Measuring Perceived Clutter in Concept Diagrams

Tie Hou
Visual Modelling Group
University of Brighton, UK
Email: t.hou@brighton.ac.uk

Peter Chapman
School of Computing
Edinburgh Napier University
Email: p.chapman@napier.ac.uk

Ian Oliver
Bell Labs, Nokia
Espoo, Finland

***Abstract*—Clutter in a diagram can be broadly defined as how visually complex the diagram is. It may be that different users perceive clutter in different ways, however. Moreover, it has been shown that, for certain types of diagrams and tasks, an increase in clutter negatively affects task performance, making quantifying clutter an important problem. In this paper we investigate the perceived clutter in concept diagrams, a visual language used for representing ontologies. Using perceptual theory and existing research on clutter for other diagrams, we propose five plausible measures for assigning clutter scores to concept diagrams. By performing an empirical study we evaluated each of these proposed measures against participants' rankings of diagrams. Whilst more than one of our measures showed strong correlation with perceived clutter, our results suggest that a measure based on the number of points where lines cross is the most appropriate way to quantify clutter for concept diagrams.**

I. Introduction

The use of diagrams, either as illustrative aids, or as foundations for visual languages, is becoming increasingly widespread. To aid the uptake of diagrams it is essential that the representation leads users quickly and consistently to the correct interpretation of the represented data. In other words, for visualization to be an effective tool it is imperative that diagrams are drawn according to guidelines that we know aid comprehension. However, existing guidelines (such as the Gestalt principles [1], Moody's Physics of Notations [2] or Miller's 7 ± 2 [3]) are theoretically-based and general. They each give a set of ideals a visual notation should adhere to that *could* aid comprehension. But, the guidelines need to be interpreted and implemented for each individual notation, and the empirical evaluation is then often omitted. Moreover, no set of guidelines discusses the visual clutter a diagram may exhibit, instead talking of shape, number of visual symbols, etc. Whilst some of these aspects no doubt *contribute* towards the clutter in a diagram, we instead focus directly on the aspect of clutter.

In particular, we look at clutter in diagrams representing ontologies. Ontologies are widely used to represent a domain of knowledge, for example in biomedicine [4], law [5] and the Semantic Web. Their application is in areas for which accuracy of information is paramount, and it is known that ontology engineering is a difficult task [6]. Various visualizations have been proposed to represent ontologies [7] of which our object of study, concept diagrams, is one. By creating a measure for clutter in concept diagrams, we can show that relatively low cluttered visualizations can be effective tools for aiding ontology engineering.

A. Structure of the paper

The structure of the remainder of the paper is as follows. We give an overview of the syntax, semantics, and usage of concept diagrams in section II. In section III, we examine the notion of clutter, especially as applied to other types of diagrams, and outline our proposed clutter measures for concept diagrams in section III-A. The design of our empirical study is explained in section IV, with the results presented and analysed in section VI. Finally, we conclude and indicate future directions in section VII.

II. Concept diagrams

This section is intended only as an overview of the syntax and semantics of concept diagrams. For a fuller exposition, see [8], [9]. As will be seen in section IV, we will be presenting concept diagrams to participants without any labels, and thus will not represent any particular context. However, we will explain the meaning of concept diagrams by reference to ontologies. A description of the language of ontologies can be found in [10], [11].

The basic elements in concept diagrams are curves, boxes and arrows. Curves represent sets of objects. Their arrangement in the plane, in particular their interaction with other curves, represents the underlying relationship between sets. Where a region exists inside two curves, we can infer that the represented set is possibly non-empty. For example, where one curve is completely contained within another, we are asserting that the set represented by the former is a subset of the set represented by the latter. In Figure 1 we have that the curve labelled A is completely inside the curve labelled B. Abusing notation, if the curve labelled A represents the set A, and similarly for B, then we are asserting that $A \subseteq B$. In the language of ontologies, we would be asserting $\mathsf{A} \sqsubseteq \mathsf{B}$. Crucially, we do not specify whether this subset relation is proper or not: where a region exists we still allow that the represented set can be empty. By contrast, where two curves do not intersect, we are asserting that the represented sets must have empty intersection. In Figure 1, we see that the curves C and D do not intersect, and thus we can infer that $C \cap D = \emptyset$

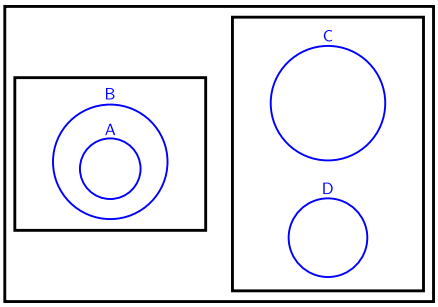


Fig. 1. Concept diagram example: curve

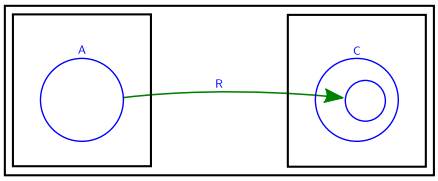


Fig. 2. Concept diagram example: arrow

for the represented sets. In the language of ontologies, we would say this situation encodes the axiom $\mathsf{C} \sqcap \mathsf{D} \sqsubseteq \bot$.

Each box represents the universe of elements. Several boxes within the same diagram (for example in Figure 1) thus represent partial information: we know that explicit relationships hold between curves inside the same box, but we are asserting nothing about the relationship between curves in separate boxes. For example, we do not assert anything about the relationship between A and C.

