Kent Academic Repository

Visualizations with Venn and Euler Diagrams

Luana Micallef
School of Computing
University of Kent
Canterbury, UK
lm304@kent.ac.uk

Abstract

Venn and Euler diagrams intuitively visualize relationships and relative cardinalities of data sets. They are used extensively in areas as biosciences, business and criminology to facilitate data reasoning and analysis. However, current automatic drawing techniques do not always produce desirable diagrams. My research aims to develop novel algorithms to automatically draw readable diagrams that facilitate data analysis. It includes software, theory, and studies demonstrating how such diagrams aid probabilistic judgment.

1.   Introduction

Venn and Euler diagrams are composed of closed curves, which when they meet, split up the plane into regions that are uniquely identified by the curves that contain them. Venn diagrams represent all the intersections amongst the given sets. Thus, a Venn-n diagram depicts all of the 2^n intersections amongst n sets using n curves [29], [31]. If any of these intersections are missing, then it is not a Venn diagram but an Euler diagram [12]. This means that an Euler diagram can additionally represent set containment and exclusion. Fig. 1 illustrates a Venn diagram and an Euler diagram. If the region areas correspond to user-specified values, the diagram is area-proportional [10].
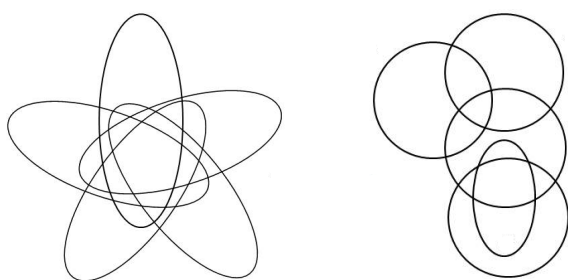

Fig. 1. A Venn diagram with 5 curves (also known as Venn-5) (*left*) and an Euler diagram with 5 curves (*right*) .

Venn and Euler diagrams are used extensively in application areas such as: medicine [21]; genetics [20]; biosciences [1]; proteomics [3]; health; ecology; criminology; segregation of duties; statistical data analysis; classification; information retrieval; search transparency. It is difficult to manually draw

these diagrams and hence, automatic drawing techniques have been proposed. However, these techniques do not always generate readable and aesthetically desirable diagrams and, if they do, they can be computationally expensive. Area-proportional diagrams are often drawn inaccurately and are usually misleading.

My PhD research aims to develop novel algorithms that draw readable Venn and Euler diagrams that facilitate data analysis. I devised a force-directed layout method for Euler diagrams (Section 2) and I developed the first area-proportional 3-set diagram drawing algorithm that uses ellipses (Section 3). I initiated theoretical work to identify region areas that can be represented accurately using area-proportional 3-set diagrams drawn with ellipses (Section 4). I question whether area-proportional Euler diagrams are the most appropriate visualization when reasoning about relative cardinalities of sets. I devised software that draws hybrid visualizations involving Euler diagrams and glyphs, and I then used these visualizations in a study to identify the most effective use of Euler diagrams for probabilistic judgment (Section 5). The design of the diagrams affects understanding and I am currently investigating ways how the design can be improved (Section 6). I conclude this paper with a summary of my contributions (Section 7) and a reflection on my research philosophy (Section 8). Unless specified, I carried out all of the mentioned work under the supervision of Peter Rodgers.

2.   Force-Directed Layout for Euler Diagrams

Current Euler diagram layout approaches involve evolutionary optimization [20], hill climbing and aesthetic layout metrics [13]. Such methods are computationally expensive and thus, impractical to lay out diagrams with a large number of curves. Large aesthetically pleasing graphs can be laid out relatively quickly using a force-directed approach, whereby a system of springs (the edges of the graph) and electrically charged particles (the vertices of the graph) forms a force model that is brought to stability in a sequence of iterations.

