Visualizing Sets: An Empirical Comparison of Diagram Types

Peter Chapman[1], Gem Stapleton[1], Peter Rodgers[2], Luana Micallef[2], and Andrew Blake[1]

[1] University of Brighton, UK
{p.b.chapman,g.e.stapleton,a.l.blake}@brighton.ac.uk
[2] University of Kent, UK
{p.j.rodgers,l.micallef}@kent.ac.uk

Abstract. There are a range of diagram types that can be used to visualize sets. However, there is a significant lack of insight into which is the most effective visualization. To address this knowledge gap, this paper empirically evaluates four diagram types: Venn diagrams, Euler diagrams with shading, Euler diagrams without shading, and the less well-known linear diagrams. By collecting performance data (time to complete tasks and error rate), through crowdsourcing, we establish that linear diagrams outperform the other three diagram types in terms of both task completion time and number of errors. Venn diagrams perform worst from both perspectives. Thus, we provide evidence that linear diagrams are the most effective of these four diagram types for representing sets.

Keywords: Set visualization, linear diagrams, Venn diagrams, Euler diagrams

1 Introduction

Sets can be represented in both sentential (textual) and visual forms and the latter is often seen as cognitively beneficial but only if the visual form is effective [1]. To-date, various different visualizations of sets have been proposed, but there is little understanding of their relative effectiveness. This paper addresses this knowledge gap by empirically comparing four visualizations: Venn diagrams, Euler diagrams with shading, Euler diagrams without shading, and linear diagrams. We do not consider the relative effectiveness of these diagrams with traditional sentential notations (such as $(A \cap B) - C = \emptyset$) as it was felt the latter would be too hard for many people to understand in a short time frame.

The Venn and the Euler variants will be familiar to most readers. All three use curves to represent sets: the area inside a curve with label A represents the set A. *Venn diagrams* (upper left, Fig. 1) require that every possible intersection between curves is present. In order assert that sets are empty, the appropriate regions (often called zones) are shaded. *Euler diagrams with shading* (upper right, Fig. 1), by contrast, can either not include or shade zones which represent the empty set. *Euler diagrams without shading* (lower left, Fig. 1) provide a

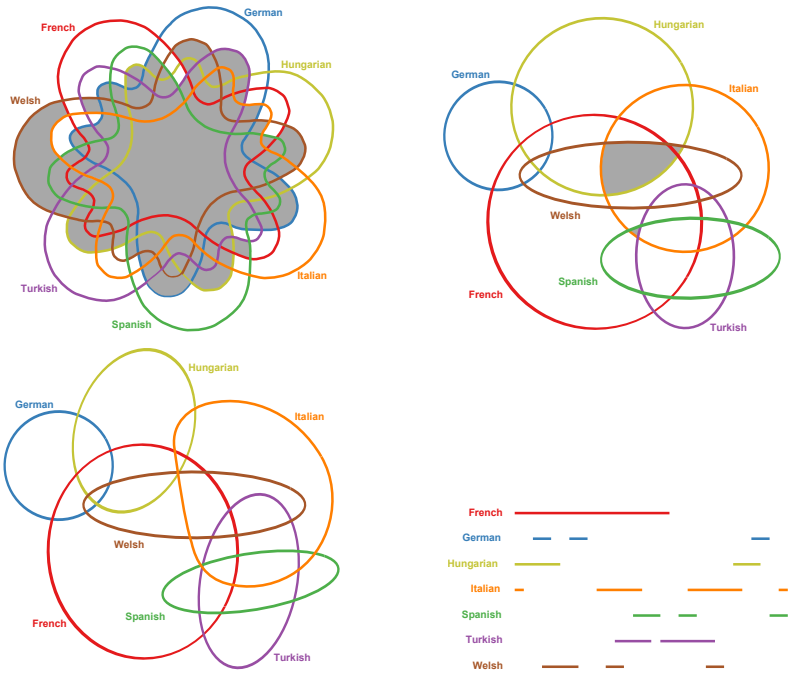

Fig. 1. The four diagram types considered in this paper.

minimal representation of the underlying sets: all, and only, zones that represent non-empty sets are included. Minimality of representation can necessitate the presence of diagrammatic features considered sub-optimal for cognition, such as three curves meeting at a point [2]; in this example, Hungarian, Italian and Welsh form such a 'triple' point. Across these three diagram types, any pair of diagrams expressing the same information will have the same number of unshaded zones; only the number of shaded zones can differ.