Arrows represent binary relations between sets, and in the context of ontologies the object properties between concepts. For example, in Figure 2 the arrow labelled R joins the curve A with an unlabelled curve. The interpretation of this is that, between them, all the elements of A are related to all, and only, the elements of the unlabelled set. For ontologies, this represents the axiom $\mathsf{A} \sqsubseteq \forall\, \mathsf{R}.\mathsf{C}$. Arrows can also be sourced, or targeted, on boxes. This situation is used to model the domains and ranges of the represented properties.

There are choices to be made when drawing a concept diagram. The same underlying information can be represented in a number of ways, and testing which of these ways is most visually appealing is the focus of this paper. As an example, consider the two diagrams in Figure 3. The left-hand diagram (d_1) uses two boxes, allowing us to assert that the curve labelled A can have any relationship with the curves labelled B, C and D. We could instead place all information in a single box, however, as shown on the right (d_2) of Figure 3. We now have to explicitly show all of the possible interactions between the curve A and the others. (Note that some represented regions may be empty.)

III. Clutter

Clutter is a difficult concept to define precisely, although Rosenholtz *et al.* [12] define it as "the state in which excess items, or their representation or organization, lead to a degradation of performance at some task." Interestingly, they define it in terms of task degradation, in other words related to performance, rather than an independent feature of the diagram. Clutter is only defined, then, in relation to a task. What may be a cluttered representation for one task may be uncluttered for another. In this paper, by contrast, we seek to find an absolute measure for clutter in a diagram-type, rather than the clutter for a diagram-task pair.

Related to clutter is the notion of visual complexity, although that is more usually applied to images in general. In [13], visual complexity was defined as "the amount of detail or intricacy of the line", whereas in [14] the visual complexity of an image was defined as the size of an image file after compression. These measures are based on some objective measure (in the case of a file size), or the pleasure a person feels when looking at the image.

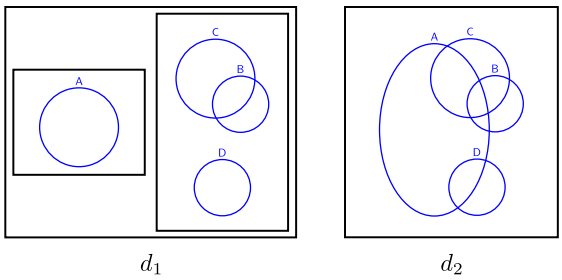
Clutter in Euler diagrams (the foundation for concept diagrams) was examined in [15], and a number of measures were proposed to quantify it. Participants were asked to assign a score for each diagram, on a scale of 1 to 100, as to how cluttered they thought it was. Each measure was compared with users' perceptions, and the one most aligned was the *contour score*, described in section III-A. Interestingly, this paper used a cardinal scale for perception. So, users were asked not to compare two diagrams and say which was more cluttered, but effectively *by how much* one was more cluttered than the other. Important to note is that John *et al.*, in [15] made a distinction between *abstract* and *concrete* clutter. Abstract clutter (or structural clutter) was clutter that was fundamental to the information being represented, independent of drawing choices. A focus on abstract clutter led them to using only black curves. By contrast, concrete clutter takes into account the drawing choices, such as using coloured curves. Since we know that concept diagrams will be drawn in practice using colour, we will focus on concrete clutter, in John *et al.*'s terminology.

Quantifying clutter is important owing to its effect on comprehension. In [16] it was found that increased clutter, according to the measure in [15], negatively affected comprehension. Similarly, in [14], it was found that there was a negative linear relationship between the visual complexity of a website and the pleasure participants had in navigating the website; in other words visually complex websites were viewed as less pleasurable to use.

A. Proposed clutter measures

We describe five plausible clutter measures to be tested, justifying each proposal.

1) Ontology complexity score (OCS): The diagrams we present to participants will have no labels and thus, in some sense, are not representative of any particular context. How-



d_1 d_2

Fig. 3. Drawing choices

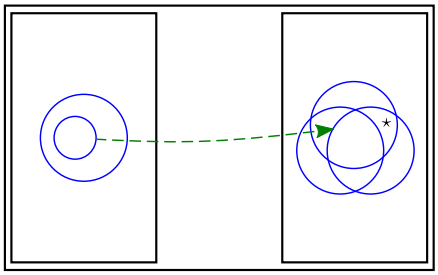


Fig. 4. Measures for a diagram

ever, the intended usage of concept diagrams is for representing ontologies. Thus, our first measure is based on the ontology complexity measure *tree impurity* (hereafter renamed OCS), found in [17]. This measure was evaluated against the criteria in [18], and found to satisfy most of them. A detailed description is given in [17]; we simply give the intuition here. Ontologies consist of hierarchical information; concepts are subclasses of other concepts, which are all subclasses of the top level Thing. The bottom-most concept is Nothing. Sometimes, these hierarchies form a tree. However, in most cases, concepts will be subclasses of more than one parent concept. For example, the concept Cat could be a subclass of has4Legs and Mammal, neither of which is a subclass of the other. Thus, the hierarchy, when drawn as a graph, will not form a proper tree. The OCS then measures "how far" from a tree a certain ontology is, with a higher score representing a more complex ontology. For our purposes, we created our study diagrams by creating small sets of ontology axioms, from which we could calculate the OCS.

The justification for this measure is to check whether the key cause of clutter in a diagram is the complexity of the underlying information. We also note that this is an existing measure that is used in the ontology community for measuring complexity, and so is widely understood.

