Radial Sets: Interactive Visual Analysis of Large Overlapping Sets

Bilal Alsallakh, Wolfgang Aigner, Silvia Miksch, and Helwig Hauser

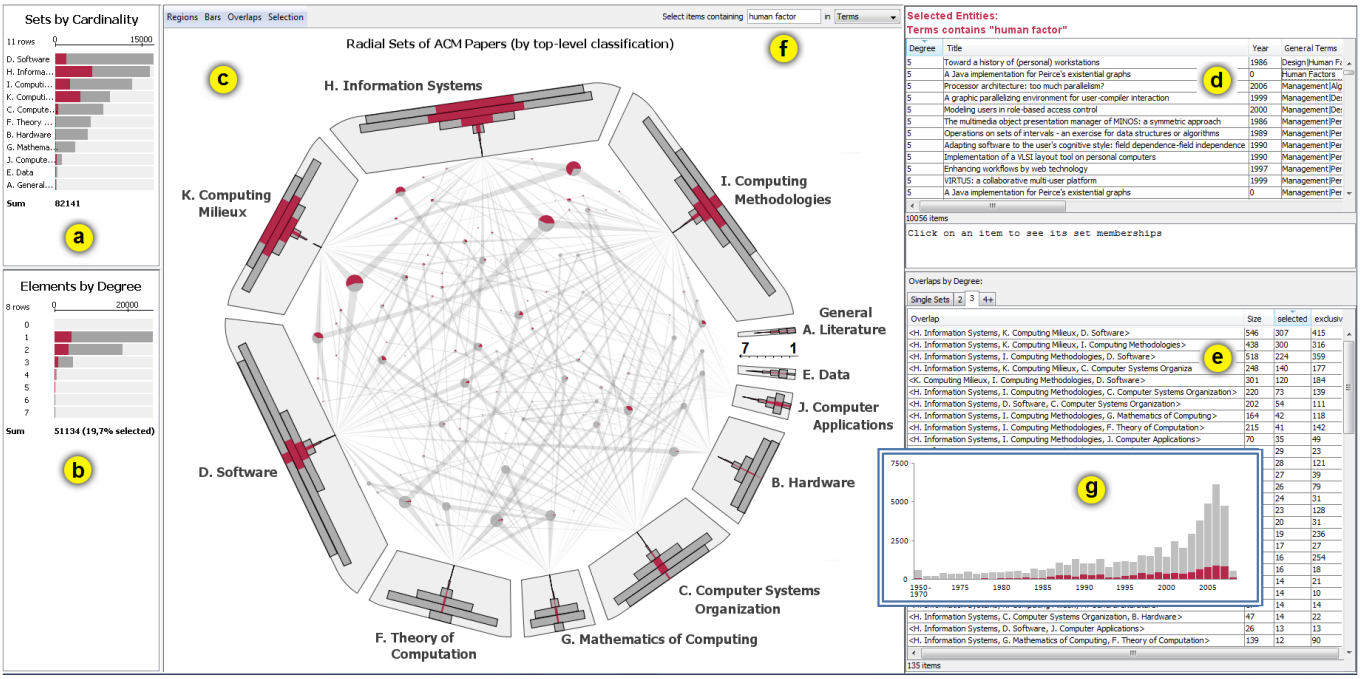

Fig. 1. The main interface of Radial Sets: (a) the sizes of the overlapping sets, (b) a histogram of the elements by degree, (c) the *Radial Sets view* showing $n > 50{,}000$ papers multi-classified into 11 ACM classes [1]; hyperedges of degree 3 are depicted to indicate overlaps between triples of sets; (d) a list of $1{,}098$ selected elements and their attributes, along with a natural text describing the selection criteria, (e) the *overlap analysis view* showing details about overlaps classified by degree into different lists, (f) a search box to select elements containing a specific text, (g) a linked view showing the publication dates for all papers and for the ones in (d).

Abstract—In many applications, data tables contain multi-valued attributes that often store the memberships of the table entities to multiple sets such as which languages a person masters, which skills an applicant documents, or which features a product comes with. With a growing number of entities, the resulting element-set membership matrix becomes very rich of information about how these sets overlap. Many analysis tasks targeted at set-typed data are concerned with these overlaps as salient features of such data. This paper presents Radial Sets, a novel visual technique to analyze set memberships for a large number of elements. Our technique uses frequency-based representations to enable quickly finding and analyzing different kinds of overlaps between the sets, and relating these overlaps to other attributes of the table entities. Furthermore, it enables various interactions to select elements of interest, find out if they are over-represented in specific sets or overlaps, and if they exhibit a different distribution for a specific attribute compared to the rest of the elements. These interactions allow formulating highly-expressive visual queries on the elements in terms of their set memberships and attribute values. As we demonstrate via two usage scenarios, Radial Sets enable revealing and analyzing a multitude of overlapping patterns between large sets, beyond the limits of state-of-the-art techniques.

Index Terms—Multi-valued attributes, set-typed data, overlapping sets, visualization technique, scalability

1 Introduction

Sets are one of the most fundamental concepts in mathematics. A set is a collection of unique objects, which are called elements of the set. Because of their simple and generic notion, sets are widely used in computer science to represent real-world concepts, query results, and the results of various algorithms. Compared to lists, sets ensure the uniqueness of their elements and impose no order on them. A set system comprises multiple sets defined over the same elements. Multiple set memberships are common in practice to represent both technical and real-world concepts. As an example, they can represent people memberships to different clubs, the markers a gene contains, or multiple tags or labels assigned manually or automatically to a set of entities. these memberships are usually stored in a database using either a multi-valued attribute or a group of Boolean attributes.

Sets defined over the same elements in a dataset potentially overlap. With a growing number of elements, these large overlapping sets contain a wealth of patterns that are worth to discover and analyze. Euler diagrams are the most common and natural way for depicting overlapping sets. However, they are inherently limited in terms of scalability.

- *Bilal Alsallakh, Wolfgang Aigner and Silvia Miksch are with Vienna University of Technology. E-mail: {alsallakh, aigner, miksch}@ifs.tuwien.ac.at*
- *Helwig Hauser is with University of Bergen. E-mail: helwig.hauser@uib.no*

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

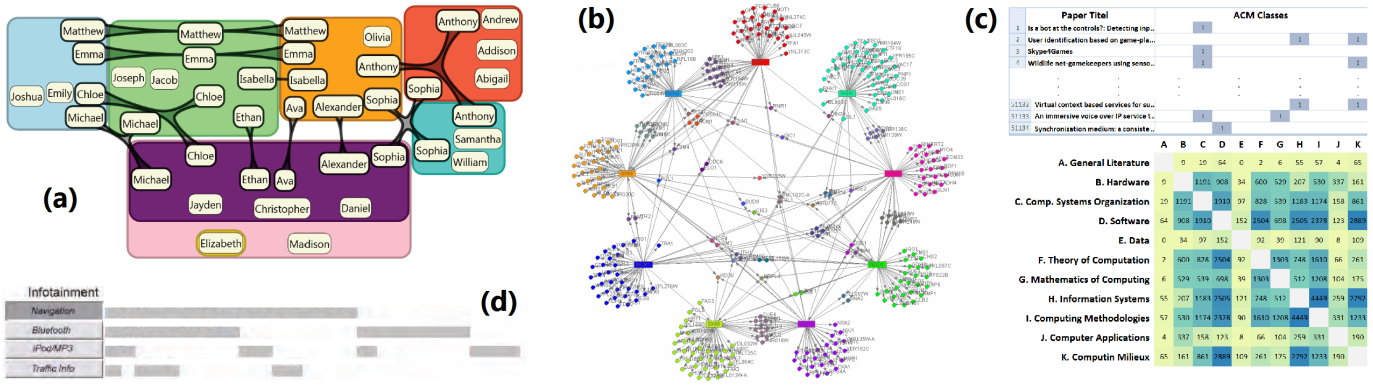

Fig. 2. Four techniques for visualizing element-set memberships: (a) untangled Euler diagrams [21] with duplications of elements that belong to multiple sets, (b) Anchored Maps [31] with sets represented as anchors on a circle and elements as free nodes, (c) two reorderable matrices [5] showing the element-set memberships and the set overlaps. (d) equal-height histograms [47] showing elements as bars in different rows,

In this paper we introduce a novel visualization technique for analyzing large overlapping sets. Our technique, called Radial Sets[1], shares several properties with state-of-the-art techniques proposed for the same purpose (Sect. 2). It builds upon selected ideas from these techniques to improve both on readability and scalability, and to support advanced analysis and pattern-finding tasks for this kind of data. In particular, given set memberships of a large number of elements in about $m \leq 30$ sets, Radial Sets enable the following analysis tasks that are common for this kind of data [18, 21, 37]:

- **T1:** Analyze the distribution of elements in each set according to their degrees (the number of sets they belong to).
- **T2:** Find elements in a specific set that are exclusive to this set, or that belong to at least, at most, or exactly k other sets.
- **T3:** Analyze overlaps (intersections) between groups of k sets.
- **T4:** Analyze overlaps between pairs of sets: find which pairs of sets exhibit higher overlap than other pairs (related to T3).
- **T5:** Find elements that belong to a specific overlap.
- **T6:** Analyze how an attribute of the elements correlates with their memberships to the sets and the overlaps.
- **T7:** Analyze how set memberships and attribute values for a selected subset of elements differ from the rest of the elements.

