Exploring the Notion of 'Clutter' in Euler Diagrams

Chris John*, Andrew Fish, John Howse, and John Taylor

Visual Modelling Group,
School of Computing, Mathematical and Information Sciences,
University of Brighton, Brighton, UK
{C.John,Andrew.Fish,John.Howse,John.Taylor}@brighton.ac.uk
www.cmis.brighton.ac.uk/research/vmg/

Abstract. Euler diagrams are an effective and intuitive way of representing relationships between sets. As the number of sets represented grows, Euler diagrams can become 'cluttered' and lose some of their intuitive appeal. In this paper we consider various measures of 'clutter' for abstract Euler diagrams and show that they compare well with results obtained from an empirical study. We also show that all abstract Euler diagrams can be constructed inductively by inserting a contour at a time and we relate this inductive description to the clutter metrics.

1 Introduction

Euler diagrams consist of contours (closed curves) which divide the plane into zones; they provide an effective and intuitive way of representing relationships between sets. Euler diagrams have numerous applications, including the visualization of statistical data, displaying the results of database queries, representing non-hierarchical computer file systems and for viewing clusters which contain concepts from multiple ontologies; closed curves are a basis for many of the UML notations, including class diagrams and statecharts. However, as the number of contours increases, Euler diagrams tend to become 'cluttered' and lose some of their appeal and effectiveness. An exploration of factors affecting clutter in Euler diagrams was given in [7] and a clutter metric was proposed.

Various additions to the basic Euler diagram notation have been proposed which allow alternative representations of set relationships; for example, the use of shading to denote empty sets [6, 11] and the use of projections [4]. In [7], these are exploited to produce a 'clutter reduction algorithm' for these (extended) Euler diagrams. This can produce semantically equivalent Euler diagrams that are considerably less cluttered.

The clutter measure used in [7] appears to provide an effective measure of clutter. However, there are alternative measures that could be used and intuition varies between individuals on the degree of clutter of various diagrams. In this paper, we present the results of an empirical study designed to test subjects' notion of clutter and consider comparisons between various clutter measures.

We also explore some of the properties of the measures in relation to diagrams comprising more than one connected component, so-called 'nested diagrams' [2] (see section 2.3). We show that any abstract Euler diagram can be constructed inductively by adding a contour at a time. This leads to a deeper understanding of the behaviour of some clutter measures for nested diagrams.

The notion of clutter is an issue in many disciplines both within the area of visual display and beyond. We subscribe to the view that reducing clutter increases the usability of diagrams [10]:

> *If we can better understand clutter, we can create tools to automatically identify it, and even eventually create systems that will advise a designer when the level of clutter is too high, and suggest techniques for reducing visual clutter.*

Metrics for measuring clutter are often specific to a particular domain. For example, [12] measures clutter in aviation displays, [8, 9] for air traffic control and [1] for scatter plots. Our measures are no exception. However, we envisage that aspects of this work will be adaptable to applications further afield; in particular, the notion of separating clutter at the concrete and abstract levels.

In section 2 we review the (concrete and abstract) syntax of Euler diagrams and the notion of 'nesting'. The clutter measure in [7] is introduced in section 3. We then present the results of the empirical study which leads to the definition of two new clutter metrics. In section 4 we show that all Euler diagrams can be built inductively a contour at a time and we relate this to our clutter metrics. We also show how to calculate the clutter measures of nested diagrams. Finally, we consider some directions for further work.

2 Euler Diagram Syntax

Following [5], we define two layers of Euler diagram syntax: an *abstract* or *type* syntax and a *concrete* or *token* syntax. The concrete syntax describes drawn diagrams (on paper or a monitor, for example) and the abstract syntax describes the underlying structure of diagrams.

2.1 Concrete Euler diagrams

A *concrete Euler diagram* is a subset of the plane bounded by a *boundary rectangle* and properly containing a number of *contours*, each with a distinct label. There are various well-formedness conditions which are usually imposed to ensure diagrams can be interpreted without ambiguity. We require that contours meet transversely (cross one another at a point of intersection) and the diagram has 'connected zones'. Examples of non-well-formed Euler diagrams are given in figure 1. The diagram d_1 contains two contours B and C that do not intersect transversely (they just touch at their point in common) whereas the diagram d_2 has a disconnected zone (shaded).
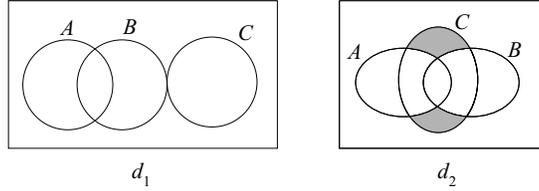
Fig. 1. Non-well-formed Euler diagrams

Definition 1. *A (well-formed) concrete Euler diagram is a quadruple $\hat{d} = \langle \hat{L}(\hat{d}), \hat{C}(\hat{d}), \hat{Z}(\hat{d}), \hat{\beta} \rangle$ defined as follows.*

*The diagram has a **bounding rectangle**, $\hat{\beta}$, whose interior contains a finite set, $\hat{C}(\hat{d})$, of **contours** (simple closed curves). Each contour has a unique label from the set $\hat{L}(\hat{d})$. Contours meet transversely and at most two contours meet at any one point.*

*The contours divide the interior of $\hat{\beta}$ into connected 'pieces' called **zones**; the set of zones is $\hat{Z}(\hat{d})$. Each zone $\hat{z}$ is uniquely identified by a subset $\hat{C}(\hat{z})$ of the contours: it comprises the region of the plane interior to the contours in the subset $\hat{C}(\hat{z})$ and exterior to the contours not in the subset.*

The last part of the definition of 'zone' prevent disconnected zones exhibited in diagram d_2 of figure 1. This also ensures that each zone is described uniquely by a pair of sets $(\hat{C}(\hat{z}), \hat{C}(\hat{d}) - \hat{C}(\hat{z}))$ that partition $\hat{C}(\hat{d})$.