Linear diagrams were introduced by Leibniz [3], with parallel bargrams [4] and double decker plots [5] being similar. Each set is represented as one or more horizontal 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 all of the overlaps represent all of the non-empty set intersections. As an example, consider the linear diagram in the lower right panel of Fig. 1. Since there is a region of the diagram where the lines French, Italian, Turkish and Welsh (and only those lines) overlap, the intersection of those four sets, less the union of Spanish, German and Hungarian, is non-empty. Further, it is not the case that Hungarian is a subset of French, because part of the line representing Hungarian does not overlap with that for French: although one segment of the Hungarian line completely overlaps with the French line, the other segment does not. Also, because there is no overlap involving all seven sets, we can infer that no element is in all of the sets.

Most existing research on diagram effectiveness evaluates notations against some cognitive framework for what *should* constitute a good diagram, such as the Physics of Notations [6] or empirically determines which aspects of a *particular* notation are most effective (e.g. [2],[7],[8]). There are some exceptions, such as [9] which shows that the most effective diagram type is task dependent. Following [9], we perform an empirical comparison of different notations. Similar studies do exist, and, by necessity, are task specific. In [10] Euler diagrams, Venn diagrams and linear diagrams were compared in the context of syllogistic reasoning (i.e. the interactions between three sets). In a more general reasoning context, a study between Euler and Venn diagrams was undertaken in [11]. In both studies, Venn diagrams were least effective, and in [10], linear diagrams were as effective as Euler diagrams. Our study is the first to assess the effectiveness of the four diagram types for visualizing sets and the first of its kind to be conducted on a large and diverse group of participants through crowdsourcing.

The structure of the paper is as follows. In section 2 we describe the experimental design, including drawing criteria for the diagrams and the crowdsourcing data collection methodology. Further details on maintaining quality of data are given in section 3, and the results are analyzed in section 4. In section 5 and section 6, we provide a discussion of the results and their validity, respectively. Finally, we conclude in section 7. All of the diagrams used in our study, and the data collected, are available from `www.eulerdiagrams.com/set`.

2 Experimental Design

We are aiming to establish the relative impact on user comprehension of four different diagrams types that visualize sets. For the purpose of this study, as with previous studies, e.g. [12–14], we measure comprehension in terms of task performance using time and error data. We adopted a between group design with one participant group for each diagram type to reducing learning effect. A further advantage of a between groups design was that participants only had to be trained in one notation. We recorded two dependent variables: the time taken to answer questions and whether the answer was correct. Each participant group was shown a set of diagrams about which they were asked a set of questions. If diagram type impacts on comprehension then we would expect to see significant differences between time taken to answer questions or error rates.

2.1 Sets to be Visualized

Each diagram represented a collection of sets with varying relationships between them. Each such collection involved either three, five, or seven sets, and we had six collections of each number of sets (thus, 18 in total). This was to ensure that the questions exhibited a range of difficulties, thus requiring varying levels of cognitive effort to answer the questions. The study included 18 questions – one for each collection of sets – and, therefore, 18 diagrams of each of the four types. Further, we wanted to ensure that, for each question, the Venn diagram, Euler

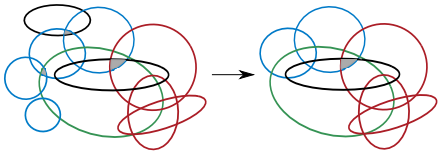

Fig. 2. Generating diagrams from real examples.

diagram with shading, and the Euler diagram without shading were different from each other. For diagrams representing three sets, we chose six combinations of three sets that ensured the diagrams were different.

The choice of diagrams representing five sets and seven sets, respectively, was not significantly limited. Rather than generate random combinations of sets, which might be unlikely to arise in real situations, we turned to Google images to choose diagrams for the five and seven set cases. We searched for examples of all diagram types that people had drawn to visualize data; we could not find any actual examples of linear diagrams and we excluded any diagrams that had been drawn by the authors of this paper. Some of the diagrams returned in this search represented more than five or seven sets. In these cases, some sets were removed, to yield the required number, in such a way as to keep the diagram connected (the curves formed a connected component) whilst ensuring that the number of zones remained as high as possible. This set removal method meant that the diagram was, roughly speaking, close in complexity to the original. An example of a diagram found through Google images can be seen on the left of Fig. 2 (re-drawn here for copyright reasons, approximating the colours used and adding shading to the zones representing empty sets; see `http://govwild.hpi-web.de/images/govwild/overlapLegalEntity.png`). It represents ten sets. Three sets were identified for removal, to yield a 7-set diagram, shown on the right. The reduced 7-set diagram corresponds to the four diagrams shown in Fig. 1 that were used in the study.

Since displaying real data can lead to bias (through the potential for prior knowledge), the names of the sets were changed to a pseudo-real context, focused on three domains: film collections, subjects studied, and languages spoken. It was anticipated that participants would have a reasonable preconception of this kind of information, but no prior knowledge of the (fictional) information visualized.

