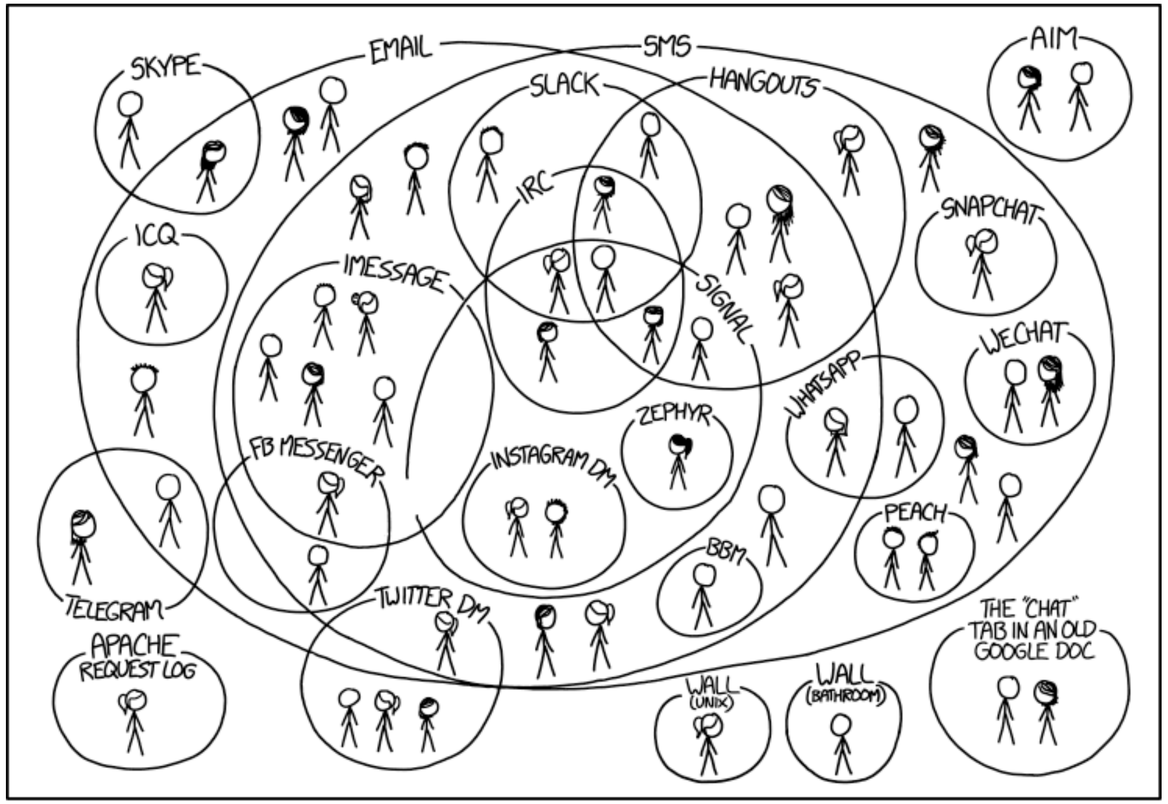
CS-5630 / CS-6630 Visualization for Data Science
Set Visualization

Alexander Lex
alex@sci.utah.edu

THE UNIVERSITY OF UTAH

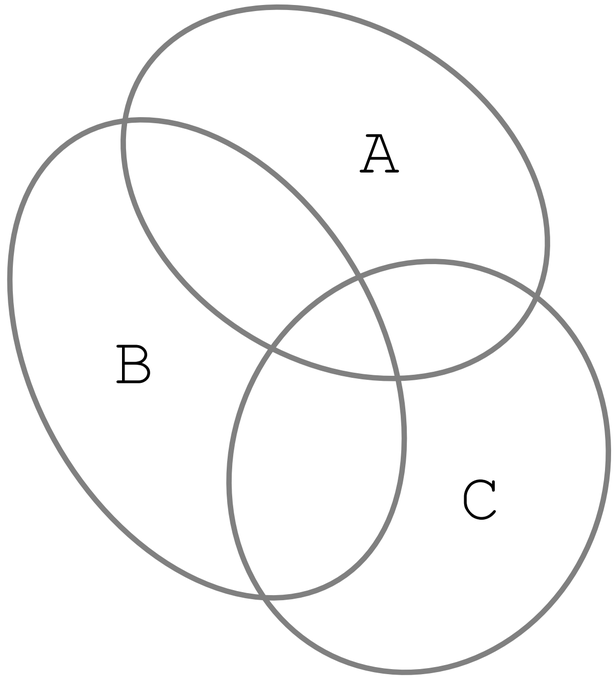


I HAVE A HARD TIME KEEPING TRACK OF WHICH CONTACTS USE WHICH CHAT SYSTEMS.

[xkcd]

Design Workshop

```
item1 : A
item2 : A
item3 : A, B
item4 : A, C
item5 : A, B, C
item6 : B
item7 : B, C
item8 : C
…
```



