. This choice is motivated by the

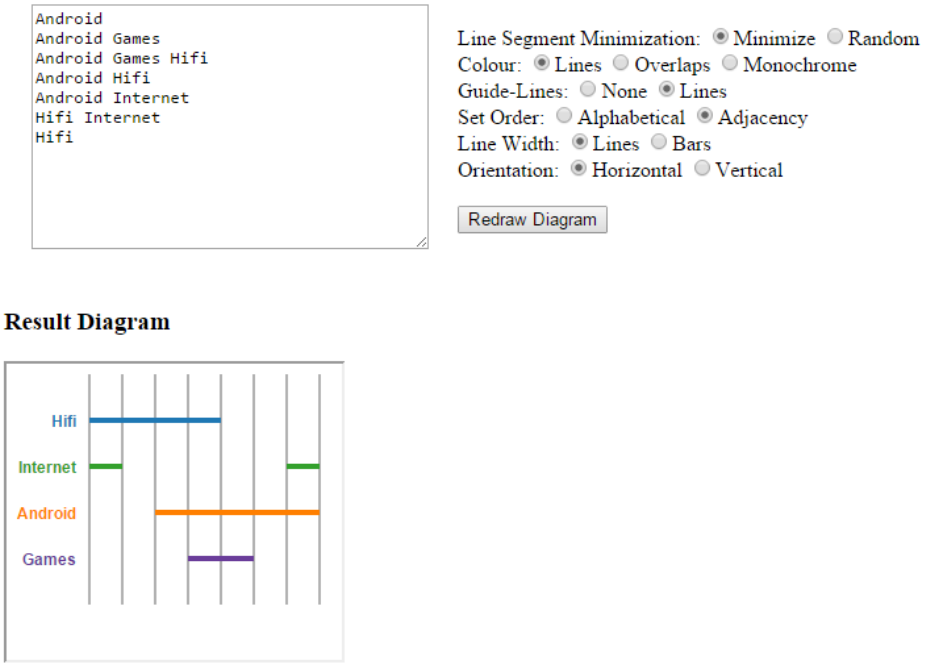

Fig. 34. A screenshot of our software.

likelihood that, because this set has the most end-points, there are going to be many other sets whose line segments' end-points are collinear with those for the drawn set. Sets are added to the diagram iteratively, with a next set to be drawn being that with the most collinear end-points with either the first-ordered or last-ordered set that has already been drawn. The to-be-drawn set is then added next to the set against which it scored highest. When there is a choice of sets to be drawn next, the tie is broken by choosing the set whose name comes first alphabetically.

A greedy method is also applied for line break minimization. Here we are choosing an order for the overlaps. We use a measure of closeness between two overlaps to reduce the number of line breaks. First, we observe that, for a pair of overlaps O_1 and O_2, the number of sets that they have in common is the same as the number of lines that *are not broken* if they are drawn next to each other. Likewise, the number of sets that are in *exactly one* of O_1 and O_2 corresponds to the number of lines that are broken when they are drawn adjacent to each other (in other words, this is the number of collinear end-points that lie 'between' O_1 and O_2). Our measure of distance between a pair of overlaps calculates the number of sets in common and subtracts the number of sets that appear in exactly one of the two overlaps.

The greedy algorithm then attempts to maximise this measure. First, the algorithm draws the overlap containing the largest number of sets. Overlaps are added to the diagram iteratively. The next overlap, O_{next}, to be drawn is one with the largest value of the measure with either the left-most, O_l, or right-most, O_r, drawn overlap, of the currently drawn diagram; so, the distance between O_{next} and O_l or O_r is the largest of all still-to-be-drawn overlaps. When there is a choice of overlap to be drawn next, the tie is broken by choosing that which was first entered into the linear generator by the user.

The linear diagram generator is implemented in Javascript, with the diagram output in svg format. It is available on the web: `www.eulerdiagrams.com/linear`.

15. CONCLUSION

Using perceptual theories relating to graphical choices, we identified graphical features, such as colour and size, that could impact users' ability to perform set-theoretic tasks using linear diagrams. These insights led to a number of hypotheses about the effect of these graphical features which we tested through a series of seven controlled empirical studies. The three statistically significant results from the first six studies are summarized as follows, each of which gives rise to three *design principles:*

— **Line Segments** Using a minimal number of line segments allows participants to perform significantly more accurately and significantly faster than using a random number of line segments.
Design Principle 1: draw linear diagrams with a minimal number of line segments.
— **Guide-Lines** Using guide-lines at the beginning and end of each overlap allowed participants to perform faster and significantly more accurately.
Design Principle 2: draw linear diagrams with guide-lines at the beginning and end of each overlap.
— **Line Width** Altering line width led to significant performance differences in accuracy: thin lines gave rise to significantly fewer errors. No significant time differences were found.
Design Principle 3: draw linear diagrams with thin lines.