The tasks **T1** and **T2** are concerned with element memberships in the sets. For example, if the sets are defined over products to represent the features they come with, a typical question about one feature is whether it tends to come exclusively, or along with one, two, or more other features. The overlap tasks **T3**, **T4**, and **T5** enable finding out which feature combinations are more common among the products, and which products belong to these combinations. The attribute analysis tasks **T6** and **T7** answer questions like how the price of a product depends on its features and whether certain feature combinations are particularly cheap or particularly expensive.

As we show in Sect. 3, the visual design of our technique is derived from the requirements of these tasks. It employs frequency-based representations of the set elements to support the memberships tasks **T1** and **T2** in a scalable way. Also, it dedicates a large portion of the screen space to emphasize the overlaps as first-order objects in the visualization, as required by the overlap tasks. Both the set elements and the overlaps are visualized using area-based representations. This supports using retinal variables [5] like color to show information about the elements, as required by the attribute-analysis tasks.

Sect. 4 presents two usage scenarios of Radial Sets to demonstrate how they can be used to perform the tasks **T1, ⋯, T7** with large sets defined over thousands, to hundreds of thousands of elements. In Sect. 5 we discuss the applicability and the limitations of Radials Sets, and outline possibilities for future work.

[1] A prototype implementation is available at www.radialsets.org

2 State of the Art

Despite the simple notion of a set system, visualizing overlapping sets is a challenging problem that has been approached in various ways. The major reason behind the complexity of this problem is the exponential growth of possible overlaps according to the number of sets: a set system with m sets can exhibit up to 2^m distinct intersections between the sets [41]. Each element lies in one of these intersections, based on its memberships to the different sets. Although a large portion of these distinct intersections is empty in practice, the number of non-empty overlaps can still be large, even with a dozen sets. These overlaps are salient features of set data with many analysis tasks typically concerned with different kind of overlaps between the sets.

Some techniques for visualizing overlapping sets bypass the complexity problem by limiting the number of sets and overlaps that can be visualized at once. Other techniques avoid visualizing the overlaps explicitly and convey more abstract information about the set system instead. In the following, we categorize existing techniques based on the visual representations they use and discuss their scalability and which of the tasks listed in Sect. 1 they support.

2.1 Euler Diagrams and Euler-like Diagrams

Euler diagrams [15] represent sets as closed regions in the plane, providing a very natural way to depict overlaps. However, they suffer from a severe limit: all possible overlaps can be depicted distinctively only with a small number of sets $m \leq 4$. Verroust and Viaud [42] showed that this limit can be increased to $m \leq 8$ by relaxing the conditions on the contours and by allowing holes in the regions.

Several techniques have been recently devised to automatically generate Euler-like diagrams. The methods of Flower et al. [16, 17] generate Euler diagrams in case of drawability. Rodgers et al. [33] and Simonetto et al. [38] presented techniques that generate an output even for undrawable instances by allowing disconnected regions. Both techniques can result in complex non-convex zones especially when the sets exhibit numerous overlaps. Henry Riche and Dwyer [21] proposed two variations to draw simplified rectangular Euler-like diagrams that also represent individual elements. Their second variation, called DupED, does not depict the intersections between the sets explicitly. It rather creates separate rectangular regions for the sets, and duplicates the elements that belong to multiple sets. Multiple instances of the same element are then linked with hyperedges (figure 2a). Recent work has focused on generating area-proportional Venn and Euler diagrams [8, 26, 44]. Such diagrams convey how large the overlaps are compared to each other without depicting the elements. However, generating these diagrams accurately is restricted to three sets.

Euler-like methods have also been employed to visualize set memberships over existing visualizations that determine the positions of the elements. BubbleSets [10], LineSets [3] and Kelp diagrams [12, 30] are examples of such methods with varying design goals and degree of

compactness. Itoh et al. [24] proposed depicting the set memberships as colored glyphs inside the visual elements. Each set is hence denoted by disconnected regions linked only by having the same color.

In summary, methods based on Euler diagrams often impose severe limits on the number of sets, elements, and overlaps they can depict, and hence can only partially cope with the tasks **T2**, **T3**, **T4** and **T5**.

2.2 Node-link Diagrams

A set system of m sets $S_{1\leq j\leq m}$ defined over n elements $e_{1\leq i\leq n}$ can be modeled as a bipartite graph $G = (V1 \cup V2, E)$. The vertices of this graph are the elements $V1 = \{e_i : 1 \leq i \leq n\}$ and the sets $V2 = \{S_j : 1 \leq j \leq m\}$. The edges $E = \{(e_j, S_i) : e_j \in S_i\}$ are the membership relations between the elements and the sets. A variety of approaches were devised both for drawing [11, 50, 32] and for visualizing [31, 35] bipartite graph as node-link diagrams. Anchored Maps [31] place the vertices of one class as anchors on a circle. The vertices of the other class are placed as free nodes with links connecting each free node with the anchors it has edges with (figure 2b). The position of these free nodes are determined by spring embedders.

A set system can be depicted as an Anchored Map by representing the sets as anchors and the elements as free nodes. This enables quickly finding which elements are exclusive to each set, and which elements are shared between multiple sets, partially solving the tasks **T2** and **T4**. However, with an increasing number of elements shared between multiple sets, the view becomes quickly cluttered making it difficult to recognize which elements belong to which overlap. This is an inherent limitation of node-link diagrams that restricts their applicability to a small number of elements.

Hypergraphs offer a more general way to model a set system with each set represented by a hyperedge that connects all element vertices in this set, or vice versa. The two general approaches to draw hypergraphs [28] roughly resemble Euler diagrams (subset standard) and node-link diagrams (edge standard).

2.3 Matrix-based Methods

A matrix can depict memberships of n elements represented as rows in m sets represented as columns (figure 2c-top). Bertin described how reordering the rows and columns can simplify such matrices [5]. This ordering has a significant impact on the ability to find patterns in the matrix, especially clusters of elements that exhibit similar patterns of memberships of the sets and vice versa [6, 45]. As the ordering problem is NP-complete [29], a large number of heuristics have been proposed for reordering matrices [27]. In addition, several interactive systems have been proposed to create and refine reorderable matrices for different purposes [22, 36, 40].

With a growing number of relations, the membership matrix outperforms node-link diagrams in several low-level reading tasks [19]. However, it falls short of solving tasks specific to set data. A separate matrix is needed to explicitly reveal the overlap between pairs of sets (task **T4**) as a heatmap (figure 2c-bottom). Henry Riche et al. [23] augmented matrices with links that show additional relations between the rows or the columns (figure 2c). Similar ideas can partially support **T4** in the membership matrix without the need for a separate matrix. Another problem with matrix representations is scalability: A large number of elements that belong to a smaller number of sets result in a skewed membership matrix. This is challenging for multi-level techniques that are usually designed for square matrices [14].

2.4 Frequency-based Methods

Node-link diagrams and memberships matrices offer item-based representations of overlapping sets that create a distinct visual item, like a node or a row for every element in the sets. In contrast to that, frequency-based representations aggregate multiple elements that belong to specific overlaps into a single visual item like a bar. This makes them potentially scalable in the number of elements they can depict.

Wittenburg proposed an extension to bargrams [46] to depict set-valued attributes [47]. The sets are represented as rows in the bargrams, sorted from the largest to the smallest. The horizontal dimension represents all the elements, sorted by their membership of the topmost set, then of the second topmost set, and so on. Bars are drawn in each row to depict the elements that belong to the corresponding set according to this sorting (figure 2d). This reveals different overlaps between the sets, however, from the perspective of the larger sets which define the elements' order. A different ordering of the rows is needed to infer the overlap between the two bottommost set.