Example 1. The concrete Euler diagram $\hat{d}$ represented in figure 2 has three contours a, b, c with labels A, B, C respectively. There are six zones which are described by the following pairs of sets that each partition $\hat{C}(\hat{d}) = \{a, b, c\}$:

$$(\{a\}, \{b, c\}), (\{a, c\}, \{b\}), (\{a, b, c\}, \varnothing), (\{b, c\}, \{a\}), (\{c\}, \{a, b\}), (\varnothing, \{a, b, c\}).$$
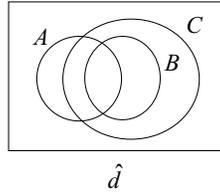
Fig. 2. An Euler diagram

2.2 Abstract Euler diagrams

An abstract Euler diagram encapsulates the *structure* in a concrete Euler diagram but discards some topological details. It includes information about the way the contours interact (as specified by the zones) but 'loses' information about the shape of the contours, for example. To describe abstractly the diagrammatic elements in a concrete diagram, we need to specify its contours and zones. Since only the labels of contours are needed to specify a zone, we can identify contours and labels.

Definition 2. *Let L be a set of contour labels. A **zone** (on L) is a pair (x, y) where $x \subseteq L$ and $y = L - x$. The zone (x, y) is **inside** the contour labels in x and **outside** the contour labels in y. An **abstract Euler diagram** d with labels L is a pair $d = \langle L, Z \rangle = \langle L(d), Z(d) \rangle$ where $Z(d)$ is a non-empty set of zones satisfying the following condition.*

(*) *There is a zone inside every contour label: $\forall\, l \in L(d)\ \exists\, (x, y) \in Z(d) : l \in x$.*

It is common (see [2, 6] for example) to require the set of zones to satisfy the following condition in addition to (*).

(**) *There is a zone outside all the contour labels: $(\varnothing, L(d)) \in Z(d)$.*

This condition is usually included as a 'drawability condition' since any diagram that fails to satisfy (**) is not well-formed. At the abstract level, however, the condition appears somewhat arbitrary. For example, the diagram $d = \langle \{A, B\}, \{(\{A\}, \{B\}), (\{B\}, \{A\})\} \rangle$ does not satisfy (**) but there is no particular reason why it should not be allowed as an abstract diagram.

2.3 Nesting in Euler diagrams

The notion of *nesting* of Euler diagrams is defined in [2]. An Euler diagram d is *nested* if it can be constructed by embedding one diagram d_2 into a zone z^* of another diagram d_1. An example is shown in figure 3 where d_2, with contours labelled P and Q, is embedded in the zone $(\{A, C\}, \{B\})$ of the diagram d_1, with contours A, B and C.

d
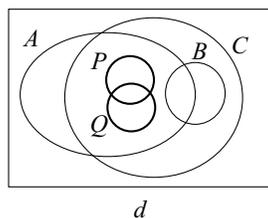
Fig. 3. A nested Euler diagram

Definition 3. *An abstract Euler diagram $d = \langle L, Z \rangle$ is **nested** if there exist Euler diagrams $d_1 = \langle L_1, Z_1 \rangle$ and $d_2 = \langle L_2, Z_2 \rangle$ and a zone $z^* = (x^*, y^*) \in Z_1$ satisfying the following conditions.*

1. *$\{L_1, L_2\}$ is a partition of L.*
2. *$Z(d) = Z_1^* \cup Z_2^*$ where $Z_1^* = \{(x, y \cup L_2) : (x, y) \in Z_1\}$ and $Z_2^* = \{(x^* \cup x, y^* \cup y) : (x, y) \in Z_2\}$.*
3. *$Z_1^* \cap Z_2^* = \{(x^*, y^* \cup L_2)\}$.*

*We say that d_2 is **embedded** in the zone z^* of d_1, and write $d = d_2 \xrightarrow{z^*} d_1$. A diagram that is not nested is called **atomic**.*

It will be convenient to define the **depth** of z^*, $depth(z^*)$, to be the number of contours enclosing it: $depth(z^*) = |x^*|$. The **components** of a diagram d are its atomic 'pieces' (obtained by successively decomposing as $d_2 \xrightarrow{z^*} d_1$).

Example 2. We shall describe the nested abstract Euler diagram illustrated in figure 3. Let d_1 and d_2 be the diagrams defined as follows:

$d_1 = \langle L_1, Z_1 \rangle$ where $L_1 = \{A, B, C\}$ and

$$Z_1 = \{(\{A\}, \{B, C\}), (\{A, C\}, \{B\}), (\{A, B, C\}, \varnothing),$$
$$(\{B, C\}, \{A\}), (\{C\}, \{A, B\}), (\varnothing, \{A, B, C\})\};$$

$d_2 = \langle L_2, Z_2 \rangle$ where $L_2 = \{P, Q\}$ and

$$Z_2 = \{(\{P\}, \{Q\}), (\{P, Q\}, \varnothing), (\{Q\}, \{P\}), (\varnothing, \{P, Q\})\}.$$

The diagram d is formed by embedding the diagram d_2 into the zone $(\{A, C\}, \{B\})$ of d_1, $d = d_2 \xrightarrow{(\{A,C\},\{B\})} d_1$.