2.2 Study Questions

As this study aims to establish which of the four notations is most effective for accessing information about sets, statements made about the diagrams were chosen to adhere to the following *templates*:

1. **Simple question templates:**
   (a) **Intersection:** Some ⟨elements in X are⟩ also ⟨in Y⟩.
   (b) **Subset:** Every ⟨element in X is⟩ also ⟨in Y⟩.

   (c) **Disjointness:** No ⟨element in X is⟩ also ⟨in Y⟩.
2. **Complex question templates:**
   (a) **Intersection:** Some ⟨elements in X and Y are⟩ also ⟨in Z⟩.
   (b) **Subset:** Every ⟨element in both X and Y are⟩ also ⟨in Z⟩.
   (c) **Disjointness:** No ⟨elements in both X and Y are⟩ also ⟨in Y⟩.

Every statement was prefixed with "This diagram shows ⟨some contextual text goes here⟩. Is the following statement true?" and participants were asked to choose the answer 'yes' or 'no'. The statement templates were populated with context-specific text by randomly choosing the sets for X, Y and, where necessary, Z. The actual phrasing of the individual statements was far less mathematical in style than the templates just given. One of each type of complex question is given here, one from each domain. The four diagrams associated with question 3 are in Fig. 1:

1. **Intersection:** This diagram shows the subjects studied by Mrs Robinson's students. Is the following statement true? Some of those studying both Geology and History are also studying Music.
2. **Subset:** This diagram shows the classifications of films owned by Grace. Is the following statement true? Every film classified as both Action and Thriller is also classified as Period.
3. **Disjointness:** This diagram shows the languages spoken by employees at Interpro Translators. Is the following statement true? No one who speaks both Welsh and Italian also speaks Turkish.

2.3 Diagram Specification and Layout Characteristics

All of the diagrams were drawn sensitive to various layout guides, used to minimize variability across types. These guides also helped ensure that each diagram type was not compromised by bad layouts, but to-date only some of these guides have been verified by empirical testing. The following conventions were adopted:

1. Curves/lines were drawn with a 6 pixel stroke width.
2. Diagrams were drawn in an area of 810 by 765 pixels.
3. Curves/lines representing a particular set were given the same colour. No two sets, appearing in the same diagram, had same colour.
4. Set names had an using upper case first letter in Sans font, 24 point size.
5. Set names were positioned closest to their corresponding curve/line and took the same colour.
6. The set names used in any one diagram started with a different first letter.
7. The same colour (grey) shading was used across diagrams where relevant.

A palette of seven colours was generated using colorbrewer2.org (accessed November 2013), in a similar fashion to [15]. Colour generation using the Brewer colour palette is recognized as a valid approach for empirical studies, such as in the context of maps [16]. So that the colours were distinguishable, but not sequential or

suggestive (e.g. increasingly vivid shades of red used to denote heat), they were generated using the 'qualitative' option, based on work by Ihaka [17].

In Fig. 1 the diagrams exhibit all of these layout choices (although they are scaled). Further layout conventions were adopted for each diagram type, using results from the literature that guide us toward effective layouts [2],[7],[8],[18]. The conventions were as follows:

Venn diagrams:

1. The curves were drawn smoothly.
2. The overlaps were drawn so that the zone areas are similar.
3. The diagrams were drawn well-formed; see [19].
4. The diagrams had rotational symmetry, using layouts given in [20].

Euler diagrams with shading:

1. The curves were circles where possible, otherwise ellipses were used.
2. The diagrams were drawn well-formed.
3. The number of shaded zones was kept minimal.

Euler diagrams without shading:

1. The curves were drawn smoothly, with recognizable geometric shapes (such as circles, ellipses or semi-circles), or with rectilinear shapes.
2. The diagrams were drawn as well-formed as possible, aiming to minimize (in this order of priority, based on [2]): concurrency between curves, non-simple curves, triple points, and brushing points (points were two curves meet but do not cross).

Linear diagrams

1. The number of line segments representing each set was kept small.
2. Favour layouts where, when reading from left to right, the number of overlapping line segments changes minimally.

In order reduce the number of line segments, the set with the largest number of intersections with the other sets was drawn using a single line segment.

2.4 Data Collection Methods

For this study, we adopted a crowdsourcing approach and we used Amazon Mechanical Turk (MTurk) [21, 22] to automatically out-source tasks to participants. In MTurk, the tasks are called HITs (Human Intelligence Tasks) which are completed by anonymous participants (called workers) who are paid if they successfully complete the HIT. The use of crowdsourcing platforms for conducting research-oriented studies is becoming more popular. Thus, as this method for collecting data has now gained recognition within the scientific community. In particular, there is evidence that it is a valid approach, where [22] compared lab-based experiments with MTurk, showing that no significant differences arise in the results. Moreover, MTurk has been specifically used to collect performance data in other scientific studies in the visualization field, such as [23].