2) Abstract scores (AS1 and AS2): These measures are based on the abstract syntax of the underlying Euler-based diagram, as defined in [15]. The curves in the diagram (including the boxes) partition the plane into a number of regions. Each region is thus inside a number of curves, and outside a number of others. The score for the diagram (without arrows) is then given by the sum of the scores for each region in the diagram. For example, in Figure 4, the highlighted region (marked with a star) is inside four curves (two curves proper, and two boxes), and thus contributes 4 to the AS. Note that, since boxes are no different to curves, at an abstract level, we treat them in the same manner.

Arrows, in the abstract syntax, are sourced on a particular curve or box, and targeted on a particular curve or box. There are two ways to calculate the contribution of an arrow. The first takes account how deeply nested the endpoints of the arrow are, which is necessarily a feature of a drawn (or concrete) diagram. Where an arrow is sourced on a curve, it touches the curve in one particular location. The contribution is then the scores of the regions where the arrow's endpoints aim at, calculated via the abstract syntax. For example, in Figure 4, the start-point of the arrow touches a region with a score of 4, and the end-point of the arrow touches a region with a score of 5. The contribution of the arrow is then 9. This method gives the score AS1.

The second method of calculation considers an arrow to add clutter whenever the curves it connects contain a large number of regions. We can calculate the score of a curve by counting the number of regions it contains. Thus, the contribution of an arrow is the number of regions in the source curve added to the number of regions in the target curve. This method of counting arrows gives the score AS2. This score relies solely on the abstract syntax of a diagram, and will be invariant under different concrete representations.

The justification for these measures is that they are simple extensions of one which has been shown to align with perceived clutter for Euler diagrams. We can have confidence that it *should* explain the clutter for the curve-based part of the diagram. The difference between the two is owing to how the clutter contribution from arrows is calculated. On the one hand, an arrow more deeply nested in the diagram could create clutter. On the other, the amount of information arrows connect could create clutter. Having both measures allows us to test which aligns with the perception of people.

3) Concrete score (CS): This measure is based on the drawn diagram. We test the assumption that the number of visual objects in a diagram is a proxy for the perceived clutter in a diagram. Each individual object (curves, boxes, arrows) adds one to the CS, and every intersection between two objects also contributes one to the CS. We define the intersection of two objects as either a point where two curves cross each other, or where an arrow crosses a curve or box. We do not count the meeting of the endpoints of an arrow with a source or target as an intersection[1]. Intuitively, the CS captures the number of points of interest in a diagram. There is a similarity between the CS and the AS1/AS2 in that curves and boxes are treated in the same fashion: boxes are just curves with a particular shape. As an example, consider Figure 4. There are 8 curves and boxes, and 1 arrow. The number of intersections between those objects is 11, and thus the CS is 20.

The justification for this measure is that it simply counts

[1]If we did, every arrow would then add at least 3 to the CS: one for the object, and one each for the endpoints.

the number of points on the diagram. Whereas the OCS will not change, and the AS1/2 may not change, depending on different representations of the same information, the CS will almost certainly change depending on how the information is drawn. Since we are asking participants to rank the clutter of diagrams, not of the underlying information, it seems plausible to base our clutter measure on the drawn diagram itself.

4) Hybrid score (HS): The hybrid score is a combination of AS1/2 and CS, which respects that curves, boxes and arrows are all distinct types of object, and so may contribute to clutter in different ways. In HS, the scoring for curves is the same as AS1/2, but each box is not counted as a curve. Rather, it is just counted as a single object, as in CS. In other words, the region marked with a star in Figure 4 would contribute 2 to the HS of the diagram. Similarly, arrows are just objects in the plane regardless of where they are sourced or targeted, and so each contributes only 1 to the HS.

The justification for this measure is that it takes into account both the underlying information (the AS1/2 gives the scores for the curves) and the representation of it (the CS gives the score for the boxes, and partially for the arrows).

IV. Study Design

An empirical study was undertaken to see how people's perception of clutter correlated with the five proposed measures of clutter. The study followed a within-group design, where each participant was shown concept diagrams and asked to rank them on the basis of how cluttered they appeared. The ranking of the diagrams, given by each participant, was the primary variable recorded. We also recorded the time taken by participants to rank the diagrams. Participants were given no further guidance on ranking the diagrams, and no time limit was imposed upon them.

The study consisted of four tasks, each of which required participants to rank 18 concept diagrams, with a ranking of 1 being least cluttered and 18 being most cluttered. Joint rankings were permitted. The first task (T_A) fixed the number of curves and boxes in a diagram, and varied the number of arrows. This task allowed us to establish how perceived relative clutter varied with the number and placement of arrows. The second task (T_B) fixed the number of arrows and curves, and varied the number of boxes. This task allowed us to establish whether merging information from several boxes into a single box affects perceived clutter. The third task (T_C) fixed the number of arrows and boxes, and varied the number of curves. Again, this task allowed us to establish how clutter would vary with the number and placement of curves. The final task (T_D) included concept diagrams with a variety of curves, boxes and arrows. This task allowed us to establish which aspects of the diagram (curves, boxes or arrows) contributes most important to perceived clutter.

A. Diagrams Created for the Study

For each task we generated 18 concept diagrams. The diagrams were designed so that the clutter measures ranked the diagrams differently. In this way, we could examine the relative merits of the proposed measures.