3 Measuring clutter

Several factors contributing to the clutter of an Euler diagram were identified in [7]. These include the number of zones, the number of contours, the ratio of zones to contours within a component (connected 'piece' of the diagram), and the degree of 'intersection' or 'disjointness' between pairs of contours – diagrams where most pairs of contours intersect tend to appear more cluttered than those where most pairs are disjoint. Various measures of clutter were considered in [7]. A preferred measure, called *contour scoring*, was proposed as a sufficiently sensitive measure of clutter which is straightforward to calculate. In this section, we consider various 'clutter measures' and some of their properties and compare how these measures perform relative to data obtained from an empirical study.
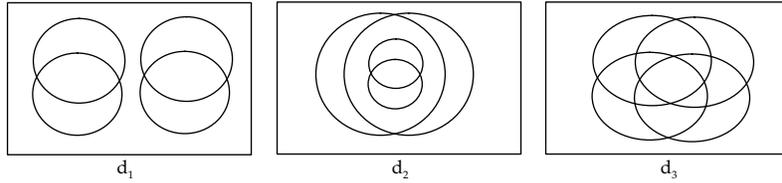
3.1 Contour scoring

In a concrete Euler diagram, the contour score is calculated simply by counting the number of zones within each contour and summing. In an abstract Euler diagram, the *contour score* is the sum over the zones of the cardinalities of containing sets:

$$\mathcal{CS}(d) = \sum_{(x,y) \in Z(d)} |x|.$$

The diagrams d_1, d_2 and d_3 in figure 4 have contour scores of 8, 14 and 28 respectively. Contour scoring is simple to compute and it is able to differentiate between diagrams where some other measures fail. In figure 4, for example, $\mathcal{CS}(d_1) \neq \mathcal{CS}(d_2)$, even though the number of zones (7), contours (4) and components (2) are the same.

In section 4 we use contour scoring to define bounds on the clutter in a diagram and we consider various patterns of diagrams and their contour scores. The diagram d_1 in figure 5 can be generalised into a pattern we call a 'chain',

Fig. 4. Illustrating 'contour scoring'.

displaying a high degree of 'disjointness' between contours; chains have a contour score which is a *linear* function of the number of contours. The diagram d_2 in figure 5 can be generalised into a pattern we call a 'crossed tunnel', displaying a higher degree of 'intersection' between contours; crossed tunnels have a contour score that is a *quadratic* function of the number of contours.
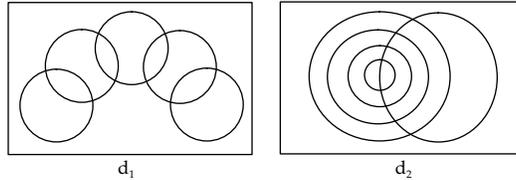

**Fig. 5.** The 'chain' and 'crossed tunnel' component patterns on 5 contours.

3.2 Empirical testing

To test whether contour scoring gives an accurate measure of clutter, an experiment was conducted to compare people's intuition with the calculated measures. Over 230 experimental subjects, ranging in age from 12 to 40, were asked to score various Euler diagrams on a scale of 1 to 30 purely on the basis of how 'cluttered' they appeared. To ensure an intuitive response, subjects were given an appropriately short time to complete the test. The experiment involved several sets of diagrams to test various properties. Here we show the results of two tests, each involving 16 diagrams. Four (appropriately scaled) measures, introduced in [7], were compared to the experimental data: $\mathcal{Z}ones$ (the number of zones), $\mathcal{Z}{:}\mathcal{C}$ (the ratio of zones to contours), $\mathcal{Z}\mathcal{S}$ (zone scoring, a measure calculated by summing the vertex degrees of the simple geometric dual graph) and $\mathcal{C}\mathcal{S}$ (contour scoring). The results for one of the sets of 16 diagrams are illustrated in figure 6; the results for the other set was similar. In each graph, the bold line represents the data (the mean score for each diagram) and the dashed line represents the measure.

Each measure appears quite accurate at measuring clutter scored by the subjects. However, closer inspection suggests that contour scoring 'follows' the empirical data most closely. In [7], it was argued that contour scoring was able to detect subtle variations in clutter the other measures were not. For example, figure 7 shows the diagrams in the experiment represented by the points 14 and 15 (last but two and last but one) on the graphs. The data shows that d_{15} is perceived as being more cluttered than d_{14}. Of the four measures, only $\mathcal{C}\mathcal{S}$ scores d_{15} greater than d_{14}; the other measures each score these diagrams equally.
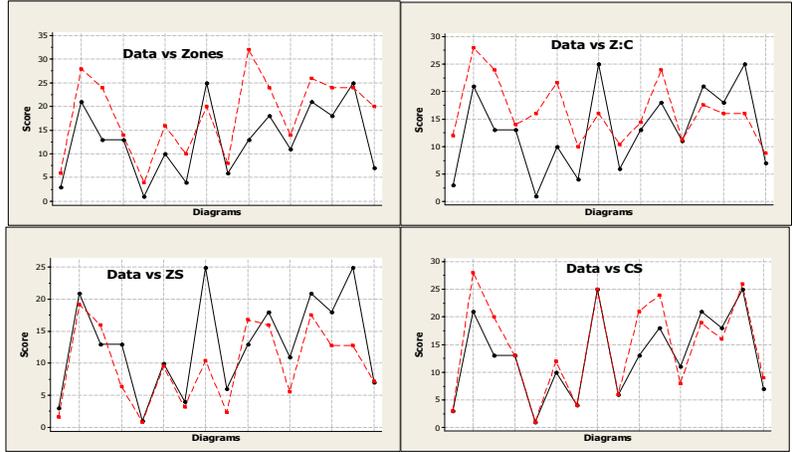
Fig. 6. Empirical data (in bold) compared to four measures (each dashed).
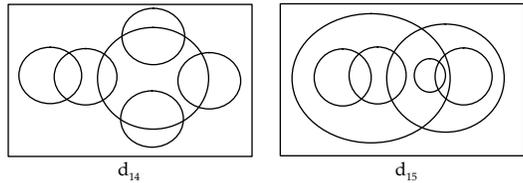

Fig. 7. Diagrams 14 and 15 from the study.