Venn diagram

LETTER

doi:10.1038/nature11241

The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants

Angélique D'Hont[1]*, France Denoeud[2,3,4]*, Jean-Marc Aury[2], Franc-Christophe Baurens[1], Françoise Carreel[1,5], Olivier Garsmeur[1], Benjamin Noel[2], Stéphanie Bocs[1], Gaëtan Droc[1], Mathieu Rouard[6], Corinne Da Silva[2], Kamel Jabbari[2,3,4], Céline Cardi[1], Julie Poulain[2], Marlène Souquet[1], Karine Labadie[2], Cyril Jourda[1], Juliette Lengellé[1], Marguerite Rodier-Goud[1], Adriana Alberti[2], Maria Bernard[2], Margot Correa[2], Saravanaraj Ayyampalayam[7], Michael R. Mckain[7], Jim Leebens-Mack[7], Diane Burgess[8], Mike Freeling[8], Didier Mbéguié-A-Mbéguié[9], Matthieu Chabannes[5], Thomas Wicker[10], Olivier Panaud[11], Jose Barbosa[11], Eva Hribova[12], Pat Heslop-Harrison[13], Rémy Habas[5], Ronan Rivallan[1], Philippe Francois[1], Claire Poiron[1], Andrzej Kilian[14], Dheema Burthia[1], Christophe Jenny[1], Frédéric Bakry[1], Spencer Brown[15], Valentin Guignon[1,6], Gert Kema[16], Miguel Dita[19], Cees Waalwijk[16], Steeve Joseph[1], Anne Dievart[1], Olivier Jaillon[2,3,4], Julie Leclercq[1], Xavier Argout[1], Eric Lyons[17], Ana Almeida[8], Mouna Jeridi[1], Jaroslav Dolezel[12], Nicolas Roux[6], Ange-Marie Risterucci[1], Jean Weissenbach[2,3,4], Manuel Ruiz[1], Jean-Christophe Glaszmann[1], Francis Quétier[18], Nabila Yahiaoui[1] & Patrick Wincker[2,3,4]

Bananas (*Musa* spp.), including dessert and cooking types, are giant perennial monocotyledonous herbs of the order Zingiberales, a sister group to the well-studied Poales, which include cereals. Bananas are vital for food security in many tropical and subtropical countries and the most popular fruit in industrialized countries[1]. The *Musa* domestication process started some 7,000 years ago in Southeast Asia. It involved hybridizations between diverse species and subspecies, fostered by human migrations[2], and selection of diploid and triploid seedless, parthenocarpic hybrids thereafter widely dispersed by vegetative propagation. Half of the current production relies on somaclones derived from a single triploid genotype (Cavendish)[1]. Pests and diseases have gradually become adapted, representing an imminent danger for global banana production[3,4]. Here we describe the draft sequence of the 523-megabase genome of a *Musa acuminata* doubled-haploid genotype, providing a crucial stepping-stone for genetic improvement of banana. We detected three rounds of whole-genome duplications in the *Musa* lineage, independently of those previously described in the Poales lineage and the one we detected in the Arecales lineage. This first monocotyledon high-continuity whole-genome sequence reported outside Poales represents an essential bridge for comparative genome analysis in plants. As such, it clarifies commelinid-

sequence errors. The assembly consisted of 24,425 contigs and 7,513 scaffolds with a total length of 472.2 Mb, which represented 90% of the estimated DH-Pahang genome size. Ninety per cent of the assembly was in 647 scaffolds, and the N50 (the scaffold size above which 50% of the total length of the sequence assembly can be found) was 1.3 Mb (Supplementary Text and Supplementary Tables 1–3). We anchored 70% of the assembly (332 Mb) along the 11 *Musa* linkage groups of the Pahang genetic map. This corresponded to 258 scaffolds and included 98.0% of the scaffolds larger than 1 Mb and 92% of the annotated genes (Supplementary Text, Supplementary Table 4 and Supplementary Fig. 1).

We identified 36,542 protein-coding gene models in the *Musa* genome (Supplementary Tables 1 and 5). A total of 235 microRNAs from 37 families were identified, including only one of the eight microRNA gene (*MIR*) families found so far solely in Poaceae[8] (Supplementary Tables 6 and 7).

Viral sequences related to the banana streak virus (BSV) dsDNA plant pararetrovirus were found to be integrated in the Pahang genome, with 24 loci spanning 10 chromosomes (Supplementary Text and Supplementary Fig. 2). They belonged to a badnavirus phylogenetic group that differed from the endogenous BSV species (eBSV) found in *M. balbisiana*[9] and most of them formed a new