I adopted this force-directed approach for Euler diagrams [26]. Noting that the closed curves in an Euler diagram can be represented as polygons, attractive and repulsive forces between vertices, lines, and the polygons themselves can be used to produce good layouts in a relatively fast time. Different from graph drawing, devising an adequate force
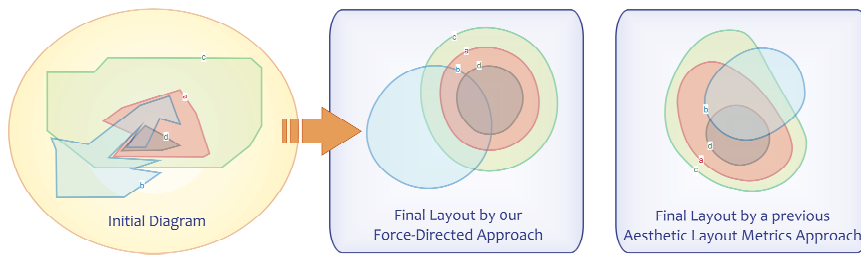
Left is an example of an initial diagram passed on to the layout algorithm

Centre is the final layout generated by my force-directed approach for the initial diagram on the left

Right is the final layout generated by the aesthetic layout metrics approach for an initial diagram similar to the one on the left.

Fig. 2. My force-directed approach vs. a previously proposed aesthetic layout metrics approach [13].

system in this case is more challenging; besides producing good layouts, the regions in the initial description have to be maintained. I devised an appropriate force model to handle all 3-, 4- and most 5-curve Euler diagrams that satisfy some aesthetic features, known as wellformedness criteria (example, not more than two curves intersect at a point). This first force-directed approach for Euler diagram layouts and its implemented prototype, demonstrate that this approach can produce good layouts in a reasonable time[1]. Fig. 2 illustrates an example and compares it to another layout approach.

However, what is a 'good' or 'optimal' layout? The only empirical study [2] that attempted to identify criteria for a good layout argues the need to avoid jagged polygons, inequality of region areas, and curves that are close to each other. These criteria are dealt with by my force model, but aesthetic metrics based on empirical studies and, general information visualization and perceptual principles [33] need to be conducted to evaluate and determine better layouts. Various experiments and expertise are required for such aesthetic metrics to be formalized. Thus, I decided to focus the major part of my PhD research on the generation of exact area-proportional Venn and Euler diagrams.

3. Drawing Area-Proportional Venn and Euler Diagrams using Ellipses

Area-proportional Venn and Euler diagrams are used in various application areas and, typically, they are made up of 3 sets. However, these diagrams (including those published in some of the world's most cited scientific journals) are often inaccurate and misleading [35]. Fig.3-O [21] is an example. For instance, regions with value 1% and 25% are respectively bigger than those with 3% and 29%. Problems of this sort are a symptom of the inherent inaccuracies when drawing such diagrams using circles. In fact, none of the 8 circle-based drawing tools (Fig.3- C1 to C8) generated an appropriate diagram for the data sets of Fig.3-O. Exact representations are possible with polygons, but their irregular shapes impede understanding. For instance, all of the 3 polygon-based methods (Fig. 3- P2, P3, P6) that generated an

---

[1] On average, 7, 26, 77 seconds for respectively 3-, 4-, 5-curve diagrams on an Intel Core 2 Duo CPU E7200 @ 2.53GHz with 3.23GB RAM running Microsoft Windows XP Professional.

exact diagram for the data sets of Fig. 3-O, produced sharply bended curves that are difficult to follow. In contrast, the curves of the diagram generated with my automated method eul*erAPE* using ellipses (Fig. 3-E1) are smooth and the region areas accurately represent the required values. Thus, having more degrees of freedom than circles, and a desirable smooth shape, unlike polygons, ellipses are more likely to produce exact, readable diagrams. However, using ellipses introduces new challenges (mainly, to compute the region areas and to position the ellipses).

3.1 Related Work

Various automatic drawing techniques using circles and polygons have been proposed, but none use ellipses.

3.1.1 Circle-Based Drawing Techniques

The formula defining the region areas in a diagram drawn with circles is non-invertible and thus, a non-algebraic mechanism must be adopted to draw the diagrams. Numerical methods have been used to produce exact area-proportional 2-circle Venn diagrams [10]. This concept has been extended for 3-circle diagrams [8](example Fig.3-C1), but in contrast to 2 circles, exact area-proportional diagrams cannot be drawn for most data [7]. Software to produce Venn and Euler diagrams is now available for various applications (diagram examples generated by the software are denoted in parenthesis): BioVenn [18](Fig.3-C2); PNNL Venn Diagram Plotter [27] (Fig.3-C6); Venn Diagram module in PatternLab for Proteomics [3] (Fig.3-C4); Google Venn Charts [16] (Fig.3-C3); R packages such as Vennerable [30] (Fig.3-C7), venneuler [35](Fig.3-C8); Stata module PVENN [17] (Fig.3-C5). In contrast to other methods, venneuler [35] uses a statistical approach and draws any *n*-circle diagrams more accurately than several other competitors. Section 3.2 compares this technique with eul*erAPE*.