Rank correlations support contour scoring as the most accurate of the four measures tested. For both sets of 16 diagrams used in the experiment, the Pearson correlation between the data and $\mathcal{CS}$ was 0.889, compared to 0.721 for $\mathcal{Z}ones$, 0.549 for $\mathcal{ZS}$, and 0.451 for $\mathcal{Z}:\mathcal{C}$. These results are not intended as an in-depth statistical analysis, but they do provide evidence of the capability of various measures to score clutter in Euler diagrams. A thorough analysis of results from this large data set is intended for future publication.

3.3 Alternative measures

The empirical data lends support to contour scoring as an accurate and reliable measure of clutter in Euler diagrams but it can give rise to some potentially undesirable results. Consider the (atomic) diagrams d_1 and d_2 in figure 8, which have equal contour scores (given in the box in the diagram). Inserting a contour into each diagram surrounding the given component produces diagrams d_3 and d_4 with different contour scores. In other words, the contour score does not preserve equality under the operation of 'enclosing a diagram with a new contour'. The reason for this is the different numbers of zones in d_1 and d_2; see corollary 1 (section 4.2). It is natural to ask whether this undermines $\mathcal{CS}$ as an measure for clutter and whether we can adapt contour scoring to avoid this. We now consider two possible variations on the contour scoring measure, each of which agrees with $\mathcal{CS}(d)$ for atomic diagrams d.
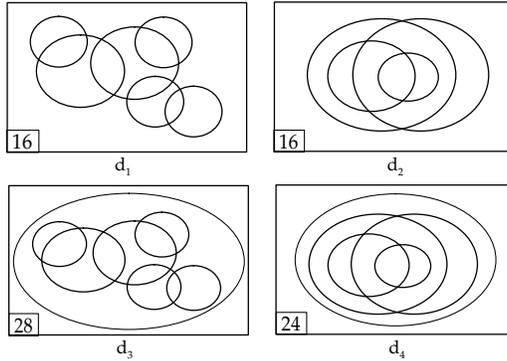
Fig. 8. Illustrating $\mathcal{CS}$ when enclosing a diagram with a new contour.

Simple component contour scoring, denoted $\mathcal{SCS}(d)$, ignores nesting and simply sums the contour scores of the components of d. In diagrams d_1 and d_2 in figure 4 the components of the diagrams are identical and hence have the same $\mathcal{SCS}$ score, even though the components are arranged differently. Simple component contour scoring consistently scores nested diagrams lower that $\mathcal{CS}$; see theorem 4. Although $\mathcal{SCS}$ clearly preserves equality under the operation of 'enclosing a diagram with a contour', the diagrams in figure 9 illustrates a weakness in $\mathcal{SCS}$. Here $\mathcal{SCS}(d_1)$ is quite a lot bigger than $\mathcal{SCS}(d_2)$ even though the two diagrams are visually very similar. This example suggests that $\mathcal{SCS}$ may not adequately represent the clutter in nested diagrams where one component is quite 'deeply embedded'.
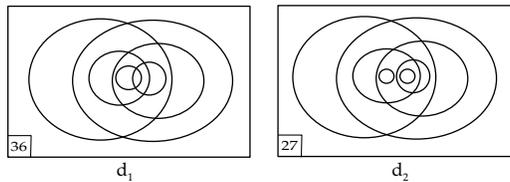
Fig. 9. Two similar diagrams scored differently under $\mathcal{SCS}$.

Nested component contour scoring, denoted $\mathcal{NCS}$, takes the sum of the components' contour scores and adds the 'nesting depth' of each component; it represents a natural compromise between $\mathcal{CS}$ and $\mathcal{SCS}$ by taking account of the depth of nesting at the component level. The diagram d_2 in figure 9 has three components with contour scores 25, 1 and 1. Their respective nesting depths are 0, 4 and 5, giving $\mathcal{NCS}(d_2) = 36$. Whilst is is straightforward to compute $\mathcal{NCS}$ for (simple) concrete diagrams, it is more difficult to formulate at the abstract level. Like $\mathcal{CS}$, $\mathcal{NCS}$ fails to preserve equality under the operation of 'enclosing a diagram with a contour' as figure 10 illustrates.