Nature 2012

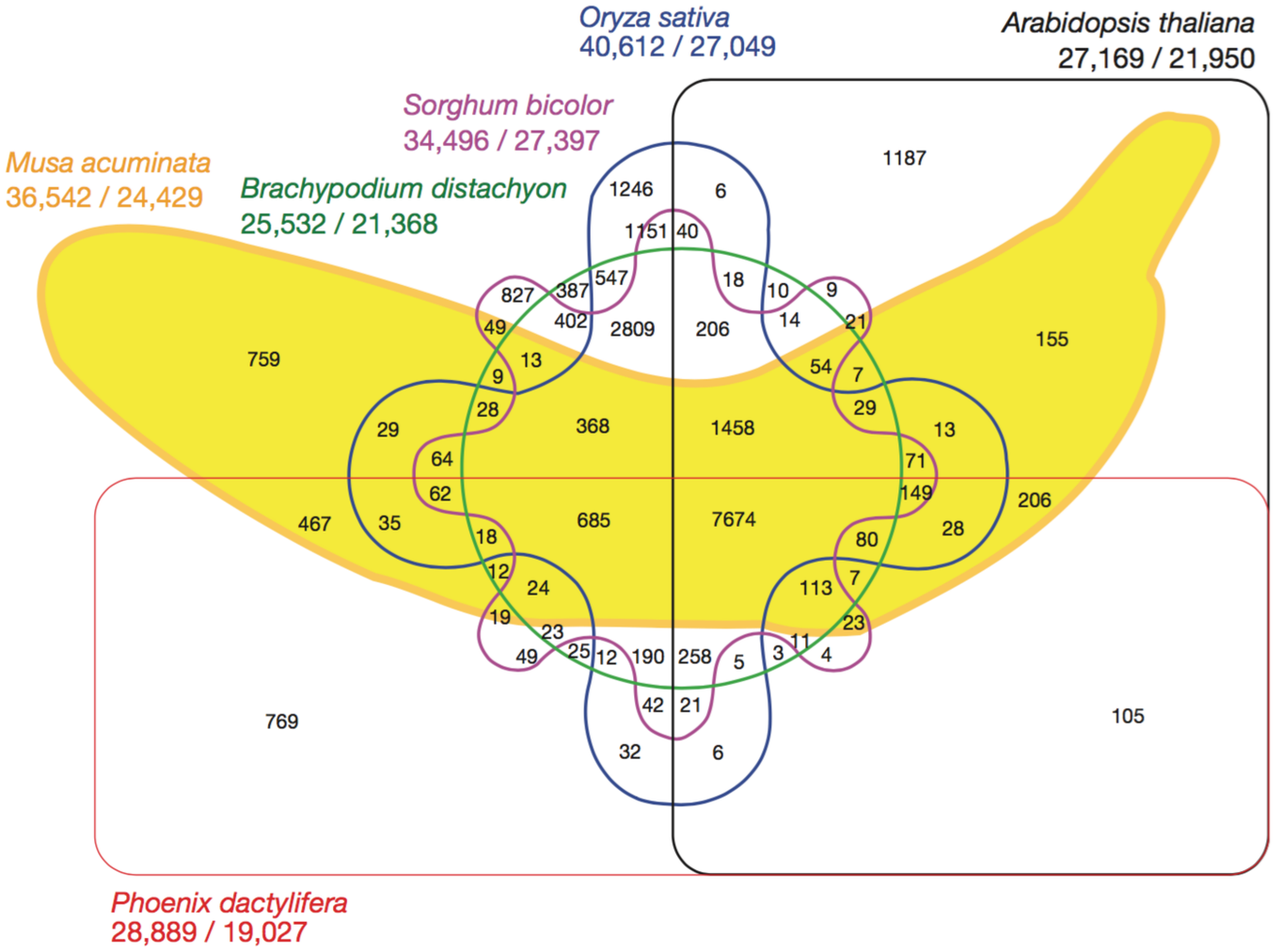
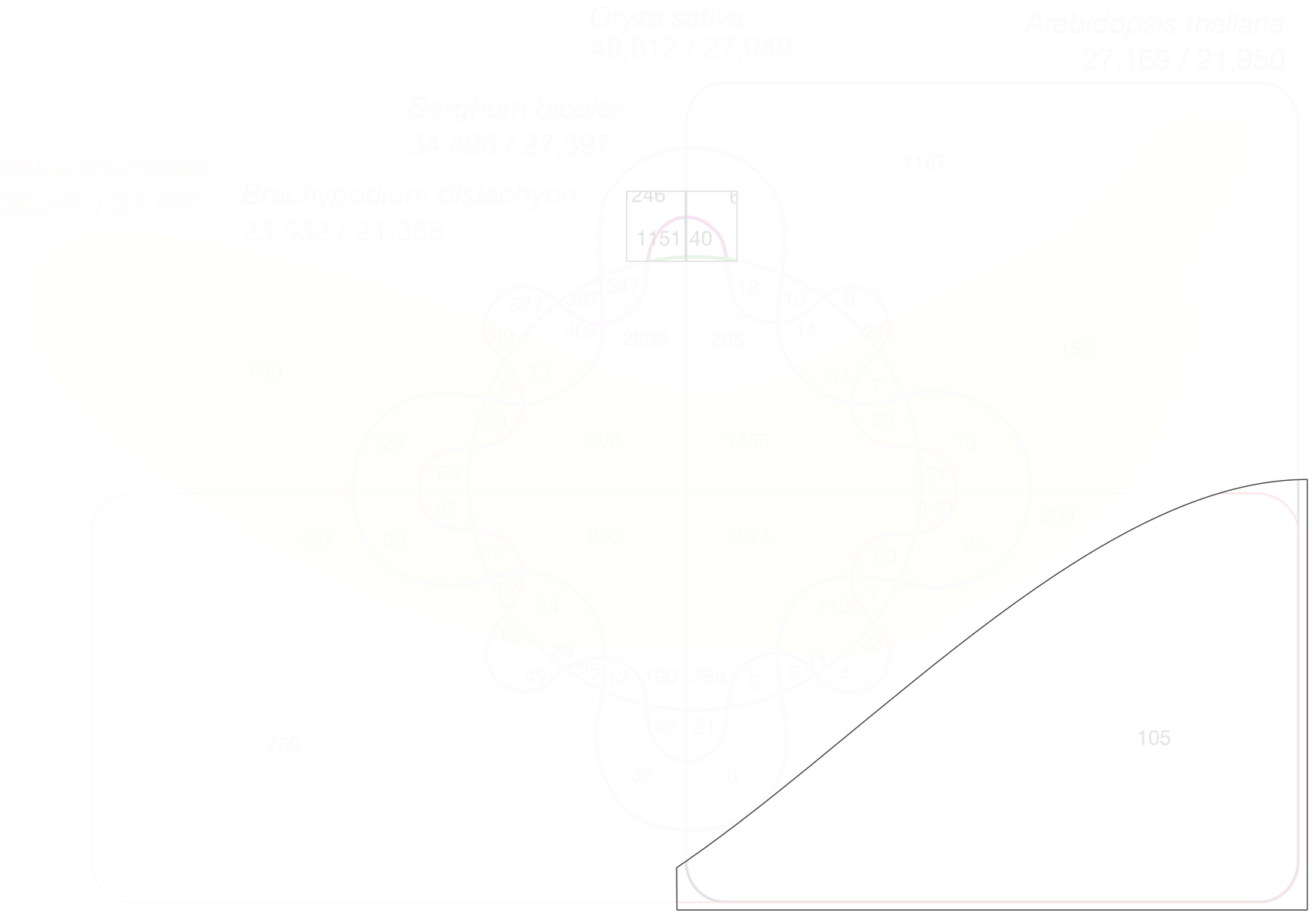