The MTurk HITs were based on the templates provided by Micallef et al.[24], at `http://www.aviz.fr/bayes`. 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. The questions in the main study were randomly ordered to reduce ordering effects.

After every five study questions, participants were asked to answer a question designed to identify inattentive workers (spammers) and those that had difficulties with the language. In MTurk there is little control over who participates in the study and, so, some workers may fail to give questions their full attention [21]. A recognized technique for identifying workers who cannot understand the language used is to include questions that require careful reading, yet are very simple to answer (e.g, [25]). In our study, these questions asked participants to click on a specific area in the diagram, whilst still presenting the participants with (redundant) radio buttons for the 'yes' and 'no' answers seen for the 18 main study questions. Participants were classified as spammers if they clicked a radio button on more than one of the four spammer-catching questions included in the study. All data obtained from spammers were removed before analysis.

3 Experiment Execution

Initially 20 participants took part in the pilot study (1 spammer). The pilot study proved the experimental design to be robust, with a few minor changes made to the wording of the questions (mostly due to typographical errors). A further 440 participants were recruited for the main study. Of note is that we only allowed MTurk workers with a HIT approval rate of at least 95% to participate. All participants were randomly allocated to one of the four diagram types in equal numbers. There were 16 participants identified as spammers, leaving each participant group with the following number of participants: Venn diagrams 107, Euler diagrams with shading 109, Euler diagrams without shading 106, and linear diagrams 102. The ID of all the workers that either completed one of our HITs or started and returned the HIT before completion was recorded. A worker whose ID was previously recorded was not assigned a HIT, so preventing multiple participation. The participants performed the experiment at a time of their choosing, in a setting of their choosing. They were told that the experiment would take approximately 20 minutes, based on participants' performance in the pilot study, and were paid $1 to take part (this was reduced from $1.50 for the pilot study, as all 20 HITs were completed within 30 minutes). For the main study, the data were collected within 24 hours, with HITs made available in sequential batches of 100, and a final batch of 40. We also note that $1 for approximately 20 minutes work is higher than is typical for MTurk workers, for example [22].

At the beginning of the study, each participant was told that they could only participate once in the study and instructed to read the questions carefully. They were further advised that they had to answer 75% of "key questions" correctly

in order to be paid (i.e. not classified as a spammer). They were further advised that the first five pages of the HIT were training, which was the first phase of the experiment. During this phase, participants attempted questions and were told whether they had answered correctly, with the answer was explained to them. An example of a training page can be seen in Fig. 3. The super-imposed rectangles highlight the two radio buttons and show the text displayed *after* the participant had clicked 'reveal answer'.

The participants then entered the data collection phase. This began with three questions, in addition to the 18 main questions. The data relating to these first three questions was not included in the analysis, in order to reduce the impact of any learning effect. Consequently, the following results are based on 424 participants each answering 18 questions, giving $18 \times 424 = 7632$ observations.

4 Statistical Analysis and Results

We now proceed to analyze the data collected by considering time taken and error rate. The raw data are available from `www.eulerdiagrams.com/set`, allowing the computation of non-essential basic statistics omitted for space reasons.

4.1 Time Data

The grand mean was 20.452 seconds (standard deviation: 23.620) and the mean times taken to answer questions by diagram type are: Venn diagrams 24.477 (sd: 24.158); Euler diagrams with shading 20.532 (sd: 26.996); Euler diagrams without shading 19.810 (sd: 20.121); and linear diagrams 16.813 (sd: 21.843). In many tables below, we abbreviate the diagrams types' names to Venn, shaded ED, unshaded ED and linear respectively. In order to establish if there is significant variation across diagram types, we conducted an ANOVA. The results of the ANOVA are summarized in table 1, which uses log (base 10) of time. Although the data are not normal, even after taking logs, the skewness of the logged data is 0.39 and the sample size is 7632, making the ANOVA test robust; the raw data have a skewness of 12.08, which is outside the permissable range

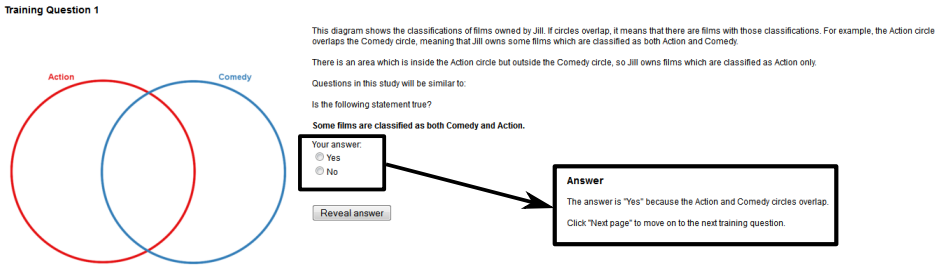

Fig. 3. The first training page for the Venn diagram group.

of ±2 for the ANOVA to be valid. We regarded p-values of less than 0.05 as significant, unless stated otherwise.