3.4 Empirical testing revisited

Since all three variations of contour scoring exhibit strengths and weaknesses, we consider the empirical study to investigate how accurately the measures cor-
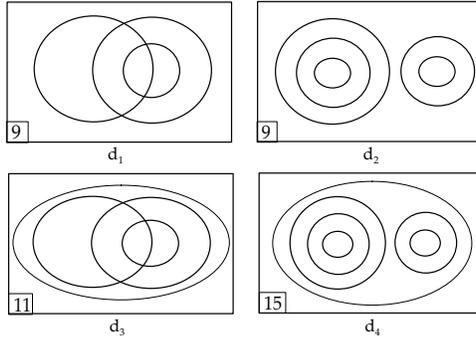
Fig. 10. Illustrating $\mathcal{NCS}$ when enclosing a diagram with a new contour.

respond to the experimental data. The study was carried out *before* these adaptations to contour scoring were considered. Since the measures differ only on nested diagrams, we consider how they compare on the 14 nested diagrams from the original 32 diagrams used in the study. The results are illustrated in figure 11, where again the experimental data is represented by bold lines and the measures by dashed lines.
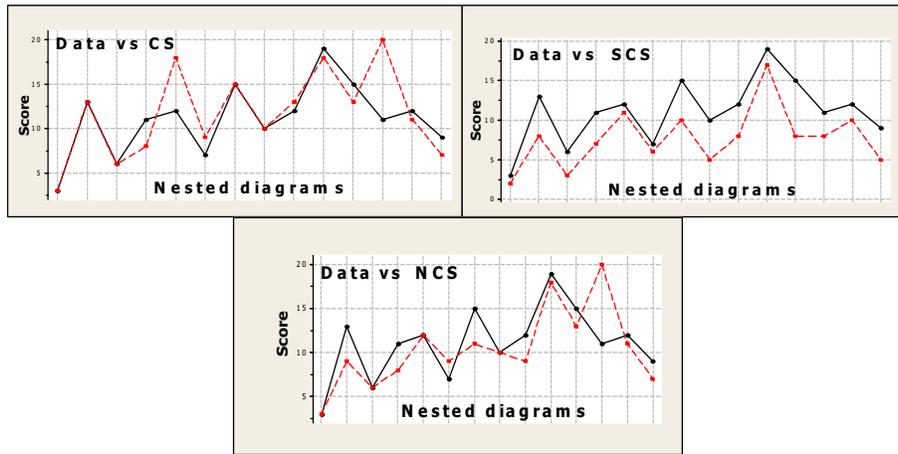

Fig. 11. Empirical data on nested diagrams compared to three measures.

All three measures perform well in relation to the data. The performance of simple contour scoring warrants a particular mention; although it consistently scores diagrams lower than the other measures, its rank correlation with the data is 0.893 (as compared to 0.761 for $\mathcal{CS}$ and 0.707 for $\mathcal{NCS}$). This is a surprisingly good match since $\mathcal{SCS}$ ignores *how* components are nested in diagrams. These values may suggest that $\mathcal{SCS}$ more accurately represents the data when scoring nested diagrams. However when all 32 (nested *and* atomic) diagrams are considered, $\mathcal{CS}$ has a slightly higher correlation with the data scores than $\mathcal{SCS}$ (0.889

and 0.857 respectively). As there is little difference in how well the measures represent the experimental data, which is adopted may depend on other factors (e.g., whether equality should be preserved under enclosing with a contour).

4 Generating Euler Diagrams Inductively

We investigate the effect certain diagrammatic transformations have on clutter scores. We define a syntactic rule which can be used to generate any Euler diagram inductively. This provides a framework to explore generic diagrammatic patterns and their clutter scores. We concentrate on $\mathcal{CS}$ but the process could be adapted to the other measures.

4.1 Inserting a contour

We describe how (abstract) Euler diagrams can be constructed inductively by inserting one contour at a time. To define the 'insert contour' rule, we need to specify how the inserted contour relates to each of the existing contours in the diagram. We begin by considering its (one-sided) inverse operation, 'remove contour', which is simpler to describe.

Rule 1 *Remove contour. Let $d = \langle L(d), Z(d) \rangle$ be an Euler diagram and let $l \in L(d)$. The diagram d with l deleted, $d' = d - l$, is defined by:*

1. $L(d') = L(d) - \{l\}$
2. $Z(d') = \{(x - \{l\}, y - \{l\}) : (x, y) \in Z(d)\}$. □

It is clear that $d - l$ satisfies condition (*) of definition 2 and so is a well-defined Euler diagram.

We now define an operation which adds a new contour label l' to an abstract Euler diagram d. Various systems based on Euler diagrams have included an 'add contour' rule which preserves the semantic content of the diagram (see, for example, [6]). The purpose of our 'insert contour' rule is different. We are not interested here in reasoning with Euler diagrams but rather in building (abstract) Euler diagrams, so our 'insert contour' rule allows a contour to be inserted into a diagram in *any* way that will create a well defined Euler diagram.

Consider inserting a contour into a concrete Euler diagram. For each existing zone z in the diagram, the inserted contour will either entirely enclose z, or will split z in two or will be disjoint from z (strictly, the *interior* of the inserted contour will be disjoint from z). For example, in figure 12, d' is obtained from d by inserting the contour labelled E. The inserted contour entirely encloses the existing zones $(\{B, C\}, \{A\})$ and $(\{C\}, \{A, B\})$, it splits in two the zones $(\{B\}, \{A, C\})$ and $(\varnothing, \{A, B, C\})$ and it is disjoint from the zones $(\{A\}, \{B, C\})$ and $(\{A, B\}, \{C\})$.