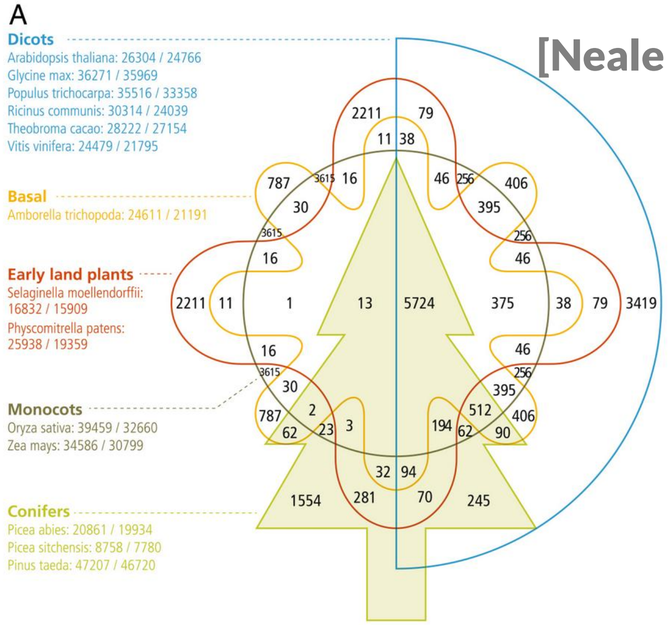
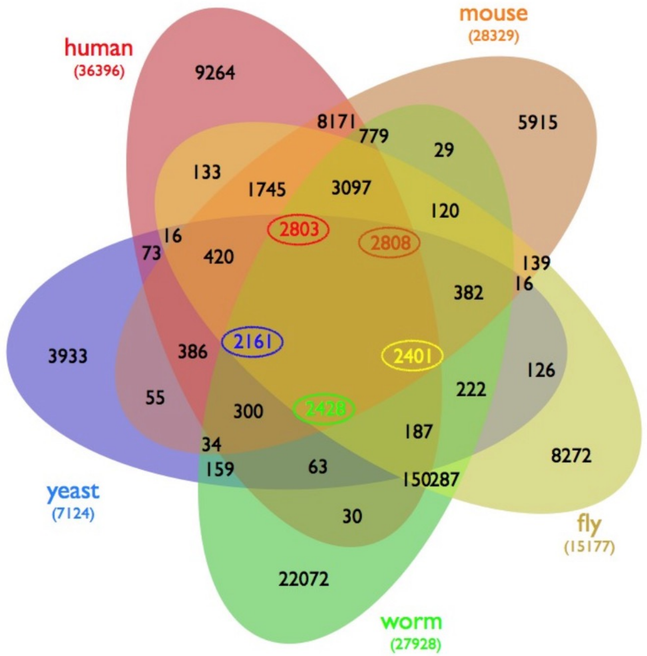