3.1.2 Polygon-Based Drawing Techniques

In contrast to circles, often, polygon-based layout techniques produce exact diagrams. Symmetric convex Venn-3 diagrams have been studied, and construction mechanisms using 4- or 5-sided convex polygons [28] (Fig.3-P3) and convex circle-like regular polygons [20](Fig.3-P1) have been proposed. Other methods draw area-proportional curves using: triangles [30] (Fig.3-P4); rectilinear orthogonal curves [1] (Fig.3-P2), [30] (Fig.3-P5); rotated parallelograms;
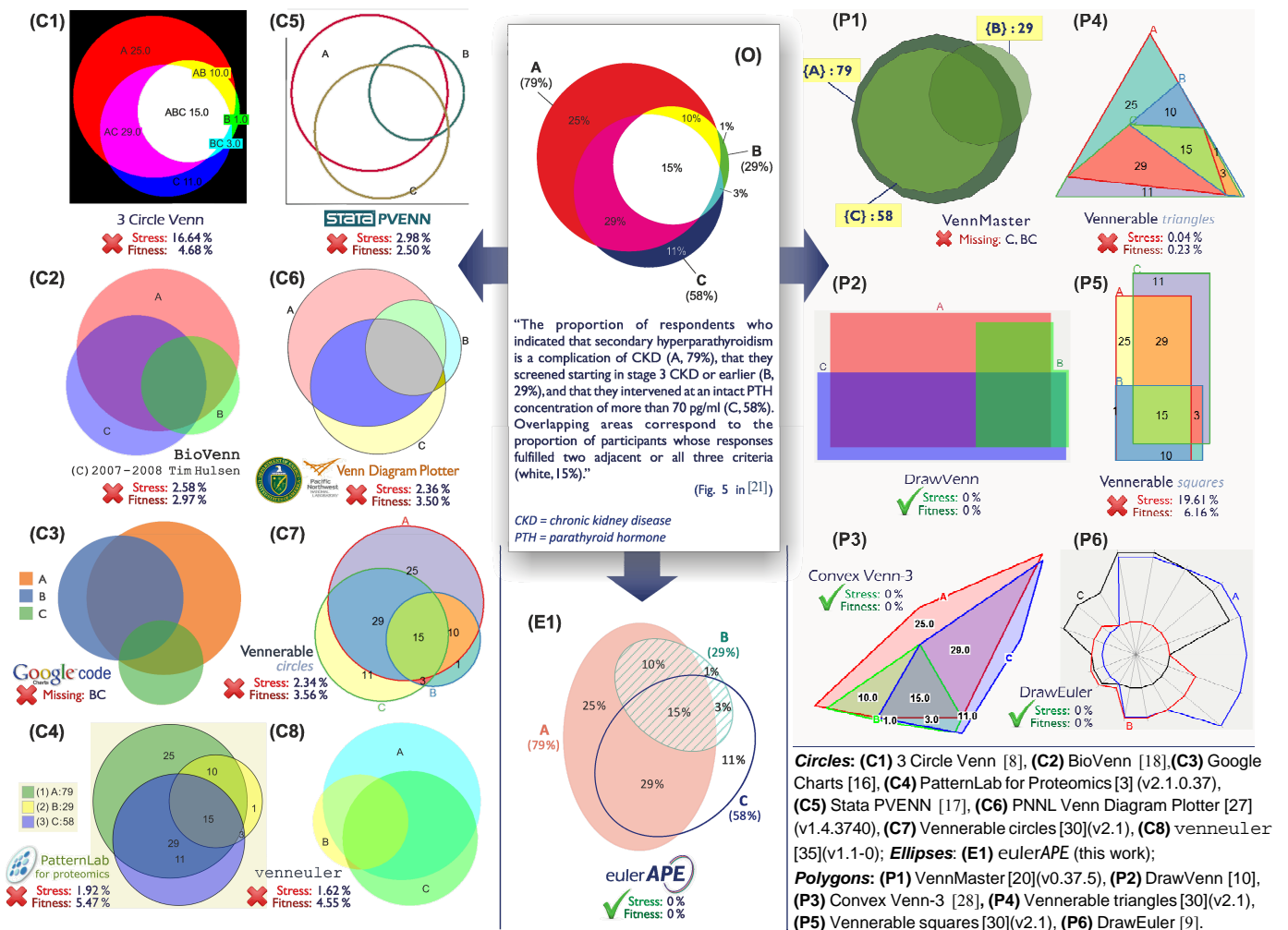
Fig. 3. Area-proportional Venn-3 diagrams for the real world data (O) [21] drawn using various circle-based (marked as C), ellipse-based (marked as E), and polygon-based (marked as P) automatic drawing tools. venneuler's stress measure [35] and eulerAPE's fitness measure are computed for the valid Venn-3 diagrams (diagrams with no missing regions) to determine the accuracy of the diagram representation.