We explain how the diagrams for the tasks were generated.[2]. For T_A and T_C the process used was similar. We give details for T_A. Three base diagrams without arrows were created from a set of ontology axioms (recall that, in order to calculate OCS, we require the diagrams to represent a set of ontology axioms), using 3 boxes and 5 curves. To each diagram we added either 1, 2 or 3 arrows in two different ways, thus each base diagram generated 6 study diagrams. The two methods for adding arrows was to join the source and target curves by the shortest distance (which may cross a number of regions), or to join the source and target curves by using a routing which followed a smooth path and intersected as few other curves and arrows as possible. For example, see Figure 5, where the left-hand diagram uses the fewest crossings and the right-hand diagram uses shorter paths. Each drawing method yields the same AS2, OCS and HS, but possibly different scores according to AS1 and CS. The scores for each diagram and measure are shown in Table I, with the associated rankings. Although the rankings for some of the measures are similar, there are also ranking differences. Thus, we should be able to distinguish which measures most accurately reflect participants' rankings.

For T_C we created three base diagrams with 1 box and 2 arrows which have the same source and target curves for each arrow. To each diagram we then added an extra 1, 2 or 3 curves using the following two methods. In the first, the newly added curves are to intersect as much as possible with the existing arrows. In the second method, as little as possible. In this way, some measures will vary, whilst others will remain constant, allowing us to differentiate them. Figure 6 shows an example of a pair of diagrams from T_C. The left-hand diagram adds a single extra curve interacting with the arrows, and in the right-hand diagram the extra curve does not interact with the arrows. The rankings for the diagrams for each method are shown in Table II. Here, we see less variation amongst the rankings. However, given the existing research, and the restriction of producing diagrams which would be used in practice, this smaller variation is to be expected.

For T_B, we must be more careful, in that we need to ensure that the information in a base diagram (drawn with 5-boxes) was the same as the information in the 4-, 3- and 2-box diagrams generated from it. A process of merging was used (sketched in Figure 3 in section II, but fully explained in [8]) to maintain the information content across the different representations of the same diagram. The side effect of this merging process is that the total number of curves and arrows may *decrease* as boxes are merged. Again, three base diagrams with 6 boxes, 13 curves and 2 arrows were drawn. A merging process then created three 4-box diagrams, three 3-box diagrams, and three 2-box diagrams, yielding 9 diagrams. The merging followed two different methods. In the first, we merge the boxes which contain the most information

[2]Tasks and raw data used in the study can be found at https://sites.google.com/site/visual4onto/file-cabinet.

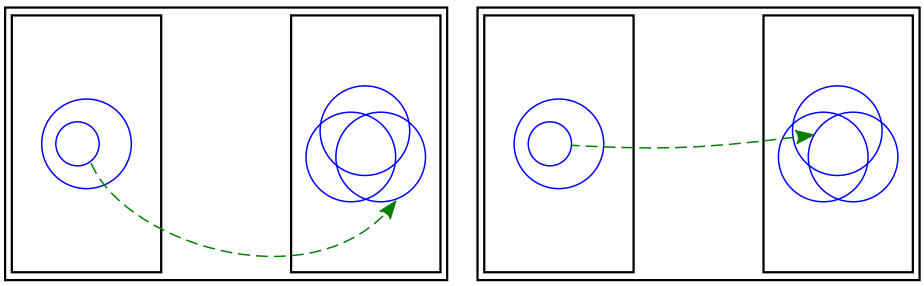


Fig. 5. Drawing choices for arrows

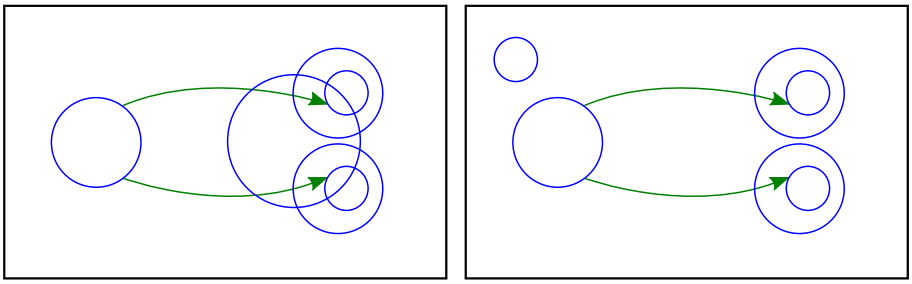


Fig. 6. Drawing choices for curves

in common, whereas in the second we merge the boxes which contain the lowest amount of common information. As an illustration, consider a diagram containing three boxes, two of which contain curves with the same labels, while the third contains none of the same labels as the other two. The first method would merge the two boxes with common curves, and the second method would merge either of the similar boxes with the dissimilar one. Examples of the different diagrams produced for the study can be seen in Figure 7, with the diagram on the left the result of merging two boxes with little information in common (resulting in the box with many curve intersections), and the diagram on the right the result of merging two boxes with common information. The rankings for the diagrams, according to the proposed measures, are found in Table II. As in T_A, there are more variation amongst the measures, as is to be expected owing to the different way each treats boxes.

For the final task T_D, we generated 18 initial diagrams, with either 1 or 3 boxes, 5, 7 or 9 curves, and 1, 2 or 3 arrows. Half were drawn with the method for drawing arrows as described for T_A. This task allowed us to vary arrows, boxes and curves at once. The rankings for the diagrams are shown in Table II. There is variation amongst the rankings as well as similarity, allowing us to see which is most aligned to the participants' ranking.