The row labelled question, with a p-value of 0.000, tells us that there are significant differences between the mean times of at least one pair of questions. This shows robustness, in that the questions are sufficiently varied in that they required different amounts of user effort to answer.

The row for diagram, with a p-value of 0.000, tells us that there are significant differences between the mean times taken to answer questions across the diagram types. That is, diagram type significantly impacts on task performance. Next, we performed a Tukey test to compare pairs of diagram types, thus establishing whether one mean is significantly greater than the other, in order to rank the diagram types. Any p-values of less than 0.01 were regarded to be significant, given multiple comparisons being made on the same data. Table 2 presents the rankings of diagram type by mean time taken. We see that **linear diagrams allow participants to perform significantly faster** on tasks than all other diagram types. There is no significant difference between shaded Euler diagrams and Euler diagrams without shading, whereas **Venn diagrams cause participants to perform significantly slower** on tasks. Thus, we conclude that linear diagrams are the most effective diagram type, with respect to time taken, followed by both unshaded Euler diagrams and shaded Euler diagrams and, lastly, Venn diagrams.

The magnitude of the significant differences is reflected in the effect sizes given in table 3. For example, the largest effect size tells us that 62% to 66% of participants were faster interpreting linear diagrams than the average person using Venn diagrams. The effect sizes all suggest that not only are the differences in mean time taken significant but, taken with the mean times, real differences in task performance will manifest through their use in practice.

Continuing now with our interpretation of the ANOVA table, the row for diagram*question, with a p-value of 0.000, tells us that there is a significant interaction between diagram type and question: the diagram type used impacts user performance for at least one question. We further investigated this manifestation by running another ANOVA, looking for an effect of question type and an interaction between diagram type and question type. This would establish whether there was any obvious systemmatic way of describing the interaction between diagram type and question. The ANOVA showed no significant effect

Source	DF	MS	F	P
question	17	4.3966	108.32	0.000
diagram	3	7.5847	13.52	0.000
diagram*question	51	0.2155	5.31	0.000
participant(diagram)	420	0.5584	13.76	0.000
Error	7140	0.0406	–	–
Total	7631			

Table 1. ANOVA for the log of time.

Diagram	Mean	Rank
Venn	24.477	C
Shaded ED	20.532	B
Unshaded ED	19.810	B
Linear	16.813	A

Table 2. Pairwise comparisons.

of question type ($p = 0.201$) and no interaction between diagram and question type ($p = 0.171$). This implies that **task performance is not affected by question type**.

4.2 Error Data

Regarding errors, of the 7632 observations there were a total of 1221 errors (error rate: 16%). The errors were distributed across the diagram types as follows: Venn diagrams 391 out of 1926 observations; Euler diagrams with shading 377 out of 1962; Euler diagrams without shading 258 out of 1908; and linear diagrams 195 out of 1836. We performed a χ^2 goodness-of-fit test to establish whether diagram type had a significant impact on the distribution of errors. The test yielded a p-value of 0.000. Thus, the number of errors accrued is significantly affected by diagram type. Investigating further, table 4 summarizes where significant differences exist. We conclude that **linear diagrams accrued significantly fewer errors** than all other diagram types. Moreover, **significantly more errors were accrued using Venn diagrams and shaded Euler diagrams** than the other two diagram types.

It is natural to ask whether question type impacts error rate by diagram type. Table 5 summarizes the raw data for error counts for each diagram type, broken down by question type. Statistically analyzing these data, by question type, reveals significant differences in all cases. In particular, our analysis revealed that, for all question types, linear diagrams lead to significantly fewer errors. For intersection and subset questions, Venn diagrams make a large contribution to the χ^2 statistic, indicating that they account for a significantly large number of errors. Lastly, for disjointness questions, Euler diagrams with shading make a large contribution to the χ^2 statistic, indicating that they account for a significantly large number of errors. In summary, our analysis of the errors suggests that linear diagrams are the most effective diagram type, with Venn diagrams being the worst, except for questions on disjointness where Euler diagrams with shading are worst.