convoluted curves [9][2] (Fig.3-P6). Typically, polygon-based approaches have irregular, jagged curves that are difficult to follow ([2], Gestalt's 'good continuation' principle [33]). Thus, the diagrams are often difficult to comprehend.

3.1.3 Ellipse-Based Drawing Techniques

The problems of circle accuracy and polygon usability motivate the use of ellipses. The authors of a few papers [1] manually drew their Venn-3 diagrams using ellipses. Developers of the circle-based [8] and polygon-based [10] drawing techniques also suggest the need of using more complex curves such as ellipses for the same reasons. However, using ellipses introduces new challenges, and thus, before eulerAPE, no technique used ellipses.

3.2 Approach and Uniqueness

I adopted a hill-climbing search approach to devise and implement eulerAPE[3], the first automated method to draw area-proportional 3-set diagrams using ellipses. Challenges included: 1) developing general analytical methods to calculate the area of regions formed when three ellipses intersect; 2) defining a fitness function to measure the diagram's accuracy in representing the user specified region areas; 3) generating a suitable starting diagram for a search optimization; 4) developing a search mechanism to explore various ellipses' configurations to find an appropriate area-proportional 3-set diagram.

I evaluated eulerAPE to determine its accuracy and examine its effectiveness for Venn-3 diagrams in comparison with the current best circle-based method. Firstly, I used eulerAPE to

---

[2] This general approach [9] draws monotone diagrams, whose curves tend to be convoluted and thus, the diagrams are not so usable.

[3] Available as open source at http://www.eulerdiagrams.org/eulerAPE.

generate diagrams for region areas obtained from randomly generated diagrams (*N=10,000*), and in the majority of the cases (98.93%), an exact solution was found. Thus, the method rarely encounters a local minimum. I then compared euler*APE* with the leading circle-based method, `venneuler` [35]. Randomly generated region areas (*N=10,000*) were drawn using both `venneuler` and euler*APE*, and the accuracy of the resultant diagrams was measured. Results demonstrate that the ellipse-based method outperforms the circle-based technique for virtually every case considered in this evaluation: using ellipses, euler*APE* generated an exact solution for 92.56% of the cases; using circles, `venneuler` failed in all of the cases. Finally, I compared our diagrams with those generated by some current circle- and polygon-based drawing tools for real world data (Fig.3). This demonstrates that, in contrast to circle-based (Fig.3- C1 to C8) and polygon-based (Fig.3- P1 to P6) methods, euler**APE** (Fig.3- E1) generates exact, readable diagrams with traceable curves and distinguishable regions.

3.3 Results and Contributions

Using ellipses, euler*APE* improves on previous Venn-3 drawing approaches by generating exact and aesthetically desirable diagrams for a large majority of randomly generated data sets (92.56%). Thus, using ellipses, diagrams are aesthetically desirable (as circles) and exact (as polygons).
A journal article [23] has been submitted in relation to euler**APE**_v2.0. A technical report [24] and a poster [25] were prepared for euler**APE**_v1.0. All the results mentioned here are related to the latest version, euler**APE**_v2.0.

Further work on the design of the diagrams and other formal evaluation investigating how effectively euler**APE** handles data sets with zero region areas is being carried out. Currently, euler**APE** draws diagrams with three curves, but it will be extended to handle more curves. Users particularly from medicine, biosciences, life sciences, proteomics and statistics have already expressed their appreciation for such a tool, and are using euler*APE* to better communicate and reason about their findings. PNNL is also recommending euler**APE** on their Venn Diagram Plotter webpage [27]. This demonstrates the need for such area-proportional ellipse-based drawing techniques to produce diagrams that intuitively visualize data set relationships, and thus, facilitate data analysis.