B. Concept Diagram Layout

When drawing the diagrams for the study, we were careful to control their layout features to minimise (so far as practically possible) unintended variation. We adopted the following layout guidelines:

1) All diagrams were drawn with black lines for boxes, blue lines for curves, and green lines for arrows. This matches the presentation in [8] and [9], and thus makes our results more applicable to concept diagrams as used in practice.
2) The stroke width for boxes was set to 3 pixels, and curves and arrows 2 pixels.
3) There were no labels in the diagrams.
4) The diagrams were printed in the center of A5-size paper.

We chose not to include labels because there we did not participants to perform any task with the diagrams other than ranking them by how cluttered they were perceived. As noted in section III, we attempted to find a measure of clutter independent of task.

V. Experiment Execution

Initially 6 participants (3 M, 3 F, ages 21-42) took part in a pilot study. The pilot study was successful and the participants finished the four tasks in less than half an hour. As no changes were deemed necessary, the pilot data was carried forward for analysis with the data collected in the main study phase. A further 63 participants were recruited, giving a total of 69 participants (30 M, 39 F, ages 18-46). All the participants were staff or students from the University of Brighton; none of them were members of the authors' research groups.

The participants performed the experiment within a room that provided a quiet environment free from interruption and noise. Each participant was alone during the experiment, except that the experiment facilitator was present. Each participant was informed not to discuss the details of the study with the other participants who are yet to perform the

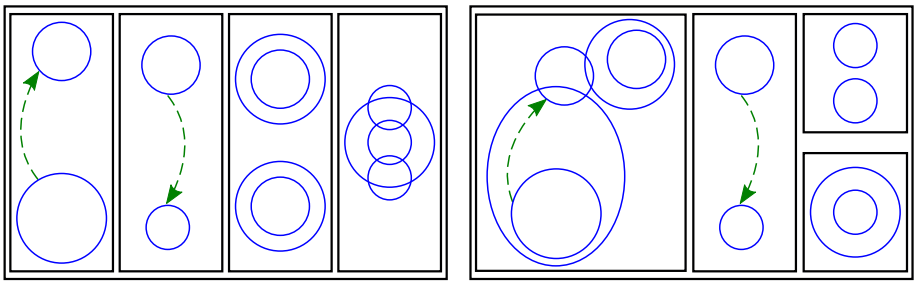


Fig. 7. Drawing choices for boxes

TABLE I
SCORES AND RANKINGS FOR DIAGRAMS IN TASK T_A.

Diagram Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
OCS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
AS1	37	38	43	45	50	53	45	47	52	55	59	62	41	42	46	51	54	57
AS2	33	33	38	38	40	40	43	43	48	48	53	53	37	37	43	43	44	44
CS	16	17	19	21	23	26	18	20	22	25	26	29	18	19	19	21	23	24
HS	15	15	16	16	17	17	19	19	20	20	21	21	17	17	18	18	19	19
OCS Ranking	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5
AS1 Ranking	1	2	5	6.5	10	13	6.5	9	12	15	17	18	3	4	8	11	14	16
AS2 Ranking	1.5	1.5	5.5	5.5	7.5	7.5	10.5	10.5	15.5	15.5	17.5	17.5	3.5	3.5	10.5	10.5	13.5	13.5
CS Ranking	1	2	6	9.5	12.5	16.5	3.5	8	11	15	16.5	18	3.5	6	6	9.5	12.5	14
HS Ranking	1.5	1.5	3.5	3.5	6.5	6.5	12.5	12.5	15.5	15.5	17.5	17.5	6.5	6.5	9.5	9.5	12.5	12.5

TABLE II
RANKINGS FOR DIAGRAMS IN TASKS T_B, T_C AND T_D.

T_B **Diagram Number**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
OCS ranking	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5
AS1 ranking	10	12	3	15	6	15	1	11	4	18	5	15	8	17	2	13	7	9
AS2 ranking	11	14	3	17.5	6	17.5	1	10	4	13	5	12	8	16	2	15	7	9
CS ranking	9	13	2.5	14.5	4	14.5	1	11	6.5	16.5	6.5	16.5	6.5	12	2.5	18	6.5	10
HS ranking	11	14.5	4	17	6.5	14.5	1	10	4	14.5	4	12	8.5	18	2	14.5	6.5	8.5
T_C **Diagram Number**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
OCS Ranking	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5
AS1 Ranking	8	16.5	2	10	4.5	14	9	15	3	12	7	13	11	18	1	4.5	6	16.5
AS2 Ranking	7.5	16	3	11	4	14	9	15	2	10	5	13	12	18	1	6	7.5	17
CS Ranking	7	17	2	10.5	4	14	8	16	3	10.5	5.5	13	9	18	1	12	5.5	15
HS Ranking	8	16	2.5	11	4	14	9	15	2.5	11	6	13	11	18	1	6	6	17
T_D **Diagram Number**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
OCS Ranking	8.5	8.5	8.5	17.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	17.5	8.5	8.5	8.5
AS1 Ranking	1	3	2	9	7	13	10	5.5	14	8	15	12	4	5.5	16	18	17	11
AS2 Ranking	1	2.5	2.5	8	7	10	12	4	14	9	15	11	6	5	16	18	17	13
CS Ranking	1	5	2	10.5	5	13	12	3	14	7.5	15.5	9	5	7.5	15.5	17	18	10.5
HS Ranking	1	3	2	9	4	10.5	14	5.5	13	7.5	15	10.5	7.5	5.5	16	18	17	12

experiment. The participants were informed that they could withdraw at any time. However, all the participants completed the experiment. Each participant was paid £6, in the form of a canteen voucher, to take part.