To tie these six individual studies together, we empirically compared various combinations of the principles: meeting all principles versus meeting most of the principles versus meeting none of the principles. We found that performance was significantly improved when most or all of the principles were met. This seventh study thus supports the adoption of the principles when using linear diagrams to perform set-theoretic tasks. Interestingly, the final study did not distinguish the guided treatment from the original treatment. We conjecture that there may be some interaction between the design principles and other graphical features. This effect is worthy of further investigation, to tease out the impact of each graphical property on the others. Indeed, whilst we have covered a major part of the design space for linear diagrams, more focussed studies could be conducted to further fine-tune their design.

To enable people to take advantage of the benefits of using linear diagrams to perform such tasks, we have provided a freely available software implementation that allows any combination of the design principles to be enforced; the software is available from `www.eulerdiagrams.com/linear`. It would be useful, in the future, to implement improved algorithms that more closely approximate the minimal number of line segments and desired set-order, whilst taking care to ensure that the run-time performance is acceptable.

We note that our empirical results are valid within the constraints imposed by the study, just as with any empirical study. In our case, the results are valid for the task types undertaken by our pool of participants drawn from the general population. Future work might examine whether these results hold when the participants are from a pool of technically trained people or with different levels of education. Differences in gender and age on task performance could also be examined. In addition, it would be possible to further explore options such as the choice of colours used and the precise shade of grey used for the guide-lines. Moreover, the result on line thickness could be linked to the presence of more white space in the diagrams, an aspect that may be worthy of further investigation.

A major avenue of future work is to extend linear diagrams to include the representation of data items and network relations, which can be represented by node-link diagrams. Complex data such as this arises in many areas, perhaps most prominently in social networks. We believe that linear diagrams augmented with node-link diagrams are likely to bring significant usability benefits over Euler-diagrams augmented with node-link diagrams, such as [Collins et al. 2009; Mutton et al. 2004; Riche and Dwyer 2010; Simonetto 2012]. It would be interesting to explore how to draw linear diagrams and node-link diagrams in combination, aiming to optimize, with respect to task performance, the layout of both notations.

Exploring interaction with linear diagrams, and extensions of them, is also a valuable avenue for further research. Examples for interactive features include zooming, filtering, highlighting and providing details on demand, which is congruent with Shneiderman's visual information-seeking mantra [Shneiderman 1996]. The ability of visualization techniques to support these features will be important for their wide-scale applicability, allowing one to exploit the value of interaction and enable new analysis possibilities.

Acknowledgements We would like to thank the participants who took part in the seven studies reported in this paper. Thanks are due to to John Howse and Jim Burton for their comments on drafts of this paper. We also thank the anonymous reviewers and the editor for their detailed and constructive comments which have greatly improved the paper.

REFERENCES

Y. Y. Ahn, J. Bagrow, and S. Lehmann. 2010. Link communities reveal multiscale complexity in networks. *Nature* 466 (2010), 761–764.

B. Alper, N. Henry Riche, G. Ramos, and M. Czerwinski. 2011. Design study of LineSets, a novel set visualization technique. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2259–2267.

B. Alsallakh, W. Aigner, S. Miksch, and H. Hauser. 2013. Radial Sets: Interactive Visual Analysis of Large Overlapping Sets. *Visualization and Computer Graphics, IEEE Transactions on* 19, 12 (2013), 2496–2505.

B. Alsallakh, L. Micallef, W. Aigner, H. Hauser, S. Miksch, and P. Rodgers. 2014. Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges. In *Eurographics Conference on Visualization (EuroVis)*. 124–138.

J. Bertin. 1983. *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press.

A. Blake, G. Stapleton, P. Rodgers, L. Cheek, and J. Howse. 2014b. The Impact of Shape on the Perception of Euler Diagrams. In *8th International Conference on the Theory and Application of Diagrams*. Springer, 124–138.

A. Blake, G. Stapleton, P. Rodgers, and J. Howse. 2014a. How Should We Use Colour in Euler Diagrams. In *7th International Symposium on Visual Information Communication and Interaction*. ACM.

S.K Card, J.D Mackinlay, and B. Shneiderman. 1999. *Readings in Information Visualisation: Using Vision to Think*. Academic Press.