4.3 Summary of Results

Linear diagrams allow users to perform most effectively in terms of both completion time and correctness: the mean time taken was significantly faster than for all other diagram types and the number of errors was significantly lower. By contrast, Venn diagrams were ranked bottom for both time taken and, jointly with

Diagram	Unshaded	Shaded	Venn
Linear	54%-58%	54%-58%	62%-66%
Shaded	–	–	54%-58%
Unshaded	–	–	54%-58%

Table 3. Effect sizes.

Diagram	Unshaded	Shaded	Venn
Linear	Y	Y	Y
Unshaded	–	Y	Y
Shaded	–	–	N

Table 4. Error differences.

	Intersection		Subset		Disjointness	
Diagram	Error	Correct	Error	Correct	Error	Correct
Venn	152	490	107	535	132	510
Shaded	122	532	96	558	159	495
Unshaded	100	536	70	566	88	548
Linear	70	542	51	561	74	538
p-values	0.000		0.000		0.000	

Table 5. Error counts and significance.

shaded Euler diagrams, error rate. Thus, the error analysis allows us to distinguish Euler diagrams with shading from Euler diagrams without shading, which were not significantly different in terms of time taken. We can therefore give a ranking of diagram types using both time and errors, in order of effectiveness:

1. linear diagrams,
2. Euler diagrams without shading,
3. Euler diagrams with shading,
4. Venn diagrams.

5 Subjective Discussion

We now seek to explain our results in the context of theories about diagrams, cognition and perception. One feature of diagrams that is thought to correlate with their effectiveness is *well-matchedness*, introduced by Gurr [18]. A diagram is *well-matched* if its syntax directly reflects its semantics. Euler diagrams without shading are well-matched because the spatial relationships between the curves directly mirror the relationships between the sets they represent. Linear diagrams are also well-matched because the spatial relationships between the lines also directly mirror the relationships between the sets. For example, in Fig. 1, the lines representing German and Italian do not overlap, and the represented sets are disjoint. By contrast, Euler diagrams with shading are not well-matched because of the additional zones used to represent empty sets. Likewise, Venn diagrams that include shading are not well-matched. That is, the spatial relationships between their curves does not directly mirror the relationships between the represented sets. Thus, our results - with linear diagrams and Euler diagrams without shading being more effective than Euler diagrams with shading and Venn diagrams, support Gurr's theory.

An interesting point is that the Euler diagrams with shading and the Venn diagrams used in the study were all *well-formed*, whereas those that did not use shading were all non-well-formed (see e.g. [2]). Rodgers et al., in [2], established that well-formed diagrams are more effective than equivalent diagrams that are not well-formed (the diagrams in [2] did not use shading). Thus, our results indicate that being well-matched is more important than being well-formed.

Another way of examining differences between diagram types is through *visual complexity*. For the Venn and Euler family, one measure of visual complexity

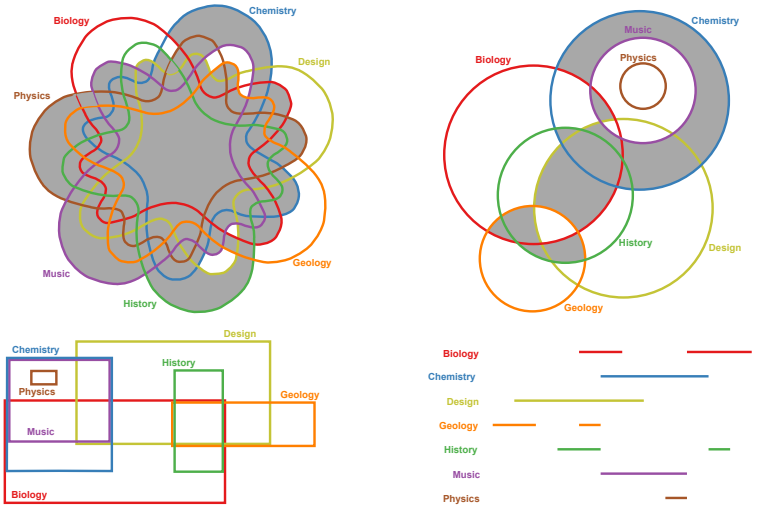

Fig. 4. Diagram complexity.

arises through the number of crossings between curves. In our study, Venn diagram exhibited more crossings than Euler diagrams with shading which, in turn, had more than Euler diagrams without shading. By contrast, linear diagrams have no line crossings. For example, in Fig. 4, the Venn diagram has 121 crossings between curves, the Euler diagram with shading has 20, the Euler diagram without shading has 9 (and a further 4 points where curves meet but do not cross), and the linear diagram has none. The results of our study lead us to hypothesize that this measure of visual complexity, at least for the diagrams in this study, correlates with diagram effectiveness and, moreover, helps to explain why linear diagrams are the most effective.