4. Which Region Areas Can Be Drawn Accurately Using Ellipses

The experimental results in Section 3 illustrate the benefits of using ellipses to draw accurate area-proportional Venn-3 diagrams. It is desirable to have analytical methods that determine whether a set of region areas can be represented with a Venn diagram drawn with ellipses. In statistical methods, a few samples of the actual population of interest are considered and, unless these samples represent the required population accurately, it is difficult to generalize reliable findings from statistical methods and, thus,

analytical methods are preferred. However, ellipses have various degrees of freedom (a semi-minor and a semi-major axis, a centre and a angle of rotation) and, defining analytical methods for Venn diagrams drawn with ellipses is challenging.
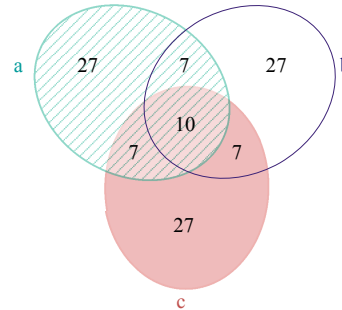

Fig. 4. A symmetry area-proportional Venn-3 diagram.

Thus, my initial theoretical work focuses on specific types of region areas and Venn diagrams, namely symmetric area-proportional Venn-3 diagrams drawn with ellipses. As shown in Fig. 4, due to symmetry, the area of the regions residing in the curves of such diagrams is one of 3 values. This is so, as one value is the area of the regions residing in exactly one curve, another value is the area of the regions residing in exactly two curves, and a third value is the area of the region residing in exactly the three curves. Region areas of this type make up a symmetric area specification.

In this work, I identify the symmetric area specifications that symmetric area-proportional Venn-3 diagrams drawn with ellipses can represent and, I define analytical methods to determine the class of symmetric area specifications for which a symmetric area-proportional Venn-3 diagram can be drawn using ellipses. This is the first theoretical work on Venn diagrams drawn with ellipses. Focusing my investigation on area specifications and Venn diagrams with restricted features, such as those which are symmetric, makes the task of analyzing them more tractable. This facilitates reason about the theoretical limits of such diagram constructions and provides guidance for further investigations of other more complex diagrams. To formalize my results, I am collaborating with Gem Stapleton (University of Brighton, UK). This work will be published.

5. Euler Diagrams for Probabilistic Judgment

Various application areas use area-proportional Venn and Euler diagrams to facilitate reasoning and analysis. Yet, are these diagrams the most appropriate visualization to reasoning about relative cardinalities of data sets? During my internship with INRIA, AVIZ team, I collaborated with Jean-Daniel Fekete and Pierre Dragicevic in a study [22] to assess the effect of different visualizations (involving Euler diagrams) for probabilistic judgment.

Both laymen and professionals struggle to understand statistical information [11]. Wrong judgments are common

and could result in overdiagnosis [34] or for instance, patients with a positive HIV test result attempting suicide before further tests turn out negative [6]. Juries can convict or acquit the wrong defendants based on unreliable evidence [15] and intrusion detection systems can give false alarms. Such scenarios involve Bayesian inference, which is known to be counterintuitive and subject to fallacious reasoning.

Though studies in psychology indicate that the representation of Bayesian problems affects comprehension, few visualizations were evaluated. In our study, we evaluated and compared a textual and 6 visual representations for 3 classical problems [14] using a diverse subject pool through crowdsourcing. The visualizations included area-proportional Euler diagrams, glyph representations and 3 different hybrid diagrams combining both, for which I devised novel algorithms to automatically generate the visualizations. Since some theories [19] indicate that subjects with low spatial abilities benefit mostly from visualizations with discrete and countable objects and, those with high spatial abilities benefit mostly from spatial representations such as Euler diagrams, we hypothesized that hybrid visualizations would be helpful to everyone, independent of their abilities.

Yet, our study demonstrated that none of the visualizations exhibited any measurable benefit and the subjects' accuracy was remarkably lower than those in previous psychology studies. However, through a second experiment, we demonstrate that simply adding a visualization to the textual problem does not aid reasoning and visualizations are more effective when the numerical values are removed from the text. Thus, besides improving our visualizations, novel ways for presenting visualizations to facilitate understanding of textual information need to be investigated and evaluated on heterogeneous populations of non-experts.