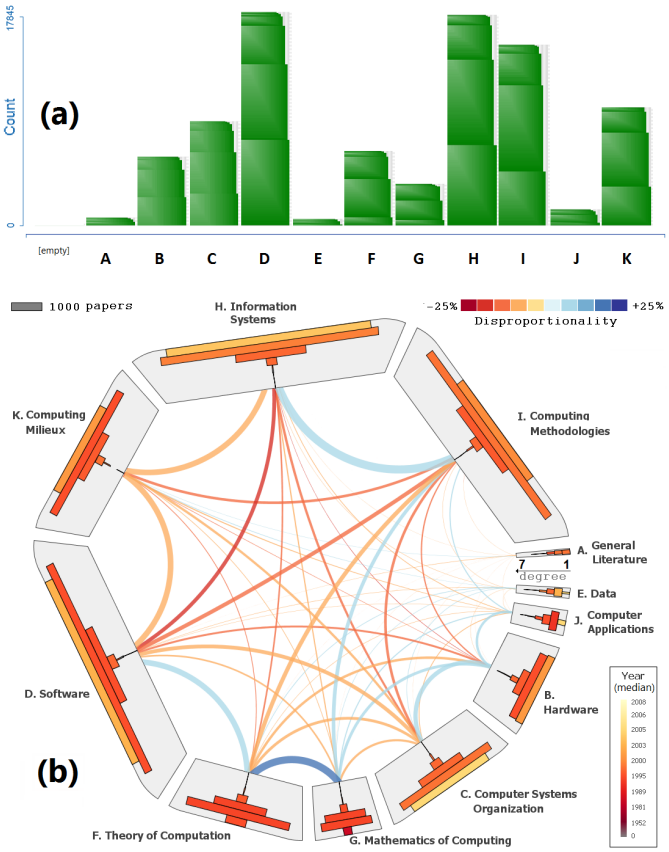

Fig. 3. (a) Set'o'grams [18] showing 11 overlapping sets as bars of proportional size, divided into groups of elements of equal degree, (b) Radial Sets showing the same data with overlaps between pairs of sets depicted as arcs. Ideally, only one color scale should be used.

Set'o'grams [18] extend bar charts to visualize overlapping sets. Each set is represented by a bar of proportional size. This bar is divided into sections that represent the different degrees of elements in the respective set (figure 3a). The degree of an element is equal the number of sets it belongs to. The sections are distinguished both from each other both by shading, and by assigning increasingly smaller widths to sections of higher degrees. Hence, it is possible to infer for each sets how many elements belong exclusively to it and how many of its elements belong to k other sets, solving exactly tasks **T1** and **T2**. Interaction by means of brushing can solve task **T5** but falls short of providing an overview of overlaps required for tasks **T3** and **T4**.

Our work extends the basic idea of Set'o'grams. It employs an alternative visual design that emphasizes the single sections in the bars and allows depicting different kinds of overlaps as we show next.

3 Radial Sets

To enable a scalable visual analysis of large overlapping sets, Radial Sets employ frequency-based representations that aggregate the elements in the sets and in their overlaps. Also, multiple views depict the information at multiple levels of detail. The main view (Sect. 3.1) shows both the distribution of elements in the sets and the overlaps between the sets. Additional views show both summary and detailed information about the elements and the overlaps (Sect. 3.2). Together, these views enable an elaborate analysis of overlapping sets.

3.1 The Visual Metaphor

To visually encode overlapping sets, Radial Sets use three types of visual elements: (1) regions to represent the sets, (2) histograms inside the regions to represent the elements in each set, and (3) links between the regions to represent overlaps between the sets. Figure 4 shows how four overlapping sets are represented as Radial Sets.

3.1.1 Visualizing the sets

Radial Sets represent the sets as uniformly-shaped non-overlapping regions. The regions are arranged radially on a circle. This arrangement aims mainly to ease the depiction of the overlaps between the sets as links inside this circle, and to emphasize them as the central part of the visualization. Moreover, it facilitates the interpretation of the histograms representing the elements in the individual regions as we explain in Sect. 3.1.2.

Unlike Set'o'grams [18], the areas of the regions are not necessarily proportional to the sizes of the sets. A dedicated view in the user interface conveys these sizes more effectively via a bar chart (Sect. 3.2.1). Depending on how the histograms are scaled, the regions can be either made of equal area or assigned different areas to fit the histograms. In the latter case, the regions are depicted as rounded parallelograms leaving equally-sized gaps between the regions. This alleviates visual artifacts and asymmetries caused by non-uniform gaps. However, the parallelograms might imply 3D cues to the regions, which impacts the accuracy of perceiving the bars insides these regions.

The use of distinct visual elements to represent the sets and the overlaps enables using simple shapes to depict the set regions. As discussed in Sect. 2.1, a similar idea was employed by Henry Riche et al. to simplify Euler diagrams [21]. They argued that the use of convex and simple regions is a primary factor impacting readability, as shown by empirical results in Gestalt psychology [25]. We also duplicate the representations of elements that belong to multiple sets, like in the untangled Euler diagrams (figure 2a). However, we aggregate these elements, and the overlaps they result in as we describe next.

3.1.2 Visualizing the elements

Like Set'o'grams [18], Radial Sets aggregate the elements of each set into groups according to their degrees. In a system of m sets $S_{1\leq j\leq m}$ and n elements $E=\{e_i : 1\leq i\leq n\}$, the degree of an element $e\in E$ is equal to the number of sets it belongs to:

$$degree(e)=|\{S_j : 1\leq j\leq m \wedge e\in S_j\}| \quad (1)$$

The elements of each set S_j are aggregated via a histogram H_j of their degrees. Each histogram consists of $b=d$ bins with d denoting the largest number of sets that share at least one item:

$$d=max\{degree(e) : e\in E\} \quad (2)$$

Hence, the number of items in bin k of histogram H_j is:

$$h_{jk}=|\{e\in S_j : degree(e)=k\}| \quad (3)$$

It is possible to use a smaller number of bins b than d. In this case the last bin b aggregates elements having degrees equal to or higher than b:

$$h_{jb}=|\{e\in S_j : degree(e)>=b\}| \quad (4)$$

This aggregation limits the analysis to overlaps between $2,3,\ldots$, till b-or-more sets. This is desirable since usually only few elements have high degrees. Aggregating them simplifies the visualization. The degree histogram retains access to these elements (Sect. 3.2).

The histograms $H_{1\leq j\leq m}$ are placed radially in the regions of their respective sets. The radial dimension encodes the elements' degrees k, with h_{j1} mapped to the outermost boundary of region S_j and h_{jb} mapped to the innermost boundary (figure 4b). This intends to emphasize that the items in outermost bar are exclusive to the respective set, while the items of the innermost bar are shared with multiple other sets. This is analogous to the magnet metaphor of Yi et al. [49] with set labels acting as magnets on the radial dimension.

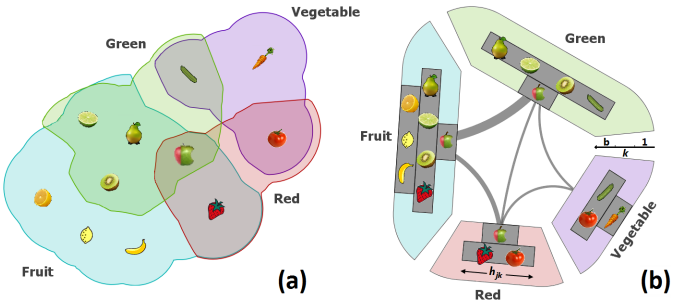

Fig. 4. (a) An Euler diagram (adapted from Wyatt [48]), (b) the equivalent representation in Radial Sets. The histograms in gray show a breakdown of the elements in each set by their degrees (Eqs. 1, 3). The arcs show overlaps between pairs of sets. The icons are for illustration only.

Bars representing the same degree k in different histograms $\{H_j\}$ are located at the same radial position in their regions. This makes it easier to identify and interact with these bars than in Set'o'grams, where sections of the same degree are located at different heights. Furthermore, gaps in the distribution can be more easily identified, since the bars do not need to be stacked like the sections in Set'o'grams.

The bars are by default centered in their regions to avoid artificial asymmetry across the histograms and to make comparing their shapes easier. Moreover, the symmetry facilitates perceiving the histograms as figures or objects in their regions following Gestalt laws [43]. This emphasizes that these objects represent elements contained in the respective sets. A similar layout was used for augmenting histograms over the axes of parallel coordinate plots [20]. However, the lack of a baseline, the radial arrangement, and the 3D visual cues (Sect. 3.1.2) impact the accuracy of comparing the length of individual bars and of estimating selected fractions of these bars (figure 1c). Therefore, Radial Sets offer an overview visualization, with precise comparisons needed to be performed on demand as we discuss in Sect. 5.

The histogram scales can be either uniform or assigned individually to fit the histograms in regions of equal area. Uniform scaling is useful for comparing the bars of different histograms in length. Nonuniform scaling is useful for comparing the shapes of the histograms especially when the sets exhibit a large variance in size. In the latter case, the different scales can be indicated via rectangles along the h_{jk} axes (figure 6) scaled differently in each region to depict the same number of elements, as suggested by Cleveland [9, p. 90].

Representing the elements in each set as a histogram of their degrees gives an idea of how much overlap this set has with how many sets. This solves the tasks **T1** and **T2**. However, histograms do not tell with which sets these overlaps are. As we show in Sect. 3.2, all 2^m possible overlaps can be analyzed on demand via interaction with the histograms. But to gain an overview of individual overlaps, additional visual elements are needed as we show in the next section.