Figure 4 | Six-way Venn diagram showing the distribution of shared gene families (sequence clusters) among *M. acuminata*, *P. dactylifera*, *Arabidopsis thaliana*, *Oryza sativa*, *Sorghum bicolor* and *Brachypodium distachyon* genomes. Numbers of clusters are provided in the intersections. The total number of sequences for each species is provided under the species name (total number of sequences/total number of clustered sequences).

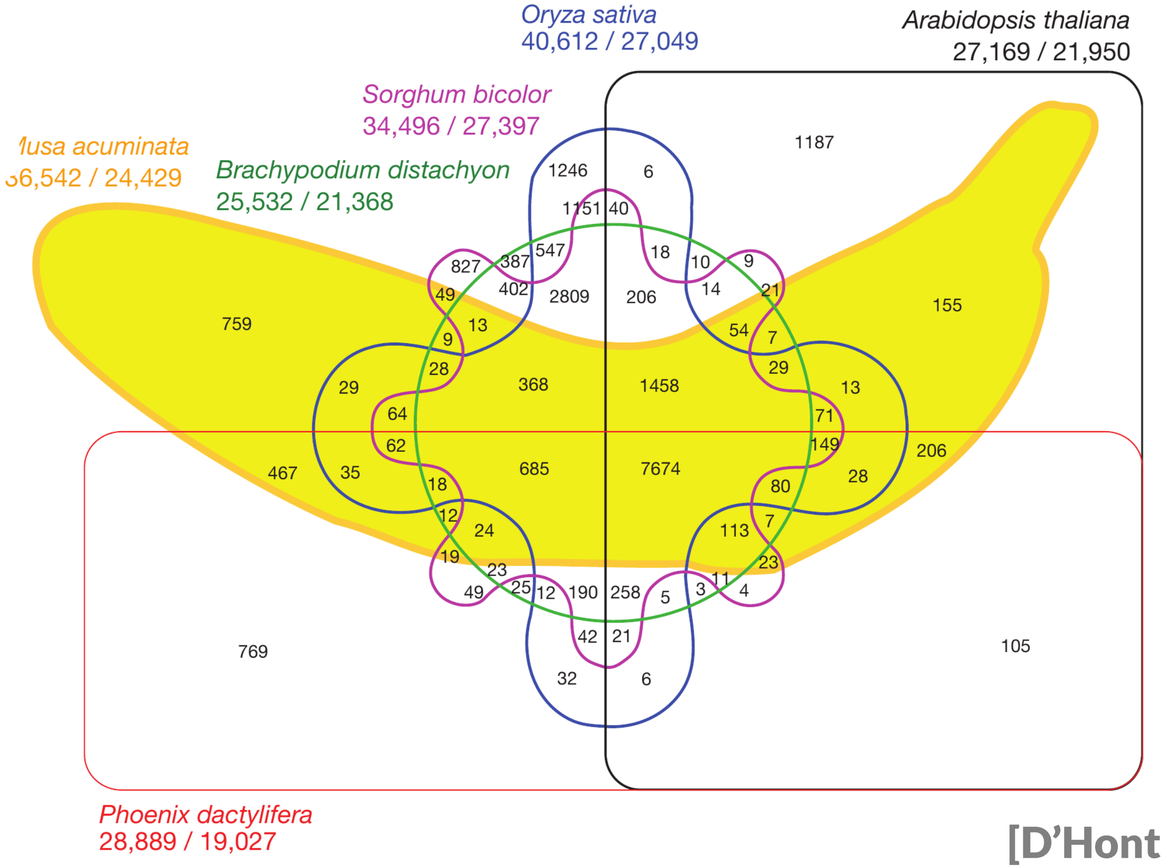

Oryza sativa
40,612 / 27,049
Arabidopsis thaliana
27,169 / 21,950
Sorghum bicolor
34,496 / 27,397
Musa acuminata
36,542 / 24,429
Brachypodium distachyon
25,532 / 21,368
Phoenix dactylifera
28,889 / 19,027
246
6
1151
40
547
827
387
402
2809
206
1187
155
759
368
1453
685
7674
467
769
105



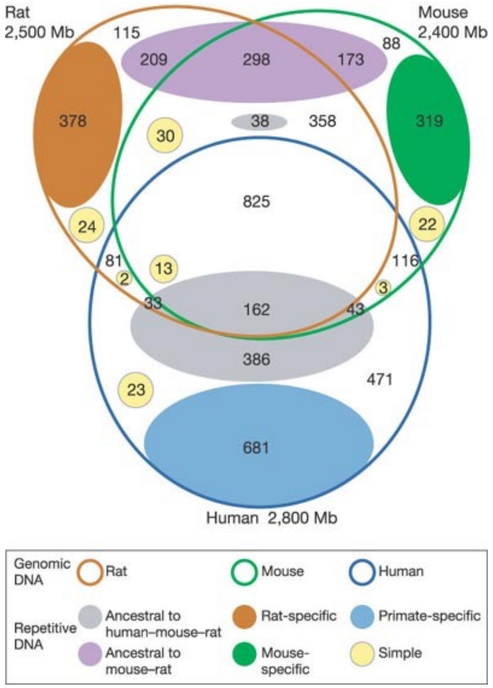
[Neale et al., BMC Genome Biology, 2014]



[Wiles et al., BMC Systems Biology]



[D'Hont et al., Nature, 2012]



[Gibbs et al., Nature, 2004]

Element ID	Sets	Attribute(s)
Name	**Characteristics**	**Age**
Lisa	School, Female	8
Bart	School, Male	10
Homer	Power Plant, Male	40
Mr. Burns	Evil, Power Plant, Male	90

What are some questions we'd like to ask?

Design Workshop

work in groups

get to know the data (5 mins)

create two (rapid!) prototypes (2x5 mins)

Write up your two favorites (5 mins) in google docs

Upload to “Bonus” Canvas Dropbox by 5pm

We’ll show you some of our solutions next time!

Element ID	Sets	Attribute(s)
Name	**Characteristics**	**Age**
Lisa	School, Female	8
Bart	School, Male	10
Homer	Power Plant, Male	40
Mr. Burns	Evil, Power Plant, Male	90

1. What is the biggest intersection?
2. Which sets make up an intersection?
3. How big is an intersection?
4. Does it work for more than four sets?
5. Does attribute value correlate with intersection