6. Design Principles of Diagrams

Having an exact diagram with desirable, traceable curves and distinguishable regions is not enough. The diagrams are often used by non-specialists and an improved design should facilitate the diagram's interpretation and comprehension. Moreover, improving the attractiveness of these diagrams is more likely to increase user engagement. As suggested by Ware [32], distinct simple features ensure a heterogeneous channel-based approach. This facilitates the identification of the sets in which any point in the diagram resides, and thus, every point would be independently searchable. For this to be possible, I use different features, namely colour, texture and outline to represent the different curves in the diagram, as illustrated in Fig.3-E1 and Fig.4. This creates an illusion of depth and layers to view the curves independently, even though they are on the same 2D plane. The subtle colours ensure minimal noise to avoid disruption of this illusion. Noting how such simple changes to the design could affect the way the diagram is perceived, I am collaborating with information designer, Angela Morelli (Central Saint Martins College of Art & Design, UK). Some of these design features include labelling strategies, multiple coordinated views to facilitate comparison of region areas, additional standard visual variables such as colour and fonts. These new features will be released in the coming months and all the work will be published at a relevant conference.

7. Conclusion

The main objective of my PhD research is to devise novel algorithms that automatically draw Venn and Euler diagrams that facilitate reasoning and data analysis. Contributions include: 1) the development of a force-directed approach to automatically lay out Euler diagrams of up to 5 curves in relatively fast time; 2) the development of the first automatic drawing tool for area-proportional 3-set diagrams using ellipses and the evaluation of the effectiveness of ellipses (in contrast to circles and polygons) in drawing accurate diagrams and region areas; 3) the formalization of theoretical work identifying region areas that can be represented accurately using an area-proportional Venn-3 diagram drawn with ellipses; 4) the development of an automatic drawing tool for hybrid diagrams combining Euler diagrams and glyphs, and providing guidance on how visualizations should be combined with textual information to facilitate reasoning and probabilistic judgment; 5) providing guidelines on how the design of the diagrams can be improved to facilitate diagram comprehension. These contributions should ultimately aid users to generate diagrams that help them communicate and reason about their data sets and, provide guidelines to developers of other novel drawing algorithms to generate readable and improved diagrams.

8. My Research Philosophy

I find research intriguing and extremely rewarding, especially when it is challenging and when it impacts a wider audience. I experienced this, particularly when I released my drawing tool, eulerAPE. Users from various application areas such as medicine, biosciences, proteomics and statistics and, from different international academic institutes and industry expressed their appreciation and provided us valuable feedback to improve next releases of the software. Some claimed that they had been years looking for automated methods that draw these diagrams using ellipses and they were delighted to have a tool that generates accurate diagrams, which they can use to help them explain their findings and reason about their data sets. PNNL is also recommending eulerAPE on their Venn Diagram Plotter webpage [27]. All of this, helped me appreciate my contribution and potential.

Publishing is a great contribution, but organizing events for your community can be extremely enriching. Together with Peter Chapman, I organized the 3rd International Workshop on Euler Diagrams. The second workshop was in 2005. Hence, it was time to bring the community together and to promote its research to other inter-disciplinary areas, with the aim of attracting new expertise

from other research institutes such as Microsoft Research and AT&T. Besides publishing the proceedings [5], we are also organizing the first special issue of a rank A journal on Euler diagrams [4].

Through research, I had the opportunity to meet various inspiring experts who provided me with valuable advice. Collaborating with mathematicians, information designers, and human computer interaction experts during my PhD helped me grow and become more open-minded. It was also very inspirational for my work. This was just the beginning. I want a research career and I hope I will be able to contribute to the needs of the greater community.

References

[1] H. el Azzouzi, et al. Peroxisome Proliferator-activated Receptor (PPAR) Gene Profiling Uncovers Insulin-like Growth Factor-1 as a PPARα Target Gene in Cardioprotection. *J Biol Chem*, 286(16), 2011.

[2] F.Benoy and P.Rodgers. Evaluating the Comprehension of Euler Diagrams. *Proc. IV 2007*, IEEE.

[3] P.C. Carvalho, et al. PatternLab for proteomics: a tool for differential shotgun proteomics. *BMC Bioinformatics*, 9(316), 2008.

[4] P. Chapman and L. Micallef (eds.). *Special Issue of the Journal of Visual Languages and Computing on Visualization and Reasoning with Euler Diagram*, 2013.

[5] P. Chapman and L. Micallef (eds.). *Proc. 3rd International Workshop on Euler Diagrams*, CEUR-WS.org/Vol-854.