3.1.3 Visualizing the overlaps

An overlap $O_{\{j_1,\ldots,j_k\}}=\bigcap_{l=1}^{l=k} S_{j_l}$ is the intersection between k specific sets $\{S_{j_1},\ldots,S_{j_k}\}$ in the set system. By k we denote the degree of the overlap. Each element e in this overlap is of $degree(e)\geq k$. Hence, this overlap contains overlaps of higher degree $O_{J\supset\{j_1,\ldots,j_k\}}$, and can intersect with other overlaps of degree k. The elements exclusive to an overlap $O_{\{j_1,\ldots,j_k\}}$ are:

$$EO_{\{j_1,\ldots,j_k\}}=\{e\in O_{\{j_1,\ldots,j_k\}} : degree(e)=k\} \quad (5)$$

Radial Sets map overlaps to frequency-based representations of proportional size. These representations can either depict the absolute sizes of the overlaps or their normalized sizes.

$$nsize(O_{\{j_1,\ldots,j_k\}})=\frac{|O_{\{j_1,\ldots,j_k\}}|}{|\bigcup_{l=1}^{l=k} S_{j_l}|} \quad (6)$$

Normalization makes it easier to compare overlaps between sets of different sizes by emphasizing the proportions of the respective sets they represent, as illustrated in figure 5. Eq. 6 computes the normalized size of an overlap by considering only the sets involved in this overlap. Disproportionality measures offer another possibility to compare two overlaps, taking into account all elements E in the set system. The disproportionality of an overlap is the deviation between the actual and expected probabilities of an element $e \in E$ to lie in this overlap:

$$disproportionality(O_{\{j_1,\ldots,j_k\}}) = \frac{|O_{\{j_1,\ldots,j_k\}}|}{n} - \prod_{l=1}^{k} \frac{|S_{j_l}|}{n} \qquad (7)$$

The expected probabilities are computed by assuming marginal independence of the sets. The resulting residuals can take either positive or negative values, and can be conveyed by coloring the overlaps using a diverging color scale. Other residuals are also possible to eliminate a possible bias in Eq. 7, caused by the sets being of different sizes [4].

To simplify overlap analysis, we restrict the visualization by default to overlaps of a certain degree k selected by the user. This is in accordance with task **T3**, where users ask questions like "which three sets exhibit disproportionally large overlap?". Moreover, this simplifies the visualization by reducing the number of visual elements needed to depict the overlaps and by making these element to have the same semantics and similar shapes. The number of possible overlaps of degree k is equal to $\binom{m}{k}$, the number of possible combinations of k objects from a set of m objects. This number can be relatively large for values of k larger than 2. Therefore, Radial Sets adopt different strategies for depicting overlaps, depending on their degrees and actual count.

Visualizing overlaps of degree $= 2$ as arcs

Radial Sets visualize overlaps between pairs of sets (task **T4**) as arcs between their regions. The thickness of an arc encodes the absolute or the normalized size of the overlap (figure 6). To alleviate clutter that results from arc crossings, the regions are ordered so that thicker arcs are kept as short as possible. For this purpose, we use a greedy algorithm that iteratively concatenates chains of regions, starting from the individual regions. At each iteration, the algorithm selects the next thickest arc between two regions and concatenates the two chains that contain these regions in one chain, optimizing on the arc length:

Algorithm 1 Compute regions' order to shorten thick arcs

```
for all j in 1...m do
    chain[j] ← {j} as list
end for
overlaps ← {O_{j1,j2} : 1 ≤ j1 < j2 ≤ m} as list
Sort overlaps in descending order of |O_{j1,j2}| or nsize(O_{j1,j2})
for all O_{j1,j2} in overlaps do
    if chain[j1] ≠ chain[j2] then
        c[1] ← concatenate(chain[j1], chain[j2])
        c[2] ← concatenate(chain[j1], reverse(chain[j2]))
        c[3] ← concatenate(chain[j2], chain[j1])
        c[4] ← concatenate(chain[j2], reverse(chain[j1]))
        {concatenate according to the shortest arc j1^ j2^}
        index ← argmin_i{j1^ j2^ computed in c[i] : 1 ≤ i ≤ 4}
        chain[j1] ← c[index]
        chain[j2] ← c[index]
        if |c[index]| = m then {all regions are in one chain}
            return c[index]
        end if
    end if
end for
```

The ordering problem resembles the seriation problem [7, 27] in reorderable matrices (Sect. 2.3). The computed order not only alleviates clutter, but also reveals clusters of sets having high overlap with each other. To analyze these overlaps more explicitly, links of higher degree are needed instead of the arcs as we explain next.

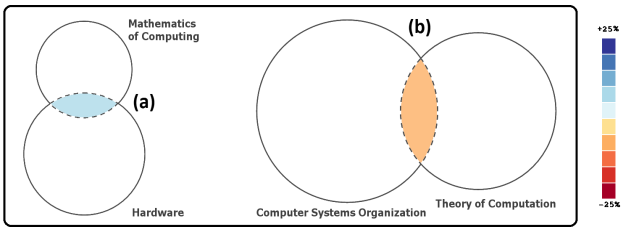

Fig. 5. Two overlaps of 2nd-degree, having different absolute sizes, but nearly equal normalized sizes (Eq. 6). The color denotes the overlap disproportionality (Eq. 7) using the same color scale as in figure 3b.

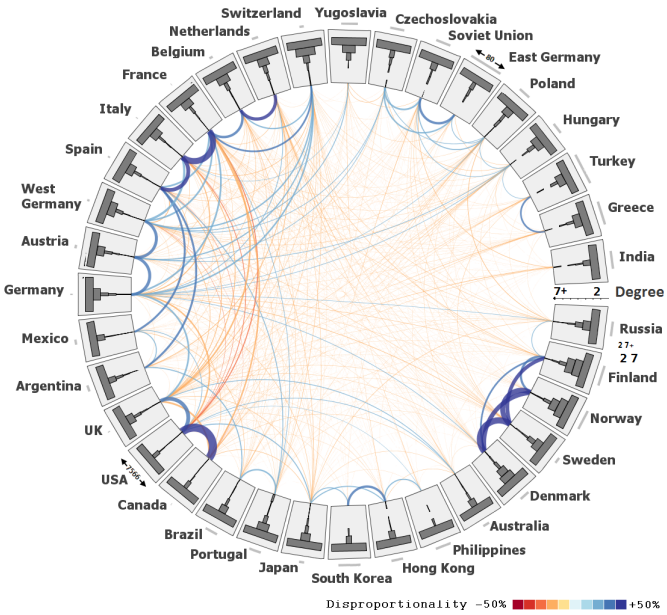

Fig. 6. Radial Sets depicting IMDb movies produced in two or more countries (including former countries). An arc between two countries represents the overlap between their movies. Its thickness and color respectively encode the normalized size (Eq. 6) and the disproportionality (Eq. 7) of this overlap. The different scales of the histograms are indicated as thin rectangles representing the same number of elements.

Visualizing overlaps of degree ≥ 3 as hyperedges

To visualize the overlap between $k \geq 3$ sets (task **T3**), Radial Sets create a bubble of proportional size in the inner area. The bubble is connected with the respective regions via elongated arrow heads (figure 1c). The bubble along with these heads form a hyperedge over m vertices denoting the sets. To fit multiple hyperedges in the inner area, a layout algorithm is needed to reduce bubble overlaps and edge crossings. Finding the optimal solution is an NP-complete problem [13]. Therefore, we use a greedy algorithm that employs a density map to place the bubbles. The algorithm iterates over the overlaps of degree k in descending order of their absolute or normalized sizes. For each overlap it creates a hyperedge centered at a point (x, y) in the map. The point is chosen so that the overall density at the pixels the hyperedge occupies is minimized. The densities at these pixels are increased to alleviate the overlap with hyperedges created in next iterations.

The design of the hyperedges intends to emphasize overlap sizes by mapping them to the bubble size. Bubbles are also appropriate for showing fractions of the overlaps to denote elements selected by the user (Sect. 3.2). The edge connecting a bubble with a region is plotted with decreasing thickness to reduce clutter. The varying thickness helps to some degree in visually separating overlapping hyperedges.

Density maps have also been used to create visual links that do not occlude the visualization [39]. The algorithm described above yields interactive performance for computing the placement of 100 hyper-

edges with a map resolution of 200×200 pixels. The bottleneck is rather its visual scalability: hyperedges are more complex objects than arcs. This imposes a severe limit on the number of hyperedges that can be visualized with sufficient readability. Figure 1c shows about 150 overlaps of 3rd degree, with the largest 10% overlaps accounting for 50% of the areas. The number and the shape complexity of the hyperedges potentially increase for overlaps of higher degree. This can rapidly increase the clutter even with a dozen sets. One way to avoid the clutter is to analyze the overlaps in a separate detail view (Sect. 3.2.4). Another way is to show the links of a hyperedge only for a few number of large overlaps, or only on demand as we explain next.