We can further explain why linear diagrams are interpreted more quickly, and with fewer errors than all other diagram types and, in particular, Euler diagrams without shading. With regard to linear diagrams, we quote Wagemans et al. "[the] comparison of features lying on pairs of line segments is significantly faster if the segments are parallel or mirror symmetric, suggesting a fast grouping of the segments based on these cues [26]", who reference Feldman [27] as the source of this insight. As we have seen, linear diagrams use parallel line segments and so are thought to be effective for this reason. However, as Wagemans et al. consider the relative effectiveness of diagrams drawn with lines, this alone does not explain why linear diagrams are more effective than the other diagram types.

To gain more insight into our observations, we consider the work of Bertin who describes graphical features consisting of elements and properties [28]. Closed curves and lines can be regarded as elements. Properties include shape and size. Bertin, recognizing our visual sensitivity to graphical properties, proposes eight visual variables. Two of these, called planar variables, are the x and y coordinates in the plane. Venn and Euler diagrams are not constrained by planar variables:

drawing their closed curves is not constrained by x or y coordinates. Therefore, each closed curve's position in the plane is arbitrary other than topological constraints imposed by the sets being visualized. Conversely, linear diagrams are constrained by planar variables. Lines are ordered vertically and run horizontally, in parallel. In the context of this study, a top down hierarchy is imposed, along the y axis, based upon the alphabetical order of set names, and this layout feature is thought to aid reading the diagram. There is no such prescribed order in the Euler and Venn family. Moreover, relationships between combinations of sets can be 'read off' along the x axis. Consequently, the prescriptive planar layout of linear diagrams, as opposed to the 'free' (disordered) positioning of curves in the other diagram types, is thought to aid comprehension.

6 Threats to Validity

Threats to validity are categorized as internal, construct and external [29]. The following discusses threats to validity, focusing primarily on those arising from using a crowdsourcing approach, that were considered and addressed to ensure the study is robust and fit for purpose. With regard to internal validity, the following factor was among a number that were considered in our study design:
Laboratory: ideally, all participants undertake the study in the same environment, ensuring each participant was exposed to the same hardware, free from noise and interruption. By adopting a crowdsourcing approach, we had no control over the environment in which each participant took part. To reduce the effect of this compromise, a large data set was collected, with over 400 participants.

Now we consider construct validity, examining the rigour of our dependent variables and independent variables for measuring comprehension:
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 crowdsourcing 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 *diversity* was introduced in the diagrams so that participants had to read and understand each diagram 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 three, five or seven sets in order to demand cognitive effort. Lastly, the study included questions to allow spammers to be identified, catching those who did not read questions carefully.

The following factor considers the limitations of the results and the extent to which the they can be generalized, thus examining their external validity:
Participant: participants were representative of the wider population, being MTurk workers. They were predominately from the USA or India. Thus, the results should be taken to be valid within these constraints.

7 Conclusion

In this paper we have examined four diagram types that are used for visualizing sets. By conducting an empirical study, we have established that task performance is significantly better when using linear diagrams over the Euler diagram family, comprising Venn diagrams, Euler diagrams with shading and Euler diagrams without shading. Furthermore, the Euler diagrams variants that we tested can be ordered: Euler diagrams without shading were most effective, Euler diagrams with shading were next and, finally, Venn diagrams proved to be least effective, having both poor time and accuracy performance. Given the prevalent use of Euler and Venn diagrams for visualizing sets, and the relative lack of use of linear diagrams, these results have implications for the use of diagrams in practice. Our results suggest that linear diagrams should be more widely adopted, at least for use by the general population. It would be interesting to establish whether these results manifest for expert users also.

Looking to the future, we plan to conduct further studies that augment the syntax of these diagrams with data items (i.e. set elements). Many diagrammatic systems, such as spider diagrams and Euler/Venn diagrams, exploit Euler diagrams with graphs to represent sets and their elements. It will be interesting to establish whether linear diagrams should instead be adopted for representing this more complex data. Moreover, the result suggest that the number of shaded zones, in the Euler and Venn diagram family, could impact on performance. However, the current study did not control this variable and further study will be needed to gain insight into their effect.