P. Chapman, G. Stapleton, P. Rodgers, L. Michallef, and A. Blake. 2014. Visualizing Sets: An Empirical Comparison of Diagram Types. In *Diagrams 2014*. Springer, 146–160.

J. Chen, N. Menezes, A. Bradley, and T. North. 2011. Opportunities for Crowdsourcing Research on Amazon Mechanical Turk. *Human Factors* 5, 3 (2011).

C. Collins, G. Penn, and M. Sheelagh T. Carpendale. 2009. Bubble Sets: Revealing Set Relations with Isocontours over Existing Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1009–1016.

L. Couturat. 1903. *Opuscules et fragments inédits de Leibniz*. Felix Alcan.

K. Dinkla, M. El-Kebir, C.-I. Bucur, M. Siderius, M. Smit, M. Westenberg, and G. Klau. 2014. eXamine: Exploring annotated modules in networks. *BMC bioinformatics* 15, 1 (2014), 201.

M. Dörk, N.H. Riche, G. Ramos, and S. Dumais. 2012. PivotPaths: Strolling Through Faceted Information Spaces. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2709–2718.

J. Feldman. 2007. Formation of visual “objects” in the early computation of spatial relations. *Perception and Psychophysics* 69, 5 (2007), 816–827.

J. Flower and J. Howse. 2002. Generating Euler Diagrams. In *2nd International Conference on the Theory and Application of Diagrams*. Springer, Georgia, USA, 61–75.

W. Freiler, K. Matković, and H. Hauser. 2008. Interactive Visual Analysis of Set-Typed Data. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1340–1347.

B. Gottfried. 2015. A Comparative Study of Linear and Region Based Diagrams. *Journal of Spatial Information Science* 2015, 10 (2015), 3–20.

J. Heer and M. Bostock. 2010. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *28th SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 203–212.

H. Hofmann, A. Siebes, and A. Wilhelm. 2000. Visualizing Association Rules with Interactive Mosaic Plots. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 227–235.

J. Huo. 2008. KMVQL: a Visual Query Interface Based on Karnaugh Map. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI2008)*. ACM, 243–250.

K. Koffka. 1935. *Principles of Gestalt Pschology*. Lund Humphries.

S. Leborg. 2006. *Visual Grammar*. Princeton Architectural Press.

J. Leskovec. 2011. Stanford large network dataset collection. *URL http:/ /snap. stanford. edu/data/index. html* (2011).

A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. 2014. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1983–1992.

R. Mazza. 2009. *Introduction to Information Visualisation*. Springer.

L. Micallef, P. Dragicevic, and J.-D. Fekete. 2012. Assessing the Effect of Visualizations on Bayesian Reasoning through Crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2536–2545.

P. Mutton, P. Rodgers, and J. Flower. 2004. Drawing Graphs in Euler Diagrams. In *3rd International Conference on the Theory and Application of Diagrams*, Vol. 2980. Springer, 66–81.

D. Oppenheimer, T. Meyvis, and N. Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45, 4 (2009), 867–872.

G. Paolacci, J. Chandler, and P. G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010), 411–419.

N. Riche and T. Dwyer. 2010. Untangling Euler diagrams. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1090–1099.

P. Rodgers, , L. Zhang, and H. Purchase. 2012. Wellformedness Properties in Euler Diagrams: Which Should be Used? *IEEE Transactions on Visualization and Computer Graphics* 18, 7 (2012), 1089–1100.

P. Rodgers, L. Zhang, and A. Fish. 2008. General Euler Diagram Generation. In *5th International Conference on the Theory and Application of Diagrams*. Springer, 13–27.

Set Visualiser. accessed March 2014. http://www-edc.eng.cam.ac.uk/tools/set_visualiser/. (accessed March 2014).

Ben Shneiderman. 1996. The Eyes Have it: A Task by Data Type Taxonomy for Information Visualizations. In *IEEE Symposium Visual Languages*. 336–343.

P. Simonetto. 2012. *Visualisation of Overlapping Sets and Clusters with Euler Diagrams*. Ph.D. Dissertation. Université Bordeaux.

P. Simonetto, D. Auber, and D. Archambault. 2009. Fully Automatic Visualisation of Overlapping Sets. *Computer Graphics Forum* 28, 3 (2009), 967–974.

G. Stapleton, P. Rodgers, J. Howse, and L. Zhang. 2009. Inductively Generating Euler Diagrams. *accepted for IEEE Transactions on Visualization and Computer Graphics* (2009).