[6] Chicago Tribune. A false HIV test caused 18 months of hell, (3/5/1993).

[7] S. Chow, *Generating and Drawing Area-Proportional Euler and Venn Diagrams*, PhD thesis, University of Victoria, Canada, 2007.

[8] S. Chow and P. Rodgers. Constructing area-proportional Venn and Euler diagrams with three circles. *Proc 2nd Intl Workshop Euler Diag*, 2005.

[9] S. Chow and F. Ruskey. Towards a General Solution to Drawing Area-Proportional Euler Diagrams. *Proc 1st Intl Workshop Euler Diag*, ENTCS 134:3-18, Springer-Verlag, 2005.

[10] S.Chow and F.Ruskey. Drawing Area-Proportional Venn and Euler Diagrams. *Proc 11th Intl Symp Graph Drawing*, LNCS 2912:466–477, Springer-Verlag, 2004.

[11] L. Cosmides and J. Tooby. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58(1), 1996.

[12] L.Euler, *Lettres à une Princesse d'Allemagne*, 2:102-108, 1761.

[13] J. Flower, et al. Layout Metrics for Euler Diagrams. *Proc. IV 2003*, pages 272–280. IEEE.

[14] G. Gigerenzer and U. Hoffrage. How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological Review*,102:684, 1995.

[15] I. J. Good. When batterer turns murderer. *Nature*, 375(6532), 1995.

[16] Google Inc. *Venn Charts*. Google Chart Tools, http://code.google.com/apis/chart/image/docs/gallery/venn_charts.html.

[17] W. Gong and J. Ostermann. *PVENN: Stata module to create proportional Venn diagram*. Statistical Software Components S457368, Nov 2011, Boston College Department of Economics.

[18] T. Hulsen, et al. BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics*, 9(488), 2008.

[19] V. Kellen, et al. Facilitating conditional probability problems with visuals. In *Conf HCI*, 2007.

[20] H.A. Kestler, et al. VennMaster: Area-proportional Euler diagrams for functional GO analysis of microarrays. *BMC Bioinformatics*, 9(67), 2008.

[21] O. Lenz and A. Fornoni. Chronic kidney disease care delivered by US family medicine and internal medicine trainees: results from an online survey. *BMC Medicine*, 4(30), 2006.

[22] L. Micallef, et al. Assessing the Effect of Visualizations on Bayesian Reasoning Through Crowdsourcing. *IEEE Trans Vis Comput Graph (Proc. IEEE InfoVis 2012)*, 2012, *to appear.*

[23] L. Micallef and P. Rodgers. Drawing Area-Proportional Venn-3 Diagrams Using Ellipses. *IEEE Trans Vis Comput Graph*, 2012, *under review.*

[24] L. Micallef and P. Rodgers. Drawing Area-Proportional Venn-3 Diagrams Using Ellipses. *TR-3-11*. University of Kent, UK, 2011.

[25] L. Micallef and P. Rodgers. euler*APE*: Drawing Area-Proportional Euler and Venn Diagrams using Ellipses. *EMEA Google Scholars' Retreat - Poster*, 2011.

[26] L. Micallef and P. Rodgers. Poster: Force-Directed Layout for Euler Diagrams. *IEEE InfoVis 2009*.

[27] PNNL and OMICS.PNL.GOV. *Venn Diagram Plotter*. http://omics.pnl.gov/software/VennDiagramPlotter.php.

[28] P. Rodgers, et al. Drawing Area-Proportional Venn-3 Diagrams with Convex Polygons. *Proc 6th Intl Conf Diag Rep Inf*, LNCS (LNAI) 6170, Springer-Verlag, 2010.

[29] F.Ruskey and M.Weston. A Survey of Venn diagrams. *Elec J Combin*, 1997. Dynamic survey, Article DS5, 2005.

[30] J. Swinton. *Venn diagrams in R with the Vennerable package*. 2011.

[31] J.Venn. On the diagrammatic and mechanical representation of propositions and reasonings. *Dublin Phil Mag, J Science*, 9(59), 1880.

[32] C. Ware. *Visual Thinking for Design*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.

[33] C.Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., USA, 2004.

[34] H. G.Welch and W. C. Black. Overdiagnosis in cancer. J *National Cancer Institute*, 102(9), 2010.

[35] L. Wilkinson. Exact and Approximate Area-proportional Circular Venn and Euler Diagrams. *IEEE Trans Vis Comput Graph*, 18(2), 2012.