In an abstract diagram, to specify how a new contour label is inserted, we need to 'partition' the set of zones into three subsets:

$Z_1(d)$, the set of zones to be enclosed by the new contour label;
$Z_2(d)$, the set of zones to be split by the new contour label;
$Z_3(d)$, the set of zones that are to be 'disjoint' from the new contour label.
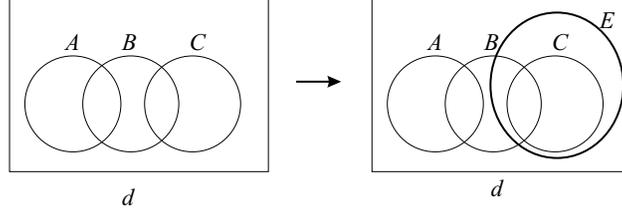
Fig. 12. Inserting a contour

In general, one or more of the sets $Z_1(d), Z_2(d), Z_3(d)$ may be empty so that, strictly, $\{Z_1(d), Z_2(d), Z_3(d)\}$ need not be a partition of $Z(d)$. However, to maintain condition (*) of definition 2, the new contour cannot be added entirely outside all the existing zones, so $Z_1(d) \cup Z_2(d) \neq \varnothing$.

In the example above, represented by figure 12, we have:

$$Z_1(d) = \{(\{B, C\}, \{A\}), (\{C\}, \{A, B\})\}$$
$$Z_2(d) = \{(\{B\}, \{A, C\}), (\varnothing, \{A, B, C\})\}$$
$$Z_3(d) = \{(\{A\}, \{B, C\}), (\{A, B\}, \{C\})\}.$$

We obtain the zones of d' as follows. Each zone (x, y) in $Z_1(d)$ gives rise to a single zone in d' by adding E to x, producing zones $(\{B, C, E\}, \{A\})$ and $(\{C, E\}, \{A, B\})$. Each zone (x, y) in $Z_2(d)$ gives rise to a pair of zones in d' by adding E to x and to y separately; this produces zones $(\{B, E\}, \{A, C\})$, $(\{B\}, \{A, C, E\})$ and $(\{E\}, \{A, B, C\})$, $(\varnothing, \{A, B, C, E\})$. Finally, each zone (x, y) in $Z_3(d)$ gives rise to a single zone in d' by adding E to y, producing zones $(\{A\}, \{B, C, E\})$ and $(\{A, B\}, \{C, E\})$.

Rule 2 *Insert contour.*
Let d be a unitary Euler diagram and let $l \notin L(d)$ be a new contour label.
Let $\{Z_1(d), Z_2(d), Z_3(d)\}$ be a 'partition' of $Z(d)$ – one or more of the sets may be empty but $Z_1(d) \cup Z_2(d) \neq \varnothing$. The diagram d with l inserted relative to the partition $\{Z_1(d), Z_2(d), Z_3(d)\}$ is d' where $L(d') = L(d) \cup \{l\}$ and the zones are defined as follows.

1. *$(x, y) \in Z_1(d) \Rightarrow (x \cup \{l\}, y) \in Z(d')$.*
2. *$(x, y) \in Z_2(d) \Rightarrow (x \cup \{l\}, y), (x, y \cup \{l\}) \in Z(d')$.*
3. *$(x, y) \in Z_3(d) \Rightarrow (x, y \cup \{l\}) \in Z(d')$.* □

Lemma 1. *Suppose the Euler diagram d_2 is obtained from the Euler diagram d_1 by removing contour label l; that is, $d_2 = d_1 - l$. Then d_1 can be obtained from d_2 by inserting l.*

Proof
The zones of d_1 can be partitioned into three sets:

$$Z_1(d_1) = \{(x, y) \in Z(d_1) : l \in x \text{ and } (x - \{l\}, y \cup \{l\}) \notin Z(d_1)\};$$
$$Z_2(d_1) = \{(x \cup \{l\}, y) \in Z(d_1) : l \notin x \cup y\} \cup \{(x, y \cup \{l\}) \in Z(d_1) : l \notin x \cup y\};$$
$$Z_3(d_1) = \{(x, y) \in Z(d_1) : l \in y \text{ and } (x \cup \{l\}, y - \{l\}) \notin Z(d_1)\}.$$

Removing l from d_1 gives the corresponding 'partition' of $Z(d_2)$:

$$Z_1(d_2) = \{(x - \{l\}, y) : (x, y) \in Z_1(d_1)\};$$
$$Z_2(d_2) = \{(x, y) : (x \cup \{l\}, y), (x, y \cup \{l\}) \in Z_2(d_1)\};$$
$$Z_3(d_2) = \{(x, y - \{l\}) : (x, y) \in Z_3(d_1)\}.$$

It is now clear that inserting the contour label l in d_2 relative to this 'partition' $\{Z_1(d_2), Z_2(d_2), Z_3(d_2)\}$ recreates the Euler diagram d_1. □

The following theorem follows by induction on the number of contour labels.

Theorem 1. *All abstract Euler diagrams can be generated from the trivial diagram $\langle \varnothing, \{(\varnothing, \varnothing)\}\rangle$ by applying rule 2.* □

Example 3. Let T_n denote the 'crossed tunnel' Euler diagram with n contours ($n \geq 2$); T_5 is shown as diagram d_2 in figure 5. To be precise, let $L(T_n) = \{l^*, l_1, l_2, \ldots, l_{n-1}\}$ where l^* is the label of the contour that intersects all the other contours and suppose that $l_1, l_2, \ldots, l_{n-1}$ label the other contours from 'outside in' so that l_1 contains l_2 which contains l_3 etc. To construct T_{n+1} from T_n, we add the contour label l_n relative to the following partition of $Z(T_n)$:

$$Z_1(T_n) = \varnothing;$$
$$Z_2(T_n) = \{(\{l_1, l_2, \ldots, l_{n-1}\}, \{l^*\}), (\{l^*, l_1, l_2, \ldots, l_{n-1}\}, \varnothing)\};$$
$$Z_3(T_n) = Z(T_n) - Z_2(T_n).$$

4.2 Using 'insert contour' to calculate contour score

Suppose that the Euler diagram d' is obtained from d by inserting the contour label l relative to the 'partition' $\{Z_1(d), Z_2(d), Z_3(d)\}$ of $Z(d)$. We aim to relate the contour score of d, $\mathcal{CS}(d)$ to that for d', $\mathcal{CS}(d')$. For $i = 1, 2, 3$, let $\mathcal{CS}_i(d)$ denote the contribution to the contour score arising from $Z_i(d)$:

$$\mathcal{CS}_i(d) = \sum_{(x,y) \in Z_i(d)} |x|.$$

Then $\mathcal{CS}(d) = \mathcal{CS}_1 + \mathcal{CS}_2 + \mathcal{CS}_3$.