Tip: Don’t always try to show all individuals

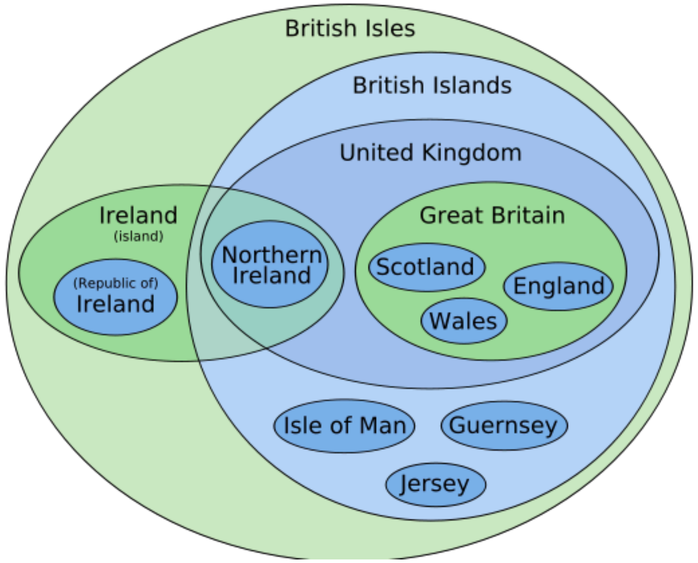
Venn and Euler Diagrams

Venn vs Euler

Euler Diagram

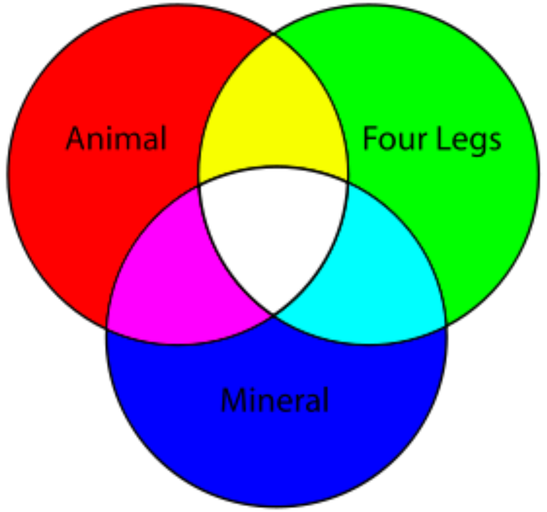
Shows logical relations

May omit empty intersections



Venn Diagram

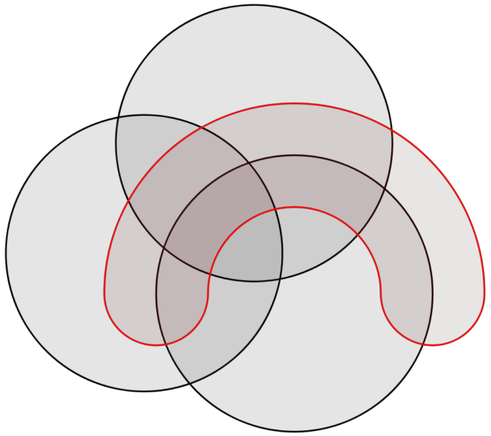
Shows all possible logical relations between sets (even if empty)

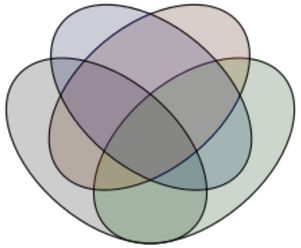


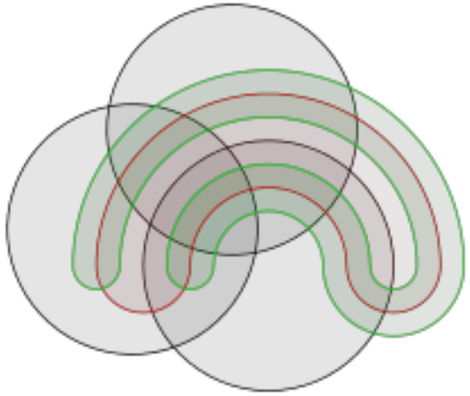
Venn Diagrams

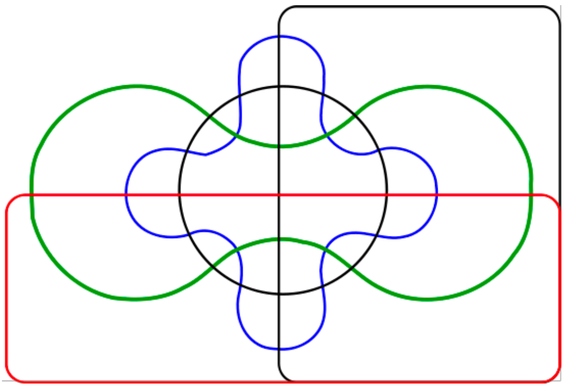
Venn diagrams for many sets are hard

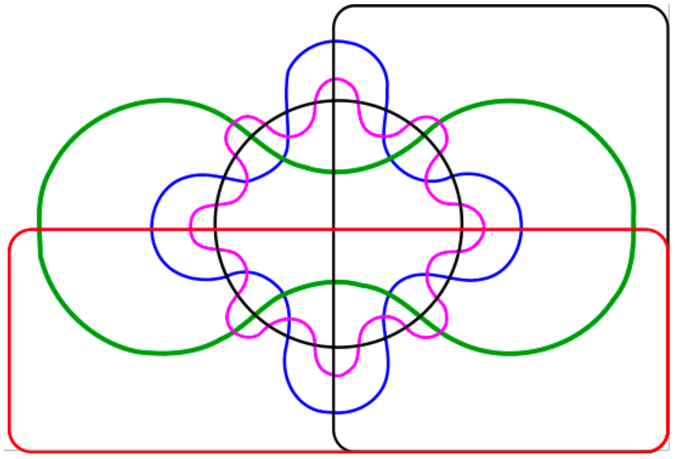
# of intersections is 2^n

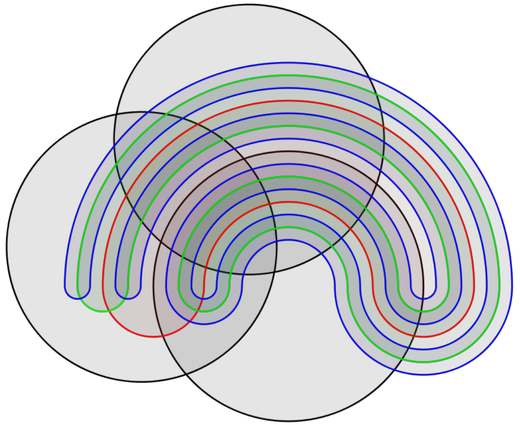


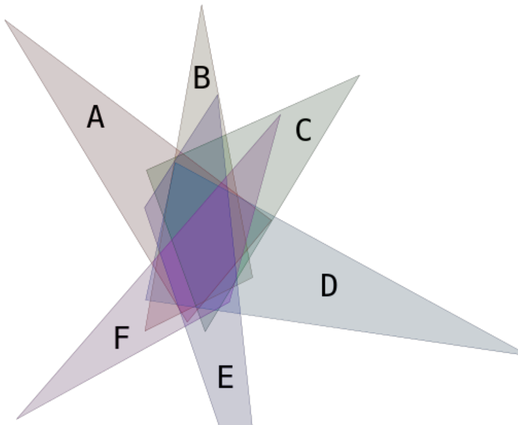












A: (-69277,-32868), (135580,121186), (70900,199427)
B: (81988,-44426), (38444,206222), (121044,165111)
C: (203271, 9619), (39604, 82683), (84652,206669)
D: (333561,225349), (61764, 76805), (38980,182461)
E: (131886,385785), (38136,111491), (94208, 24690)
F: (-60184,274046), (142476, 39903), (103276,183962)