Visualizing overlaps as bubbles

Showing only the bubbles of the hyperedges described above results in a "bubble chart" of the overlaps. Pointing over a bubble reveals the links to the sets involved in the corresponding overlap. In case the histograms are scaled uniformly, the bubbles can be scaled using the same scaling factor. This facilitates perceiving an overlap in proportion of the involved sets. Alternatively, the bubbles can be scaled to fit in the inner area, to efficiently use this area in supporting the interaction with the bubbles and the comparison of their sizes (figure 7).

The compactness and the uniform shape of the bubbles allow showing overlaps of multiple degrees $2 \leq k \leq b$ at once by dividing the inner area into concentric rings. Starting from the outermost, each ring k contains bubbles that represent overlaps of degree $k+1$. A bubble can represent either all the elements in the overlap, or the elements exclusive to it (Eq. 5). The latter case avoids the redundancy of representing the same element in multiple overlaps. The former case allows comparing absolute overlap sizes across multiple degrees to analyze, for example, the satisfaction of increasing set membership requirements. Both color and interaction allow analyzing the exclusiveness of these overlaps and the intersections they exhibit between each other, as we explain next.

3.1.4 Visualizing information about the elements via color

Each arc, bubble, and histogram bar in Radial Sets represents a subset of the elements E whose size is encoded by its area or thickness. Further information about the elements in this subset can be communicated by coloring this area. When the user performs a select operation over the elements (Sect. 3.2), Radial Sets use color to depict selected fractions in each of the above-mentioned subsets. If no selection exists, the user can specify which information to encode via color.

By choosing an attribute of the elements as source of the color information, the user can gain an overview of the distribution of its values in the different subsets (figure 7). As we show in the usage scenarios (Sect. 4), this provides insights into how this attribute correlates with the elements' membership of different sets and overlaps (task **T6**).

Color can also be used to depict *relative information* about the subsets. As can be seen in figure 6, color reveals the disproportionality of the overlaps. Likewise, while the length of a histogram bar encodes the absolute size h_{jk} of the corresponding subset (Eq. 3), its color can encode the disproportionality of this subset, defined as follows:

$$disproportionality(h_{jk}) = \frac{h_{jk}}{|S_j|} - \frac{k \cdot |E_k|}{\sum_{j_2=1}^{m} |S_{j_2}|} \tag{8}$$

In the above equation, E_k is the set of elements of degree k:

$$E_k = \{e \in E : degree(e) = k\} \tag{9}$$

This disproportionality measure compares the actual histograms with the ones that would result if all histograms exhibit the same distribution [2]. This reveals, for example, which sets tend to have more (or less) exclusive elements or 2nd-degree overlaps than the other sets. The exclusiveness of an overlap (Eq. 5) can be analyzed by coloring its visual element by the average degree of its elements. An exclusive overlap receives a color that correspond to the overlap degree. Alternatively, the exclusiveness of an overlap can be analyzed via interaction, by selecting the elements E_k as we show next.

[2] See the supplemental materials for more explanation of this measure.

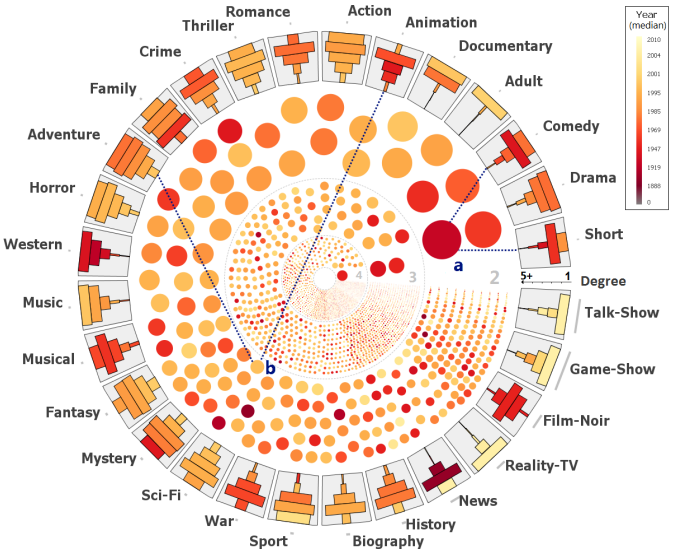

Fig. 7. Radial Sets depicting IMDb movies according to their genres. The bubbles encode the overlaps of degrees 2, 3, and 4 between the genres and are scaled to fit in the inner area. The area of a bubble encodes the normalized size of the overlap (Eq. 6). The color represents the median release date for the movies aggregated both in the bubbles and in the histograms. The sets involved in an overlap can be inferred by hovering over the respective bubble (a, b).

3.2 The Interactive Exploration Environment

The main user interface of Radial Sets comprises coordinated and multiple views that show information at different levels of detail. The *Radial Sets view* is the central part of the interface. The additional views show both summary and detailed information about the sets, the elements, and the overlaps. Together, these views enable formulating highly-expressive and visually-guided queries on the elements iteratively, and analyzing the query results in detail as we show next.

3.2.1 Summary views

Two views show summary information about the set system:

The *sets bar chart* depicts the set sizes $\{|S_{1 \leq j \leq m}|\}$ in descending order, along with the selected fractions of these sets (figure 1a). Since the sets can overlap, the bars do not sum up to the number of elements n, but to the number of their set memberships $\sum_{j=1}^{m} |S_j|$.

The *degree histogram* D (figure 1b) depicts a breakdown $\{|E_{0 \leq k \leq d}|\}$ of the set elements by their degrees (Eqs. 1, 9). The histogram bins sum up to the number of elements $n = \sum_{k=0}^{d} |E_k|$, with E_0 containing elements that belong to none of the sets of the set system. A sub histogram $D_{selected}$ depicts selected elements by their degrees.

Summary views are also essential to define which sets to depict in the Radial Sets view (show/hide) and which elements to incorporate in the computations (include/exclude). Furthermore, they are vital for gaining an overview on the elements under selection as well as for defining or refining the selection. Finally, both views are very useful for understanding the metaphor of Radial Sets as we explain next.

3.2.2 Radial Sets view

The Radial Sets view (figure 1c) can be thought of as a cross representation of both summary views: For each set S_j represented by a bar in the sets bar chart, Radial Sets show the breakdown of its elements by degree as a histogram H_j in the set's region (Eq. 3). When the selection is equal to S_j, the sub histogram $D_{selected}$ in the degree histogram is equal to H_j, assuming no aggregation of degrees, i.e. $b = d$ (Eq. 2).

The visual design of Radial Sets aims to provide an overview of a set system, emphasizing how the sets overlap and how the elements are distributed in them. More details about the elements and the overlaps can be obtained on demand either via tooltips or in the detail views.

Hovering the mouse pointer over a visual element in Radial Sets shows a tooltip with more information about the elements in the respective subset (figure 8). This comprises a short description of the subset, the absolute and relative sizes of the subset and of the selected fraction in it, and further statistics such as disproportionality or aggregated attribute values. More details about the individual elements in the subsets can be obtained using brushing and linking (Sect. 3.2.3).

In addition, the Radial Sets view supports direct manipulation to merge the sets or change their order using drag and drop. Merging two sets replaces them by their union and updates the visualization accordingly. The order of the sets can also be configured from the menu bar in the top of the view. The commands in this bar allow specifying color mappings (Sect. 3.1.4), histogram scaling, and overlaps' degree and sizes (absolute / relative). The selection commands allow manipulating the selected elements as we explain next.

3.2.3 Brushing the elements for details on demand

The Radial Sets view along with the summary views expose several subsets of the elements E in the set system. Brushing these subsets enables defining a selection over E. This selection can be specified iteratively using set operations to represent a variety of combinations of these subsets. This allows a highly expressive selection of elements by their set memberships and degrees. Furthermore, the selected fractions depicted in Radial Sets and in the summary views are updated during the iterative selection. This gives an immediate feedback to the user on how the selected elements belong to the different sets and overlaps, and offers guidance on how to refine this selection[3].

Brushing the elements in a set region can be performed either by clicking on the individual bars or by defining a range over the degree axis using mouse dragging. Similar interactions are possible in the summary views and with the overlaps. If no keyboard modifier is active during the brushing operation, the selection is set to the newly brushed elements. Specific keyboard modifier can be used to specify if the brushed elements should be added to (set union), intersected with, or subtracted from the existing selection. In addition to defining the selection based on set memberships, the elements can be selected based on their attribute values. Radial Sets supports this both via textual search in the attribute values (figure 1f), or via coordinated views that enable brushing elements having certain attribute values.

The selection view shows detailed information about the selected element (figure 1d). The top of this view shows a formula that details how the selection was specified. The formula text is composed using the common set-theory notation, with extensions to express further conditions on the elements' degrees and attribute values. The body of the selection views is a tabular list of the elements in the selection, showing their attribute values. The list can be sorted by one of these attributes. These attributes can also be analyzed in detail via additional views (figure 1g). Clicking on an element in the tabular list highlights this element and shows its set memberships both graphically and in text. The text is shown at the bottom of the selection view as a comma-separated list of these set memberships. Additionally, these memberships are indicated graphically as a star graph over the Radial Sets view. This graph shows in which region and in which bars in these regions the highlighted element is present.