References

1. Larkin, J., Simon, H. Why a diagram is (sometimes) worth ten thousand words. *J. of Cognitive Science*, 11:65–99, 1987.
2. Rodgers, P. , Zhang, L. , Purchase, H. Wellformedness properties in Euler diagrams: Which should be used? *IEEE Trans. on Visualization and Computer Graphics*, 18(7):1089–1100, 2012.
3. Couturat, L. *Opuscules et fragments inédits de Leibniz.* Felix Alcan, 1903.
4. Wittenburg, K., Lanning, T., Heinrichs, M., Stanton, M. Parallel Bargrams for Consumer-based Information Exploration and Choice. *14th ACM symposium on User interface software and technology*, pp. 51–60, 2001. 1985.
5. Hofmann, H., Siebes, A., Wilhelm, A. Visualizing Association Rules with Interactive Mosaic Plots *ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 227–235, 2000.
6. Moody, D. The "Physics" of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering. *IEEE Trans. on Software Engineering*, 35(6):756–779, 2009.
7. Benoy, F., Rodgers, P. Evaluating the comprehension of Euler diagrams. In *11th Int. Conf. on Information Visualization*, pp. 771–778. IEEE, 2007.
8. Blake, A., Stapleton, G., Rodgers, P., Cheek, L., Howse, J. Does the orientation of an Euler diagram affect user comprehension? In *18th Int. Conf. on Distributed Multimedia Systems*, pp. 185–190. Knowledge Systems Institute, 2012.

9. Grawemeyer, B.. Evaluation of erst – an external representation selection tutor. In *Proc. Diagrams*, pp. 154–167. Springer, 2006.
10. Sato, Y., Mineshima, K.. The Efficacy of Diagrams in Syllogistic Reasoning: A Case of Linear Diagrams. In *Diagrams*, pp. 352–355. Springer, 2012.
11. Sato, Y., Mineshima, K., Takemura, R. The Efficacy of Euler and Venn Diagrams in Deductive Reasoning: Empirical Findings. In *Diagrams*, pp. 6–22. Springer, 2010.
12. Isenberg, P., Bezerianos, A., Dragicevic, P., Fekete, J. A study on dual-scale data charts. In *IEEE Tran. on Visualization and Computer Graphics*, pp. 2469 – 2478. IEEE, 2011.
13. Puchase, H.. Which aesthetic has the greatest effect on human understanding? In *5th Int. Symp. on Graph Drawing*, pp. 248–261. Springer, 1997.
14. Riche, N., Dwyer, T.. Untangling Euler diagrams. *IEEE Tran. on Visualization and Computer Graphics*, 16(6):1090–1099, 2010.
15. Harrower, M., Brewer, C. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *Cartographic Journal, The*, 40(1):27–37, 2003.
16. Silva, S., Madeira, J., Santos, B.S. There is more to color scales than meets the eye: A review on the use of color in visualization. In *Information Visualization* , pp. 943–950. IEEE, 2007.
17. Ihaka, R. Colour for presentation graphics. In *3rd Int. Workshop on Distributed Statistical Computing*, 2003.
18. Gurr, C.. Effective diagrammatic communication: Syntactic, semantic and pragmatic issues. *J. of Visual Languages and Computing*, 10(4):317–342, 1999.
19. Stapleton, G., Rodgers, P., Howse, J., Taylor, J. Properties of Euler diagrams. In *Proc. of Layout of Software Engineering Diagrams*, pp. 2–16. EASST, 2007.
20. Ruskey, F. A survey of Venn diagrams. *Electronic J. of Combinatorics*, 1997. www.combinatorics.org/Surveys/ds5/VennEJC.html.
21. Chen, J., Menezes, N., Bradley, A., North, T. Opportunities for crowdsourcing research on amazon mechanical turk. *Human Factors*, 5(3), 2011.
22. Paolacci, G., Chandler, J., Ipeirotis, P.G. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419, 2010.
23. Heer, J., Bostock, M. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *ACM Conf. on Human Factors in Computing Systems*, pp. 203–212, 2010.
24. Micallef, L., Dragicevic, P., Fekete, J.D. Assessing the Effect of Visualizations on Bayesian Reasoning through Crowdsourcing. *IEEE Trans. on Visualization and Computer Graphics*, 18(12):2536-2545, 2012.
25. Oppenheimer, D., Meyvis, T., Davidenko, N. Instructional manipulation checks: Detecting satisficing to increase statistical power. *J. of Experimental Social Psychology*, 45(4):867–872, 2009.
26. Wagemans, J., Elder, J., Kubovy, M., Palmer, S., Peterson, M., Singh, M. A century of gestalt psychology in visual perception: I. perceptual grouping and figure-ground organisation. *Computer Vision, Graphics, and Image Processing*, 31:156–177, 1985.
27. Feldman, J. Formation of visual "objects" in the early computation of spatial relations. *Perception and Psychophysics*, 69(5):816–827, 2007.
28. Bertin, J. Semiology of Graphics, *Uni. of Wisconsin Press*, 1983.
29. Purchase, H. *Experimental Human Computer Interaction: A Practical Guide with Visual Examples.* Cambridge University Press, 2012.