https://en.wikipedia.org/wiki/Venn_diagram

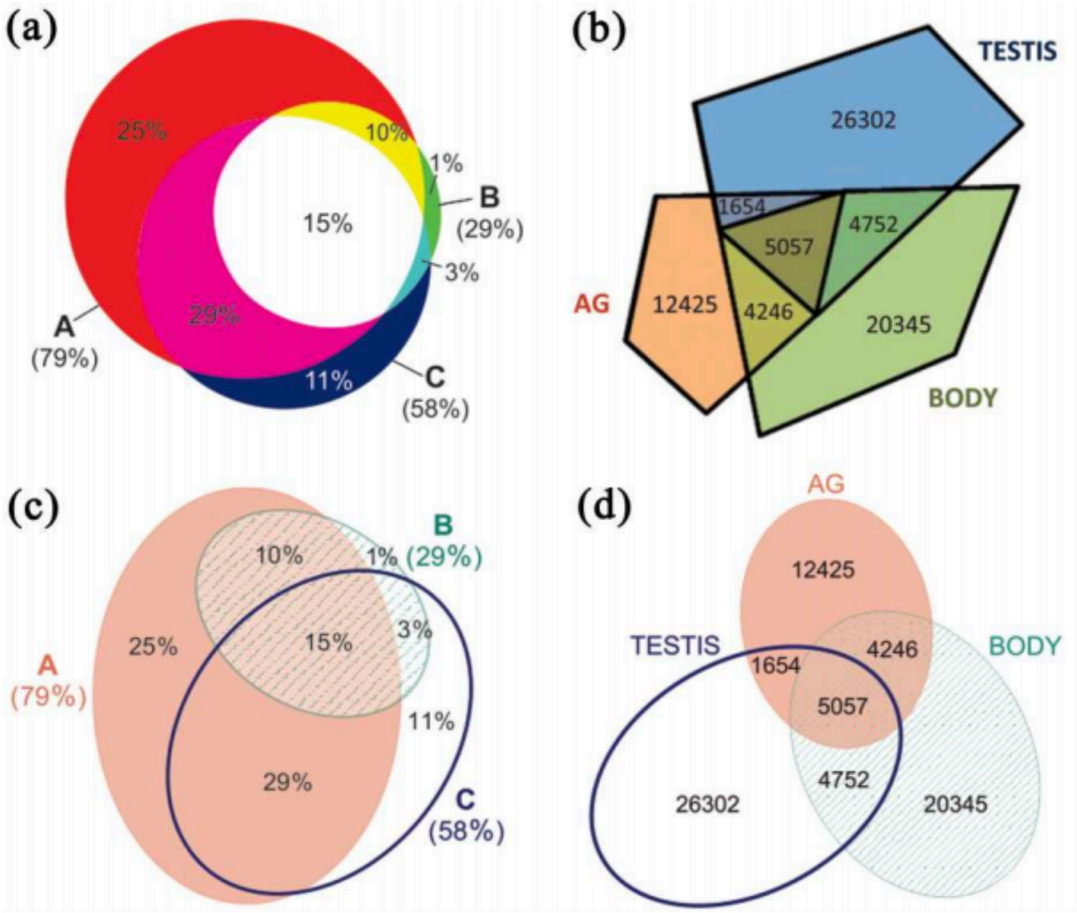
Area-Proportional Euler Diagrams

Problem with Venn: size doesn't correspond to the data.

Creating area-proportional Euler diagrams is hard.

Layout criteria:

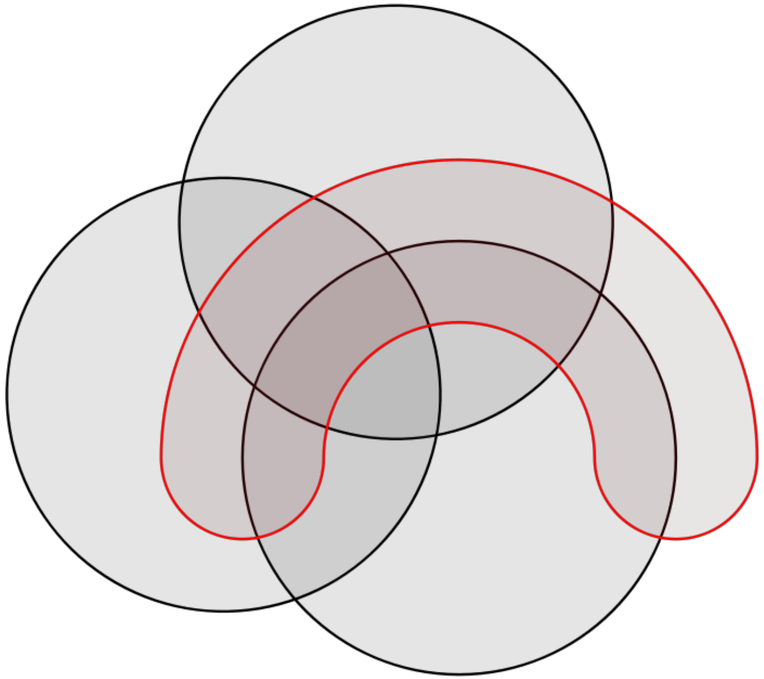
- area proportional
- simple curves (circles are best)
  - makes it easy to identify which sets are participating in intersection
  - Gestalt-principle: good continuation