Besides gaining details into specific elements, interactive selection is also useful for filtering and manipulating the data. it can be used to hide or exclude certain elements from the analysis based on their attributes, degrees, and set memberships. This is useful for dealing with real-world datasets that often exhibit highly skewed distributions of set sizes (few sets comprise the majority of the elements) or of element degrees (most elements are exclusive in their sets). Filtering out such elements reveals finer details about the rest of the data.

The expressive power of the interactive selection possibilities and the immediate feedback on selected fractions in Radial Sets, enable an elaborate analysis of the set memberships and the attribute values of certain elements in the set system (task **T7**). These possibilities constitute a visual query language for set-typed data. This language covers all possible 2^m overlaps between the sets, and goes beyond by allowing the selection of exclusive parts of these overlaps, parts having specific degrees, or parts containing certain values for selected attributes. Furthermore, the memberships of selected elements in different overlaps can be analyzed in detail as we explain next.

[3]The supplementary video demonstrates the interactive selection of elements in Radial Sets in detail.

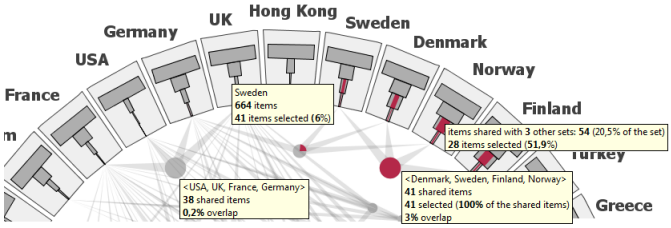

Fig. 8. Tooltips showing various information about the subsets represented by the regions, the bars and the links in Radial Sets.

3.2.4 Overlap analysis view

The arcs and bubbles in Radial Sets give a compact overview of existing overlaps and the sets involved in them. They are also suited to revealing overlap patterns such as clusters of highly overlapping sets and to quickly select a specific overlap. To analyze and compare the overlaps in more detail, Radial Sets employ a coordinated view that shows these overlaps in tabular lists (figure 1e). Each list L_k in this *overlap analysis view* contains the overlaps of a specific degree $k \geq 2$. An additional list L_1 contains the sets like in the summary sets bar chart (Sect. 3.2.1), along with further statistics about the sets. For each overlap $O_{\{j_1,\ldots,j_k\}}$, the list L_k textually shows the sets $\{S_{j_1},\ldots,S_{j_k}\}$ involved in this overlap, separated by commas and ordered by their order in L_1. Additionally, L_k can show the absolute and normalize sizes of the overlap, the fractions of selected elements in it, a summary value of the color attribute in the whole overlap and in the selected portion, and the disproportionality of the overlap and of its selected portion. These statistics can be shown either textually or graphically using color and/or bar charts. The overlaps list L_k can be sorted according to these statistics. This enables a detailed analysis of the overlaps in the lists and quickly finding large or overrepresented overlaps at different degrees, without having space limitations or clutter issues.

The overlap analysis view is interactively updated when the selection changes. Also, the Radial Sets view is updated when an overlap in one of the lists $L_{k\geq 2}$ is clicked: In case the view already includes a visual element for this overlap, it becomes highlighted. Otherwise, a new visual element is overlaid in the Radial Sets view to indicate involved sets and the size of the overlap in proportion to them.

4 Usage Scenarios

To demonstrate Radial Sets, we report insights we gained in two real-world set-typed datasets using some of the features described in Sect. 3. The datasets are of different scales and skewness, and deal resp. with multi-label classifications and with multi-valued attributes.

4.1 ACM Paper Classification

The ACM digital library comprises computer science papers tagged with multiple index terms from the ACM classification system [1]. We define a set system over a collection of more than 50,000 ACM papers extracted by Santos and Rodrigues in 2008 [34] . The sets of this system are the top-level index terms (*A.* to *K.*), also called classes. Figure 3b depicts the Radial Sets of these index terms. Each histogram bar is colored by the median publication date of the papers it represents. The arcs depict the overlaps between the index terms, with thickness and color representing the normalized size (Eq. 6) and disproportionality (Eq. 7) respectively. From the histograms it can be easily seen that the index terms vary in their exclusiveness: few computer-science papers are exclusive to class G (Mathematics of Computing); while 92.2% of the papers in this class have other index terms. On the contrary, 42% of "Hardware" papers did not have other terms assigned.

It is also noticeable that the index terms vary in the recency of their papers, indicated by the median publication date. The median date varies between 1994 (classes F and G) and 2001 (classes C and E). Also, papers that belong to one class tend to be more recent than papers that belong to multiple classes, with medians at 2003 and 1997 respectively. This variance can be easily inferred by coloring the bars in the summary charts (Sect. 3.2.1) with the median dates. However, by examining the Radial Sets view, finer details about this variance can be observed, compared to the summary views. For example, contrary to the global trend, papers exclusive to class G have a median date of 1984, which is significantly older than the class median 1994. On the other hand, while class J has also a relatively old median date of 1995, the small fraction of papers exclusive to it have a very recent median date of 2005. A similar contrast between exclusive and shared papers is noticeable in class C. To verify the above observations, we plot the distribution of publication date in each of these paper classes as histograms, along with sub-histograms that represent the papers exclusive to them (figure 9). This confirms the recency trend of class C with exclusive papers in this class being an increasing trend, constituting 67% of the papers in 2007 (up from 10% in 2000). A similar observation holds for papers exclusive to class J: they started to appear in 2002, and made up 40% of "Computer Application" papers in 2007. To get more details about these papers, we select them in the Radial Sets view and examine the venues they were published in using the detail view (Sect. 3.2.3). Most of them were published in conference series that started in the past decade on topics like "mobile computing", "genetic and evolutionary computation", "electronic governance", "future play", and "advances in computer entertainment technology" to mention a few. The long tradition of class G is observable, with papers exclusive to it being an old trend that disappears in the 1990s and reappears in the past decade. To investigate this trend, we select the G-exclusive papers whose publication dates are newer than 2000 and observe their venues. While some of these venues are recent like "Symbolic-Numeric Computation", the majority of them are established yearly conferences that were started in the 1980s or earlier on topics like "symbolic and algebraic computation", "theory of computing", "computational geometry", "parallelism in algorithms and architectures" and "supercomputing". By searching for all papers of these conferences in the dataset and examining their publication dates, we consistently found full or large gaps in the 1990s. This explains the gap we observed for the G-exclusive papers (figure 9) and reveals a sampling bias in the dataset.

The insights gained so far are focused on set-membership tasks (**T1** and **T2**) and attribute-analysis tasks (**T6** and **T7**). To explicitly analyze set overlaps (tasks **T3**, **T4** and **T5**), we observe the arcs in figure 3b and the hyperedges in figure 1c. From the arcs we immediately notice a significant overlap between "Mathematics of Computing" and "Theory of Computing". This overlap constitutes 15.5% of the union of these classes; up from 5% expected overlap in case of statistical independence. Many other disproportionally-high overlaps are noticeable such as "Information Systems" $\cap$ "Computer Methodologies" and "Hardware" $\cap$ "Computer Systems Organization". On the other hand, there are classes that exhibit only a small overlap such as "Hardware" $\cap$ "Information Systems". By examining the 207 papers in this overlap in the detail views, we observe that many of them were published in conferences on "Design Automation" (40 papers), "Human Factors in Computing" (24), and "Management of Data" (19).

Hyperedges with large bubbles in figure 1c indicate significant overlap between three classes, such as $D \cap H \cap K$ and $F \cap G \cap H$. In this figure, papers having "Human Factor" in their general terms are selected, comprising about 19.6% of the dataset. The bubbles are colored to indicate selected fractions in the overlaps. Certain overlaps have disproportionally-large selected fractions. For example, 66% of papers on "Computing Milieux", "Computing Methodologies", and "Information Systems" address issues of "Human Factors". This ratio is higher in the overlap than in its individual classes, as can be observed in the summary view (figure 1a). These papers were published in conferences like ACM CHI (48 papers), SIGACCESS (40), SIGCSE (26) SIGGRAPH (16), and IUI (9).

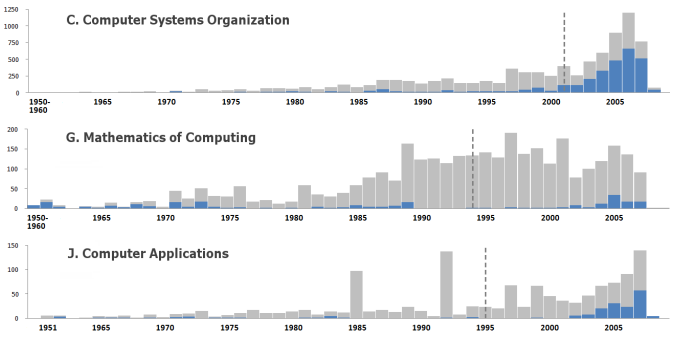

Fig. 9. The number of papers over time for three classes in the ACM digital library. Exclusive papers in each class are highlighted in blue. The dashed lines indicate the median publication date in each class.