The contour label l is added to the containing set x of each zone (x, y) in $Z_1(d)$; since there are $|Z_1(d)|$ such zones, $\mathcal{CS}_1(d') = \mathcal{CS}_1(d) + |Z_1(d)|$.

Each zone (x, y) in $Z_2(d)$ gives rise to two zones in d', one of which has the additional contour l added to its containing set x; since there are $|Z_2(d)|$ such zones, $\mathcal{CS}_2(d') = 2\mathcal{CS}_2(d) + |Z_2(d)|$.

Finally, each zone (x, y) in $Z_3(d)$ gives rise to a single zone in d' with the same containing set x, so $\mathcal{CS}_3(d') = \mathcal{CS}_3(d)$. Combining these equations gives the following theorem.
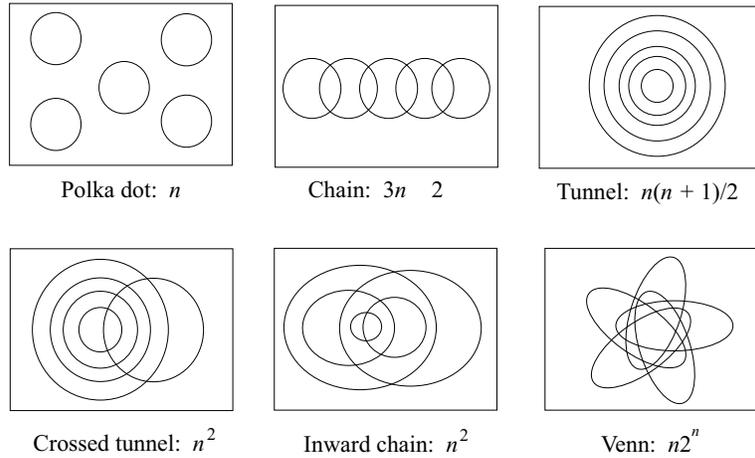
Theorem 2. *Let d' be obtained from d by inserting the contour label l relative to the 'partition' $\{Z_1(d), Z_2(d), Z_3(d)\}$ of $Z(d)$. With the notation above,*

$$\mathcal{CS}(d') = \mathcal{CS}(d) + \mathcal{CS}_2(d) + |Z_1(d)| + |Z_2(d)|.$$

□

Corollary 1. *Let d be an abstract Euler diagram and let d' be obtained from d by 'enclosing' it with a new contour (see section 3.3). Then $\mathcal{CS}(d') = \mathcal{CS}(d) + |Z(d)|$.*
□

Figure 13 shows various patterns of Euler diagram that can be generated by inserting one contour at a time. The figure shows a representative diagram for each pattern with 5 contours, together with the contour score for a diagram with n contours. In each case, the formula for $\mathcal{CS}(d)$ can be proved by a simple induction argument using theorem 2.

Polka dot: n Chain: $3n - 2$ Tunnel: $n(n+1)/2$

Crossed tunnel: n^2 Inward chain: n^2 Venn: $n2^n$

Fig. 13. Patterns of Euler diagrams and their contour scores

The 'polka dot' and Venn diagrams with n contours give the lower and upper bounds respectively on the contour score for a diagram with n contours.

Theorem 3. *Let d be an abstract Euler diagram with n contour labels. Then $n \leq \mathcal{CS}(d) \leq n2^{n-1}$.*
□

4.3 Clutter in nested diagrams

The next theorem gives the contour score and its variations for a nested diagram; part 1 generalises corollary 1. The theorem relates to some of the properties of the measures discussed in section 3.3.

Theorem 4. *Let $d = d_2 \xrightarrow{z^*} d_1$ be an abstract Euler diagram obtained by embedding d_2 in the zone $z^* = (x^*, y^*)$ of d_1 as in definition 3. Then*

1. $\mathcal{CS}(d) = \mathcal{CS}(d_1) + \mathcal{CS}(d_2) + (|Z_2| - 1)depth(z^*)$.
2. $\mathcal{SCS}(d) = \mathcal{SCS}(d_1) + \mathcal{SCS}(d_2)$.
3. $\mathcal{NCS}(d) = \mathcal{NCS}(d_1) + \mathcal{NCS}(d_2) + depth(z^*)$, provided d_2 is atomic.