Each of the four tasks were conducted consecutively in a random order by the participants. Moreover, for each task, the printed diagrams were given to the participants in a random order, one at a time. The participants were asked to order the diagrams on the table from left to right where those on the left are least cluttered and those on the right are the most cluttered. If the participants thought that any diagrams were equally cluttered, then they were told that they could place them at the same point. They were also told that there was no limit on how many diagrams can be placed at the same point. The participants were given the opportunity to reorder the diagrams on the table at any time during the task. The participants were not given any help with their decisions during any of the tasks.

VI. RESULTS AND ANALYSIS

The goal of our work is to compare participants' rankings with our proposed measures. In order to do that, we first had to combine the 69 individual rankings for each task into a single ranking. A Friedman test was used for this purpose, giving estimated median ranks, from which we calculated the participants' rank. For each task, the Friedman test had a p-value of 0.000, indicated that further analysis using the participants' ranking was justified. The participants' ranks for

each task are given in table III, with the estimated median ranks given for task T_A only.

Using these participants' rankings for each task, we performed a Pearson correlation test against our proposed measures. This approach is consistent with that used by others studying visual complexity, for example [15], [19]. A higher number indicates a better (positive) correlation between two rankings. Table IV shows the correlation coefficients for each proposed measure with the participants' ranking by task. The figures in brackets are the p-values for the test. The scatter plot of measure rankings against participants' ranking for T_A can be seen in Figure 8 (without OCS).

We observe that OCS is a very poor measure for clutter in a diagram. The asterisks in Table IV in the OCS column indicate that there was no evidence of any relationship between OCS rankings and the participants' rankings. This is not surprising: if data can be presented in a number of ways, it is highly likely that the representation will affect how clutter the data appears. Since this measure was derived from the context of ontologies, we can conclude that one cannot look solely at an ontology's complexity to determine the visual complexity of the concept diagram used to represent it. All further sub-analysis then does not involve OCS.

All other measures showed correlation with participants' perceived clutter. However, they all exhibited different strengths of correlation. The measure AS1 showed the strongest correlation in two tasks, T_B and T_D, whereas the measure CS showed the strongest correlation in tasks T_A and T_C. Using the Fisher r-z transformation, we can test for significant differences between the correlation coefficients. For task T_A, pairwise tests indicate that the ranking coefficients for CS and AS1 are significantly different to those for AS2 and HS (CS vs. AS2: $z = 4.92$, $p = 0.0000$, CS vs. HS: $z = 6.39$, $p = 0.0000$, AS1 vs. AS2: $z = 3.56$, $p = 0.0002$, AS1 vs. HS: $z = 5.03$, $p = 0.0000$), whereas there is no difference between AS1 and CS ($z = 1.36$, $p = 0.0869$) and AS2 and HS ($z = 1.47$, $p = 0.0708$). For T_B, pairwise tests revelated no significant differences between any of the rankings. For T_C, the correlation coefficient for CS was significantly higher than that of all other rankings, with AS2 being the closest (vs. AS2: $z = 2.95$, $p = 0.0016$). For T_D, there was no significant differences between any of the correlation coefficients. Thus, we can conclude that CS is the best measure: it is never worse than any other measure, and in certain situations is significantly better.

That all measures (except OCS) are correlated with participant rankings suggests that the key component for clutter in a concept diagram is the clutter in the underlying Euler diagram. Three of the measures (AS1, AS2, HS) all calculate the curve-based clutter in the same way, the only difference being in how boxes are scored. The remaining measure (CS) counts the intersections between objects in a diagram, the majority of which are intersections between curves. The more regions an Euler diagram has, the greater the number of intersections between curves. The notable exception to this is with nested curves (the tunnels of [15]). CS and AS1 score arrows in similar ways: the former counts the number of curves an arrow crosses between source and target, whereas the latter counts how deeply nested the end-points are. In certain circumstances, these may give similar results. For example, a deeply nested source connected to a deeply nested target in another box will cross several curves. However, two deeply nested curves could be connected by an arrow which crosses no other curves. In this situation, the contribution to the CS by the arrow would be 1, whereas the contribution to AS1 would be larger. Since CS outperforms AS1, we can conclude that the number of crossing points is the best predictor for clutter.

We also noted the time taken for participants to perform each task. The averages are shown in Table V, where the time is given in seconds. The data were transformed using logarithms to ensure normality. An ANOVA test was conducted on the log of time taken by task ($F_{(3,68)} = 6.56, p = 0.000$), and pairwise Tukey tests at the 99% significance level gives a difference between tasks T_B and T_C. In other words, participants were quicker ranking diagrams which differed in the number of curves, than they were in ranking diagrams which differed by the number of boxes. This observation supports our conclusion that the underlying Euler diagram is the largest component of clutter score: when the underlying Euler diagram changed, participants could quickly rank the diagrams. However, when other syntax varied, greater cognitive effort was required to rank diagrams by clutter.