4.2 IMDb Movies

Information about movies comprises several multi-valued attributes such as genres, production countries, and languages. To illustrate the insights gained by Radial Sets in such attributes, we consider two set systems that can be defined over a 2010 snapshot of the IMDb database [2] comprising over 525,000 movies.

The sets of the first system are the top 35 production countries of the movies. The sets exhibit a large skewness in their sizes with the US being involved in 38% of the movies, followed by the UK (7.7%). The smallest sets are East Germany and Russia, each involved in about 0.4% of the movies. Another large skewness exist in the distribution of the element degrees: 96% of the movies were produced in one country. These elements do not contribute to any overlaps, and hence are less important for analyzing co-production patterns between the countries. Including them obscures finer information about the overlaps. Similarly, very few movies (0.03%) were produced in five or more countries, with only one movie having the largest element degree of 13. Therefore, we group elements of degree ≤ 5 to increase the resolution of the histogram bars. Depicting absolute values in the histograms will assign the majority of the available space to the few top-5 countries and obscure the rest of the data. Therefore, we assign the regions equal areas to enable relative comparison of the distributions in these histograms (figure 6). This reveals a variety of patterns in the data: pairs of countries that produced *relatively* more joint movies than other pairs become visible (**T4**). Such countries often have a common language or a common border. The ordering algorithm reveals groups of countries that exhibit high mutual overlaps, most noticeably the Scandinavian countries. By checking the 4^{th}-degree overlaps in the overlap-analysis view, we immediately notice that 41 movies were produced jointly by all of Denmark, Finland, Norway, and Sweden, making this the largest overlap of 4^{th} degree (**T3**). Figure 8 shows the absolute sizes of these overlaps graphically using hyperedges. The 2^{nd}-largest overlap is between USA, UK, France and Germany, the four largest sets comprising 56.5% of all movies. This points to a very significant disproportionality of the Scandinavian overlap, given the small sizes of the involved sets (summing up to 3.5% of all movies).

The sets of the second systems are the 28 IMDb movie genres. Figure 7 depicts the Radial Sets of the genres set system. The bubbles in the different rings represent normalized overlaps of degree 2, 3, and 4. We also employ relative analysis both for the histograms and for the bubbles due to the high skewness between the set sizes. We easily notice that the genres vary in their exclusiveness (**T1** and **T2**): 94.1% of Animation movies had other genres, whereas 93.1% of Adult movies were exclusive to this genre. The elements are colored by the median release date of the movies they represent. This reveals a significant variance in the recency of the genres and their combinations (**T6**). For example, movies exclusive to Mystery were predominantly old (median date 1944), whereas Mystery movies that have other genres are more recent (median 1988). The opposite holds for genre News. This is revealed by contrasting the first bar with the other bars in the regions.

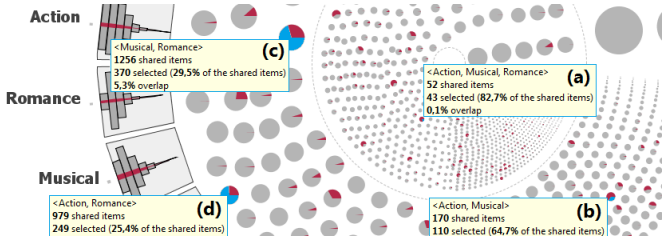

Fig. 10. Selected genre overlaps that exhibit disproportionately high presence of Indian movies (highlighted in red).

The combination Comedy ∩ Short contains mainly the older movies (median 1926) from both genres (figure 7a) that have individually more recent median dates (1967 and 1966). In contrast, Animation ∩ Adventure contains mainly the newer movies (median 1997) from both genres (figure 7b) that have older median dates (1966 and 1986).

The two set systems (countries and genres) can be analyzed against each other to find disproportionalities in the overlaps (**T7**). For example, while Indian movies comprise 2% of the dataset, selecting them in the Radial Sets of Genres reveals higher percentages in specific overlaps (Fig 10). In particular, these movies comprise 83% of Musical ∩ Action ∩ Romance (figure 10a), and 65% of Musical ∩ Action (figure 10b). The other two pairs of these three genres (figure 10c-d) exhibit less percentages. These findings can be analyzed and compared against each other in more details in the overlap analysis view.

5 Discussion

Radial Sets build upon and extend several ideas from state-of-the-art techniques to enable advanced visual analysis of large overlapping sets. Our technique extends the frequency-based aggregation of Set'o'grams [18], which accounts for high scalability in the number of elements of the set system. Also, it uses separate visual elements for the sets and for the overlaps, similar to the untangled Euler diagrams [21]. The hyperedges between radially-arranged regions are inspired from the free nodes in Anchored Maps [31]. The radial layout is adopted from Contingency Wheel++ [4] which was designed to visualize skewed contingency tables having few columns but a large number of rows. These tables have a similar structure and dimensionality as the elements-set membership matrix. Nevertheless, Radial Sets use different aggregation for the elements in the histograms and in the overlaps, and introduce additional visual elements to address the characteristics of set data and support the tasks specific to them.

The visual design of Radial Sets is a compromise between information richness and effectiveness. For example, an $m \times m$ heatmap can be more effective at showing the 2nd-degree overlaps than crossing arcs with a limited range of varying thicknesses. Also, standard bar charts are more precise at showing the elements by degree in each set than non-aligned bars depicted in radially-arranged parallelograms. Finally, color is sub-optimal for showing the values of an attribute in the elements aggregated in a bar or in a bubble. Nevertheless, depicting all this information together enables gaining a high-level overview of the distributions of the elements, the overlaps, and the attributes, in relation to each other. Using separate visualizations such as an overlap matrix, element histograms and attribute histograms makes it harder to visually link between related elements. Our interaction techniques allow certain elements in Radial Sets to be investigated at greater detail on demand using simpler and more precise visualizations. Hence, Radial Sets serve as a starting point of the analysis and as a means to detect extreme differences and to quickly formulate queries to select these elements. However, in an informal pilot feedback session with 10 engineers from different disciplines, three subjects reported that Radial Sets of movies are showing too much information at once. One of them recommended showing the arcs for one selected set only. Nevertheless, the subjects were able to interpret the visual metaphor correctly and use interaction to perform set operations on the elements and to answer questions on the relations between the sets.

The visual complexity of Radial Sets imposes a limit on the number of sets it can depict. For example, using the 2nd-level classes of the ACM classification in Sect. 4.1 results in 89 sets, each receiving 4 degrees of the angular resolution on average. A higher angular resolution is needed to ensure a sufficient readability of the histograms, which limits the number of sets that can be visualized at once to about $m \leq 30$. On the other hand, Radial Sets can handle a large number of elements, at the order of 1 million, thanks to the frequency-based aggregations and to the relative analysis possibilities. For example, figure 7 depicts information about 525,000 movies using non-uniform scaling for the histograms and normalized sizes for the bubbles.

Another limitation is the number of hyperedges that can be visualized at once being ≤ 100 (assuming a normal distribution of overlap sizes), which is only 2% of all possible 3rd-degree overlaps between 30 sets. The remaining possibilities in these cases are to show the bubbles only or to analyze the overlaps separately in the detail view.

Finally, using separate visual representations for the overlaps and for the sets hinders the depiction of containment relations between the sets. Such relations are pre-attentive in Euler diagrams [26], even in a composite form such as $S_1 \subset S_2 \cup S_3$ or $O_{\{1,2\}} \subset S3$. Also, both the arcs and the bubbles show the absolute or normalized size of an overlap, without indicating the different fractions it constitutes in the involved sets. This information needs to be investigated on demand by selecting this overlap and checking these fractions individually.

Future Work One way to compensate for the visual limitations of Radial Sets and their low sensitivity to small differences between attribute values or overlap sizes is to employ complementary computational methods. These methods can pre-compute significant disproportionalities in the overlaps and in attribute distributions among all elements or in selected subsets. We are investigating statistics-based and computationally-efficient measures for this purpose along with possibilities to communicate their results visually, and steer the calculation interactively. Additionally, we are considering different placements of histograms and hyperedges to address the perceptual issues of the current layout as well as alternative visual representations based on heatmaps to visualize the same information in a more scalable way in the number of sets. Finally, to confirm our informal findings on the understandability of our visual design, we are currently conducting a formal evaluation of Radial Sets that will assess how well they support the tasks intended in comparison with other alternatives.