Proof. In the notation of definition 3,

$$
\begin{aligned}
\mathcal{CS}(d) &= \sum_{(x,y)\in Z(d)} |x| \\
&= \sum_{(x,y\cup L_2)\in Z_1^*} |x| + \sum_{(x^*\cup x,\, y^*\cup y)\in Z_2^*} |x^* \cup x| - \sum_{(x^*,\, y^*\cup L_2)\in Z_1^*\cap Z_2^*} |x^*| \\
&= \mathcal{CS}(d_1) + \sum_{(x^*\cup x,\, y^*\cup y)\in Z_2^*} |x| + \sum_{(x^*\cup x,\, y^*\cup y)\in Z_2^*} |x^*| - depth(z^*) \\
&= \mathcal{CS}(d_1) + \mathcal{CS}(d_2) + |Z_2| depth(z^*) - depth(z^*).
\end{aligned}
$$

The result in part 1 now follows. The proofs of parts 2 and 3 follow immediately from the definitions. □

In example 2 (figure 3), we have $\mathcal{CS}(d_1) = 9$, $\mathcal{CS}(d_2) = 4$, $depth(z^*) = 2$ and $|Z_2| = 4$. Hence $\mathcal{CS}(d) = 9 + 4 + 3 \times 2 = 19$, $\mathcal{SCS}(d) = 9 + 4 = 13$ and $\mathcal{NCS}(d) = 9 + 4 + 2 = 15$. These results are easily verified by direct calculation.

5   Conclusion and Further Work

The empirical work and exploration of the properties of various clutter measures, confirms the claim, made in [7], that contour scoring provides an effective measure. Consideration of the properties of the contour score led to the definition of alternative measures (for nested diagrams). Further and more detailed statistical analysis of the experimental data is needed to confirm the initial findings presented in this paper. The experiment was designed and carried out before the additional measures of simple component contour scoring and nested component contour scoring were defined. A further study to explore subjects' scoring of the clutter of nested diagrams is desirable.

In attempting to understand the properties of clutter measures relating to nested diagrams, it was shown that all abstract Euler diagrams can be built inductively by inserting one contour at a time. This could produce an efficient method of generating a library of Euler diagrams. It is well known that not all concrete diagrams can be built in this way. For example, the Venn-5 diagram in figure 13 has the property that removing any contour results in a non-well-formed Euler diagram with disconnected zones. There are, of course, concrete versions of Venn-5 which can be constructed by adding one contour at a time. It is an open question whether, given any drawable abstract Euler diagram, there exists a concrete instantiation that can be constructed one contour at a time. In considering this question, describing a concrete version of our 'insert contour' rule and exploring its properties will be important. For example, it is not clear under what conditions a contour can be inserted into a concrete diagram to produce a well-formed diagram.

In this paper we consider 'abstract clutter' – clutter arising from the *structure* of the diagram. However, it is clear that various aspects of the layout of concrete diagrams can influence how cluttered it appears. Factors that appear to have an effect include the shape, size and relative positions of contours and whether there

is symmetry in the diagram. Further experimental and theoretical work is needed to explore the factors affecting what we might call 'concrete clutter'. This would include investigating various well-formedness conditions which provide different ways of representing the same information content.

Automatic theorem proving algorithms based on Euler diagrams are being adapted to consider minimising the length of a proof [3]. An alternative or additional criterion, designed to improve understandability, might be to use the measures to reduce clutter in proofs.

The ideas presented in this paper could have wider applicability beyond Euler diagram systems. Reducing visual complexity whilst maintaining information content could be seen as a desirable goal for many diagrammatic systems. The idea suggested here – to separate 'structural' (abstract) from 'visual' (concrete) clutter – may prove a useful framework in fulfilling this goal.

References

1. Bertini, E. and Santucci, G. *Improving 2D Scatterplots Effectiveness through Sampling, Displacement, and User Perception*, Ninth International Conference on Information Visualisation (IV'05), 2005, 826-834.
2. Flower J., Howse J. and Taylor J. *Nesting in Euler diagrams: syntax, semantics and construction.* J. Software Systems Modeling, 3, 2004, 55-67.
3. Flower J, Masthoff J, Stapleton G. *Generating Readable Proofs: A Heuristic Approach to Theorem Proving With Spider Diagrams*, Proc. Diagrams 2004, LNAI 2980, Springer-Verlag, 2004, 166-181.
4. Gil J., Howse J., Kent S. and Taylor J. *Projections in Venn-Euler Diagrams*, Proc. IEEE Symposium on Visual Languages (VL2000), IEEE, 2000, 119-126.
5. Howse J., Molina F., Shin S-J. and Taylor J. *On Diagram Tokens and Types*, Proc. Diagrams 2002, LNAI 2317, Springer-Verlag, 2002, 146-160.
6. Howse J., Stapleton G. and Taylor J. *Spider Diagrams*, LMS J. Comput. Math., 8, 2005,145-194.
7. John C. *Measuring and reducing clutter in Euler diagrams*, Proc. 1st International workshop on Euler diagrams, ENTCS 134, Elsevier, 2005, 103-126.
8. Lloyd N. *Clutter measurement & reduction for enhanced information visualization.* Masters thesis, Worcester Polytechnic Institute. Available at www.wpi.edu/Pubs/ETD/Available/etd-011206-232808/unrestricted/nlloyd.pdf
9. Peng W., Ward M. and Rundensteiner E. *Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering*, infovis, IEEE Symposium on Information Visualization (INFOVIS'04), 2004, 89-96.
10. Rosenholtz R., Li Y., Mansfield J., and Jin, Z. *Feature Congestion: A Measure of Display Clutter.* Proc. SIGCHI conference on Human Factors in Computing Systems, 2005, 761-770.
11. Shin S.-J. *The Logical Status of Diagrams*, Cambridge University Press, 1994.
12. Wickens, C. *Display Formatting and Situation Awareness Model (DFSAM): An approach to aviation display.* Technical report AHFD-05-14/NASA-05-5, 2005.