VII. Conclusions and Future Work

We investigated the perceived clutter in concept diagrams. By producing five plausible measures that could explain clutter in a diagram, and evaluating each of these measures against participants' rankings of diagrams, we found that a measure based on the number of line crossings in a diagram was the most effective predictor of clutter. Moreover, the results suggested that the largest part of clutter in a concept diagram comes from the underlying Euler diagram. Our results are also consistent with the findings of [15], in that the measures AS1, AS2 and HS, all based on the clutter scores for Euler diagrams, are all correlated with clutter in concept diagrams. Moreover, since the measure CS counts the number of points of interest in a diagram, our findings are consistent with [20], where a study on clutter in linear diagrams found that the number of lines in the plane was the best predictor of clutter. We can also give a recommendation for arrows in concept diagrams to reduce clutter: have the arrow cross as few curves as possible.

This work can be taken in two future directions. Firstly, the effect of clutter on task performance using concept diagrams can be investigated. In [21], it was conjectured that more cluttered concept diagrams negatively affected identifying empty classes in a set of ontology axioms. The study reported there was not directly investigating clutter, however, and was only focused on one particular task. To validate or reject the conjecture of [21], then, additional studies need to be undertaken.

Secondly, there is now a growing corpus of research on clutter in various visual languages. In this paper, we have seen

TABLE III
PARTICIPANT RANKINGS FOR ALL TASKS.

Diagram Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
T_A **Est. med. rank**	3.653	4.986	7.347	8.181	13.958	14.181	6.181	5.819	10.181	13.042	14.292	14.569	5.014	5.514	8.458	10.625	12.736	13.514
T_A **ranking**	1	2	7	8	15	16	6	5	10	13	17	18	3	4	9	11	12	14
T_B **ranking**	11	10	3	17	5	15	1	9	2	16	4	18	14	13	7	12	8	6
T_C **ranking**	8	17	4	11	7	15	5	16	2	10	3	13	9	18	1	12	6	14
T_D **ranking**	1	7	3	9	5	13	12	6	14	8	15	10	2	4	18	16	17	11

TABLE IV
CORRELATIONS BETWEEN CLUTTER MEASURES AND PERCEPTION, BY TASK.

	OCS	AS1	AS2	CS	HS
T_A	0 (*)	0.924 (0.000)	0.760 (0.000)	0.952 (0.000)	0.629 (0.005)
T_B	0 (*)	0.876 (0.000)	0.831 (0.000)	0.807 (0.000)	0.831 (0.000)
T_C	0 (*)	0.879 (0.000)	0.908 (0.000)	0.966 (0.000)	0.907 (0.000)
T_D	0.273 (0.274)	0.955 (0.000)	0.928 (0.000)	0.951 (0.000)	0.926 (0.000)

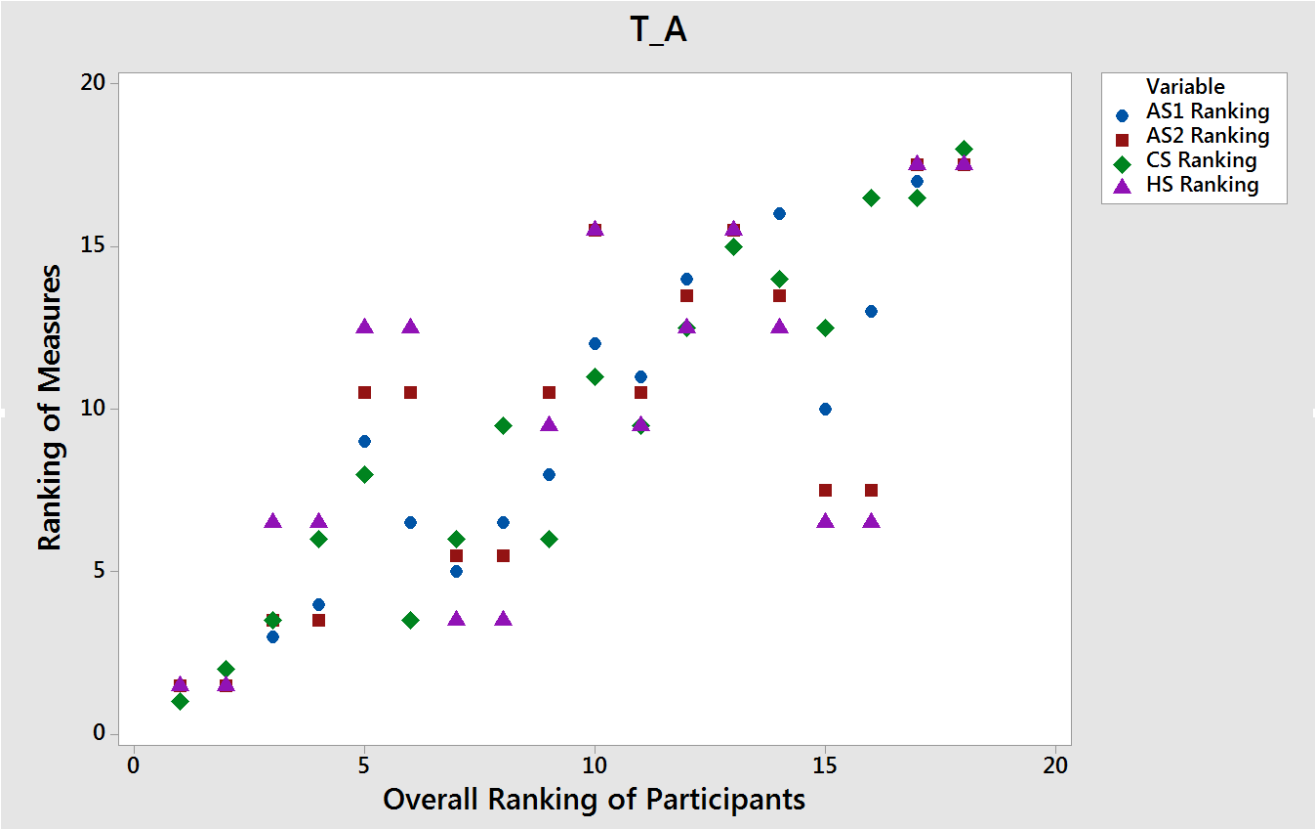


Fig. 8. Scatter plot for T_A

TABLE V
AVERAGE PARTICIPANT TIME, BY TASK.