G. Stapleton, L. Zhang, J. Howse, and P. Rodgers. 2011. Drawing Euler Diagrams with Circles: The Theory of Piercings. *IEEE Transactions on Visualization and Computer Graphics* 17, 7 (2011), 1020–1032.

J. Wagemans, J. Elder, M. Kubovy, S. Palmer, M. Peterson, and M. Singh. 2012. A Century of Gestalt Psychology in Visual Perception: I. Perceptual Grouping and Figure-Ground Organisation. *Psychol Bull.* 138, 6 (2012), 1172–1217.

S. Wasserman and K. Faust. 1994. *Social Network Analysis*. Cambridge Univ. Press.

K. Wittenburg, T. Lanning, M. Heinrichs, and M. Stanton. 2001. Parallel Bargrams for Consumer-based Information Exploration and Choice. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*. ACM, 51–60.

APPENDIX

In what follows, there is one subsection for each empirical study, except for the first study, presenting a table on diagram complexity and information on participant demographics.

A.1. Colour Study

Table X indicates the level of complexity of the linear diagrams used in this study.

Table X. Diagram Complexity for Study 2: Black lines versus Coloured Lines versus Coloured Overlaps

Data Set	1	2	3	4	5	6	7	8	9	10	11	12
Sets	6	6	6	10	10	10	6	6	6	10	10	10
1-set overlaps	6	5	6	3	4	9	2	4	6	3	9	4
2-set overlaps	4	4	2	8	9	5	5	4	5	8	6	10
3-set overlaps	0	0	2	10	0	1	1	2	1	5	2	5
4-set overlaps	0	0	0	3	0	0	0	2	0	0	2	1
5-set overlaps	0	0	0	2	0	0	0	0	0	0	1	3
6-set overlaps	0	0	0	2	0	0	0	0	0	0	0	1
Total number of overlaps	10	9	10	28	13	15	8	12	12	16	20	24
Task Type	I	S	D	I	S	D	CI	CS	CD	CI	CS	CD

The demographics of the participants were as follows:

— **gender:** 128M, 164F, 0 other, 1 not stated
— **age range, in years:** 19 to 70 (mean: 34)
— **qualification level:** 0 not stated, 2 some high school, 16 high school graduate, 84 some college, 29 associates degree, 117 Bachelors degree, 34 Masters degree, 11 doctorate degree.

There were six colour blind participants, two in the monochrome group and four in the coloured overlaps group.

A.2. Guide-Lines

Table XI indicates the level of complexity of the linear diagrams used in this study.

Table XI. Diagram Complexity for Study 3: Perpendicular Guides versus No Guides

Data Set	1	2	3	4	5	6	7	8	9	10	11	12
Sets	6	6	6	10	10	10	6	6	6	10	10	10
1-set overlaps	6	4	5	10	3	9	5	5	6	9	8	4
2-set overlaps	5	6	3	13	8	9	7	6	8	6	7	5
3-set overlaps	0	3	0	0	5	1	2	3	1	0	0	5
4-set overlaps	0	3	0	0	0	0	0	0	0	0	0	1
5-set overlaps	0	1	0	0	0	0	0	1	0	0	0	0
6-set overlaps	0	0	0	0	0	0	0	0	0	0	0	0
$>$ 6-set overlaps	0	0	0	0	0	0	0	0	0	0	0	0
Total number of overlaps	11	17	8	23	16	19	14	14	15	15	15	15
Task Type	I	S	D	I	S	D	CI	CS	CD	CI	CS	CD

The demographics of the participants were as follows:

— **gender:** 95M, 104 F, 0 other, 0 not stated
— **age range, in years:** 18 to 71 (mean: 35)
— **qualification level:** 0 not stated, 1 some high school, 22 high school graduate, 54 some college, 18 associates degree, 73 Bachelors degree, 28 Masters degree, 3 doctorate degree.

A.3. Set-Order

Table XII indicates the level of complexity of the linear diagrams used in this study.

Table XII. Diagram Complexity for Study 4: Alphabetic versus Adjacency-Driven Set-Order

Data Set	1	2	3	4	5	6	7	8	9	10	11	12
Sets	6	6	6	10	10	10	6	6	6	10	10	10
1-set overlaps	5	5	6	9	5	9	3	6	5	5	7	10
2-set overlaps	2	2	2	6	2	9	2	4	2	10	7	12
3-set overlaps	0	0	2	0	4	1	3	0	2	2	2	4
4-set overlaps	0	0	0	0	3	0	1	0	0	2	0	2
5-set overlaps	0	0	0	0	0	0	1	0	0	0	0	0
6-set overlaps	0	0	0	0	0	0	1	0	0	0	0	0
$>$ 6-set overlaps	0	0	0	0	0	0	0	0	0	0	0	0
Total number of overlaps	7	7	10	15	14	19	11	10	9	19	16	28
Task Type	I	S	D	I	S	D	CI	CS	CD	CI	CS	CD

The demographics of the participants were as follows:

— **gender:** 102 M, 89 F, 1 other, 1 not stated
— **age range, in years:** 18 to 69 (mean: 32)
— **qualification level:** 0 not stated, 3 some high school,14 high school graduate, 66 some college, 16 associates degree, 75 Bachelors degree, 16 Masters degree, 3 doctorate degree.