6 Conclusion

Radial Sets is a novel interactive technique for the visual analysis of large overlapping sets, designed to provide insights into different kinds of overlaps between the sets. These overlaps are salient features of set-typed data and are central to relevant analysis tasks. Radial Sets builds upon selected ideas from existing techniques to support these tasks in a scalable way using several aggregation methods and a multi-level overview+detail exploration environment. In particular, our technique enables (1) gaining insights into different kinds of overlaps between the sets and into the disproportionalities they represent, (2) analyzing the element memberships of the sets and the overlaps in relation to other attributes of the elements, and (3) interactively querying the elements by their set memberships and attribute values, and analyzing how selected elements differ from the rest of the elements in their memberships of the sets and the overlaps. As the usage scenarios demonstrate, Radial Sets enable conducting elaborate analysis workflows in large set-typed data using expressive visual queries. These queries allow set-theoretic operations to select and analyze specific elements of interest in the data. Compared with existing visual representations, Radial Sets offer richer information in the overview but at lower precision and sensitivity to small differences. Nevertheless, using interaction and complementary views, Radial Sets reveal a variety of overlapping patterns in large overlapping sets, beyond the limits of state-of-the-art techniques.

Acknowledgement: This work was supported by the Austrian Federal Ministry of Economy, Family and Youth via CVAST, a Laura Bassi Centre of Excellence (No. 822746).

REFERENCES

[1] The ACM computing classification system [1998 version]. http://www.acm.org/about/class/1998. accessed: March 2013.

[2] The IMDB database [snapshot in sept. 2012]. http://www.imdb.com/interfaces. accessed: March 2013.

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

[4] B. Alsallakh, W. Aigner, S. Miksch, and M. E. Gröller. Reinventing the contingency wheel: Scalable visual analytics of large categorical data. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2849–2858, 2012.

[5] J. Bertin. *Graphics and graphic information processing*. de Gruyter, 1981.

[6] J. Bertin and M. Daru. Matrix theory of graphics: Jacques Bertin's theories. *Information Design Journal*, 10(1):5–19, 2000.

[7] C.-H. Chen. Generalized association plots: Information visualization via iteratively generated correlation matrices. *Statistica Sinica*, 12(1):7–30, 2002.

[8] S. C. Chow. *Generating and drawing area-proportional Euler and Venn diagrams*. PhD thesis, University of Victoria, 2007.

[9] W. Cleveland. *The elements of graphing data*. AT&T Bell Laboratories, 1994.

[10] C. Collins, G. Penn, and S. Carpendale. Bubble sets: Revealing set relations with isocontours over existing visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1009–1016, 2009.

[11] E. Di Giacomo, L. Grilli, and G. Liotta. Drawing bipartite graphs on two curves. In *Graph Drawing*, pages 380–385. Springer, 2007.

[12] K. Dinkla, M. van Kreveld, B. Speckmann, and M. Westenberg. Kelp diagrams: Point set membership visualization. In *Computer Graphics Forum*, volume 31, pages 875–884. Wiley Online Library, 2012.

[13] P. Eades and N. C. Wormald. Edge crossings in drawings of bipartite graphs. *Algorithmica*, 11(4):379–403, 1994.

[14] N. Elmqvist, T.-N. Do, H. Goodell, N. Henry Riche, and J.-D. Fekete. ZAME: Interactive large-scale graph visualization. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 215–222. IEEE, 2008.

[15] L. Euler. *Lettres à une princesse d'Allemagne sur divers sujets de physique et de philosophie*, volume 1 letters no. 102-108. Courcier, 1772.

[16] J. Flower, A. Fish, and J. Howse. Euler diagram generation. *Journal of Visual Languages & Computing*, 19(6):675–694, 2008.

[17] J. Flower and J. Howse. Generating Euler diagrams. *Diagrammatic Representation and Inference*, pages 285–285, 2002.

[18] W. Freiler, K. Matkovic, and H. Hauser. Interactive visual analysis of set-typed data. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1340 – 1347, Nov. 2008.

[19] M. Ghoniem, J.-D. Fekete, and P. Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE Symposium on Information Visualization (INFOVIS)*, pages 17–24. IEEE, 2004.

[20] H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *IEEE Symposium on Information Visualization (INFOVIS)*, pages 127–130. IEEE, 2002.

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

[22] N. Henry Riche and J.-D. Fekete. MatrixExplorer: a dual-representation system to explore social networks. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):677–684, 2006.

[23] N. Henry Riche and J.-D. Fekete. MatLink: Enhanced matrix visualization for analyzing social networks. *Human-Computer Interaction–INTERACT 2007*, pages 288–302, 2007.

[24] T. Itoh, C. Muelder, K.-L. Ma, and J. Sese. A hybrid space-filling and force-directed layout method for visualizing multiple-category graphs. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 121–128, 2009.

[25] G. Kanizsa and W. Gerbino. Convexity and symmetry in figure-ground organization. *Vision and artifact*, pages 25–32, 1976.

[26] H. Kestler, A. Müller, J. Kraus, M. Buchholz, T. Gress, H. Liu, D. Kane, B. Zeeberg, and J. Weinstein. VennMaster: area-proportional Euler diagrams for functional GO analysis of microarrays. *BMC bioinformatics*, 9(1):67, 2008.

[27] I. Liiv. Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining*, 3(2):70–91, 2010.

[28] E. Mäkinen. How to draw a hypergraph. *International Journal of Computer Mathematics*, 34(3-4):177–185, 1990.

[29] E. Mäkinen and H. Siirtola. Reordering the reorderable matrix as an algorithmic problem. *Theory and Application of Diagrams*, pages 453–468, 2000.

[30] W. Meulemans, N. Henry Riche, B. Speckmann, B. Alper, and T. Dwyer. KelpFusion: a hybrid set visualization technique. *Visualization and Computer Graphics, IEEE Transactions on*, 2013. to appear.

[31] K. Misue. Drawing bipartite graphs as anchored maps. In *Proceedings of the Asia-Pacific Symposium on Information Visualisation (APVIS)*, pages 169–177. Australian Computer Society, Inc., 2006.

[32] M. Newton, O. Sỳkora, and I. Vrto. Two new heuristics for two-sided bipartite graph drawing. In *Graph Drawing*, pages 465–485. Springer, 2002.

[33] P. Rodgers, L. Zhang, and A. Fish. General Euler diagram generation. *Diagrammatic Representation and Inference*, pages 13–27, 2008.

[34] A. P. Santos and F. Rodrigues. Multi-label hierarchical text classification using the acm taxonomy. *14th Portuguese Conference on Artificial Intelligence (EPIA)*, pages 553–564, 2009.

[35] H.-J. Schulz, M. John, A. Unger, H. Schumann, et al. Visual analysis of bipartite biological networks. In *Eurographics Workshop on Visual Computing for Biomedicine*, 2008.

[36] H. Siirtola and E. Mäkinen. Constructing and reconstructing the reorderable matrix. *Information Visualization*, 4(1):32–48, 2005.

[37] P. Simonetto and D. Auber. Visualise undrawable Euler diagrams. In *12th International Conference Information Visualisation (IV)*, pages 594–599. IEEE, 2008.

[38] P. Simonetto, D. Auber, and D. Archambault. Fully automatic visualisation of overlapping sets. *Computer Graphics Forum*, 28(3):967–974, 2009.

[39] M. Steinberger, M. Waldner, M. Streit, A. Lex, and D. Schmalstieg. Context-preserving visual links. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2249–2258, 2011.

[40] J. Talbot, B. Lee, A. Kapoor, and D. S. Tan. EnsembleMatrix: Interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 1283–1292. ACM, 2009.

[41] J. Venn. On the diagrammatic and mechanical representation of propositions and reasonings. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 10(59):1–18, 1880.

[42] A. Verroust and M.-L. Viaud. Ensuring the drawability of extended Euler diagrams for up to 8 sets. *Diagrammatic Representation and Inference*, pages 271–281, 2004.

[43] M. Wertheimer. Laws of organization in perceptual forms. In W. D. Ellis, editor, *A sourcebook of Gestalt psychology*, pages 71–88. Routledge and Kegan Paul, 1938.

[44] L. Wilkinson. Exact and approximate area-proportional circular Venn and Euler diagrams. *Visualization and Computer Graphics, IEEE Transactions on*, 18(2):321–331, 2012.

[45] L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[46] K. Wittenburg, T. Lanning, M. Heinrichs, and M. Stanton. Parallel bargrams for consumer-based information exploration and choice. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 51–60. ACM, 2001.

[47] K. Wittenburg, A. Malizia, L. Lupo, and G. Pekhteryev. Visualizing set-valued attributes in parallel with equal-height histograms. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI)*, pages 632–635. ACM, 2012.

[48] D. F. Wyatt. http://www-edc.eng.cam.ac.uk/tools/set_visualiser/. accessed: March 2013.

[49] J. S. Yi, R. Melton, J. Stasko, and J. A. Jacko. Dust & magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, 4(4):239–256, 2005.

[50] L. Zheng, L. Song, and P. Eades. Crossing minimization problems of drawing bipartite graphs in two clusters. In *Proceedings of the Asia-Pacific Symposium on Information Visualisation (APVIS)*, pages 33–37. Australian Computer Society, 2005.