T_A	T_B	T_C	T_D
204.0	215.1	168.3	185.3

that the concrete syntax score of a diagram is consistent with perceived clutter in Euler diagrams, concept diagrams, and linear diagrams. We conjectured that the number of points in a diagram (either objects or interactions between objects) was what caused clutter. It would be relatively straightforward to test this conjecture by using an eye-tracking experiment. Such an experiment forms part of our future research directions.

Other aspects of diagrams could also cause them to appear cluttered. The relative proportion of white-space in a diagram, as well as the placement of objects relative to each other, could both affect perceived clutter. Whilst this paper is an important step in providing a unifying framework for clutter, more work is needed to be able to develop a general theory of clutter.

ACKNOWLEDGMENT

The first and second authors were supported by the EPSRC grant "Visual Justifications for Ontologies" [EP/M016323/1].

REFERENCES

[1] K. Koffka, *Principles of Gestalt Pschology*. Lund Humphries, 1935.

[2] D. Moody, "The "Physics" of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering," *Software Engineering, IEEE Transactions on*, vol. 35, no. 6, pp. 756–779, 2009.

[3] G. A. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information." *Psychological review*, vol. 63, no. 2, p. 81, 1956.

[4] O. Bodenreider and R. Stevens, "Bio-ontologies: current trends and future directions," *Briefings in bioinformatics*, vol. 7, no. 3, pp. 256–274, 2006.

[5] J. Hage and B. Verheij, "The law as a dynamic interconnected system of states of affairs: a legal top ontology," *International Journal of Human-Computer Studies*, vol. 51, no. 6, pp. 1043–1077, 1999.

[6] N. Guarino and C. Welty, "Evaluating ontological decisions with ontoclean," *Communications of the ACM*, vol. 45, no. 2, pp. 61–65, 2002.

[7] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis, and E. Giannopoulou, "Ontology Visualization Methods - a Survey," *ACM Computing Surveys (CSUR)*, vol. 39, no. 4, p. 10, 2007.

[8] G. Stapleton, J. Howse, K. Taylor, A. Delaney, J. Burton, and P. Chapman, "Towards diagrammatic ontology patterns," in *4th Workshop on Ontology and Semantic Web Patterns*, ser. CEUR Workshop Proceedings, vol. 1188, 2013.

[9] J. Howse, G. Stapleton, K. Taylor, and P. Chapman, "Visualizing Ontologies: A Case Study," in *International Semantic Web Conference*, ser. LNCS, no. 7031. Springer, 2011, pp. 257–272.

[10] F. Baader, D. Calvanese, D. McGuinness, D. Nadi, and P. P.-S. (eds), *The Description Logic Handbook*. CUP, 2003.

[11] U. S. F. Baader, I. Horrocks, "Description logics as ontology languages for the semantic web," in *Mechanizing Mathematical Reasoning*. Springer, 2005, pp. 228–248.

[12] R. Rosenholtz, Y. Li, J. Mansfield, and Z. Jin, "Feature congestion: a measure of display clutter," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005, pp. 761–770.

[13] J. G. Snodgrass and M. Vanderwart, "A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity." *Journal of experimental psychology: Human learning and memory*, vol. 6, no. 2, p. 174, 1980.

[14] A. N. Tuch, J. A. Bargas-Avila, K. Opwis, and F. H. Wilhelm, "Visual complexity of websites: Effects on users experience, physiology, performance, and memory," *International Journal of Human-Computer Studies*, vol. 67, no. 9, pp. 703–715, 2009.

[15] C. John, A. Fish, J. Howse, and J. Taylor, "Exploring the notion of clutter in Euler diagrams," in *4th International Conference on the Theory and Application of Diagrams*. Stanford, USA: Springer, 2006, pp. 267–282.

[16] M. Alqadah, G. Stapleton, J. Howse, and P. Chapman, "Evaluating the Impact of Clutter in Euler Diagrams," in *Diagrammatic Representation and Inference*, ser. LNAI. Springer, 2014, vol. 8578, pp. 108–122.

[17] H. Zhang, Y.-F. Li, and H. B. K. Tan, "Measuring design complexity of semantic web ontologies," *Journal of Systems and Software*, vol. 83, no. 5, pp. 803–814, 2010.

[18] E. J. Weyuker, "Evaluating software complexity measures," *Software Engineering, IEEE Transactions on*, vol. 14, no. 9, pp. 1357–1365, 1988.

[19] H. Purchase, E. Freeman, and J. Hamer, "An Exploration of Visual Complexity," in *Diagrammatic Representation and Inference*. Springer, 2012, pp. 200–213.

[20] M. Alqadah, G. Stapleton, J. Howse, and P. Chapman, "The Perception of Clutter in Linear Diagrams," accepted for presentation at Diagrams 2016.

[21] T. Hou, P. Chapman, and A. Blake, "Antipattern Comprehension: An Empirical Evaluation," submitted to Formal Ontology of Information Structures 2016.