A.4. Line Width

Table XIII indicates the level of complexity of the linear diagrams used in this study.

Table XIII. Diagram Complexity for Study 5: Bars versus Lines

Data Set	1	2	3	4	5	6	7	8	9	10	11	12
Sets	6	6	6	10	10	10	6	6	6	10	10	10
1-set overlaps	6	3	6	7	4	6	5	5	6	9	5	1
2-set overlaps	3	2	6	7	10	7	7	2	6	6	2	4
3-set overlaps	0	3	0	2	4	2	2	0	7	2	2	5
4-set overlaps	0	1	0	0	2	3	0	0	1	2	4	4
5-set overlaps	0	1	0	0	0	1	0	0	0	1	0	3
6-set overlaps	0	1	0	0	0	0	0	0	0	0	0	0
$>$ 6-set overlaps	0	0	0	0	0	0	0	0	0	0	0	$1 \times 7, 1 \times 10$
Total number of overlaps	9	11	12	16	20	20	14	7	20	20	13	19
Task Type	I	S	D	I	S	D	CI	CS	CD	CI	CS	CD

The demographics of the participants were as follows:

— **gender:** 97 M, 94 F, 1 other, 0 not stated
— **age range, in years:** 18 to 71 (mean: 32)
— **qualification level:** not stated, 1 some high school, 9 high school graduate, 65 some college, 20 associates degree, 71 Bachelors degree, 22 Masters degree, 4 doctorate degree.

A.5. Orientation

Table XIV indicates the level of complexity of the linear diagrams used in this study. The demographics of the participants were as follows:

— **gender:** 95 M, 101 F, 0 other, 0 not stated
— **age range, in years:** 18 to 66 (mean: 33)

Table XIV. Diagram Complexity for Study 6: Vertical versus Horizontal Orientation

Data Set	1	2	3	4	5	6	7	8	9	10	11	12
Sets	6	6	6	10	10	10	6	6	6	10	10	10
1-set overlaps	5	4	5	4	7	10	6	4	6	9	4	2
2-set overlaps	7	9	7	9	7	6	2	4	6	9	10	10
3-set overlaps	2	6	6	4	2	1	2	2	7	1	4	3
4-set overlaps	0	1	1	0	0	0	0	2	1	0	2	3
5-set overlaps	0	0	1	0	0	0	0	0	0	0	0	2
Total number of overlaps	14	20	20	13	16	17	10	12	20	19	20	20
Task Type	I	S	D	I	S	D	CI	CS	CD	CI	CS	CD

— **qualification level:** 0 not stated, 2 some high school, 20 high school graduate, 61 some college, 15 associates degree, 70 Bachelors degree, 22 Masters degree, 6 doctorate degree.

A.6. Final Study

Table XV indicates the level of complexity of the linear diagrams used in this study.

Table XV. Diagram Complexity for the Final Study

Data Set	1	2	3	4	5	6	7	8	9	10	11	12
Sets	6	6	6	10	10	10	6	6	6	10	10	10
1-set overlaps	5	2	5	6	9	9	5	6	2	5	7	3
2-set overlaps	7	5	3	7	6	9	7	5	4	2	7	8
3-set overlaps	6	2	0	2	2	1	2	7	4	2	2	5
4-set overlaps	1	0	0	0	2	0	0	1	3	4	0	0
5-set overlaps	1	0	0	0	1	0	0	0	1	0	0	0
Total number of overlaps	20	9	8	15	20	19	14	19	14	13	16	16
Task Type	I	S	D	I	S	D	CI	CS	CD	CI	CS	CD

The demographics of the participants were as follows:

— **gender:** 113 M, 174 F, 0 other, 0 not stated
— **age range, in years:** 18 to 80 (mean: 33)
— **qualification level:** 1 not stated, 4 some high school, 20 high school graduate, 78 some college, 33 associates degree, 87 Bachelors degree, 57 Masters degree, 7 doctorate degree.