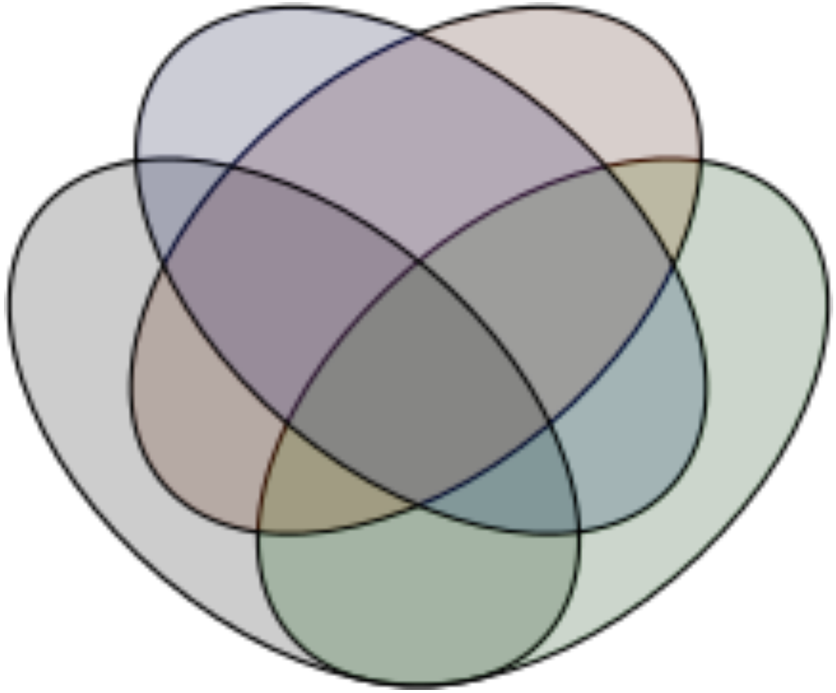
[Alsallakh 2015]

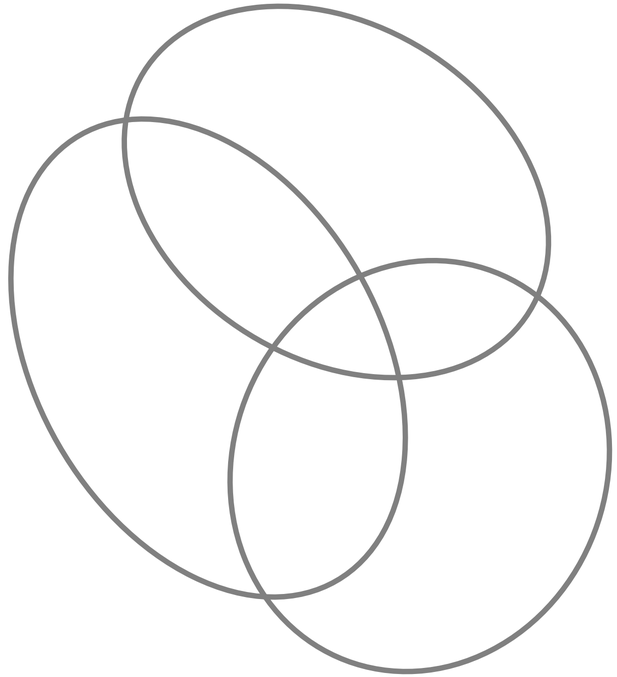
Compare Simple vs Complex Shape

Complex



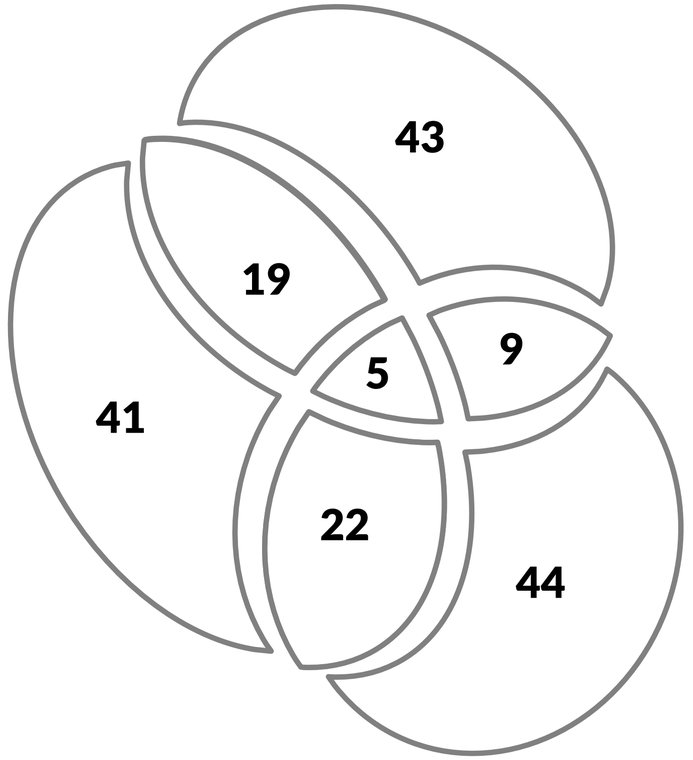
Simple





[created with EulerAPE]

22 ? 19



[created with EulerAPE]

Venn-Euler Pros/Cons

Pros

Familiar

Intuitive

Work well for 2-4 sets

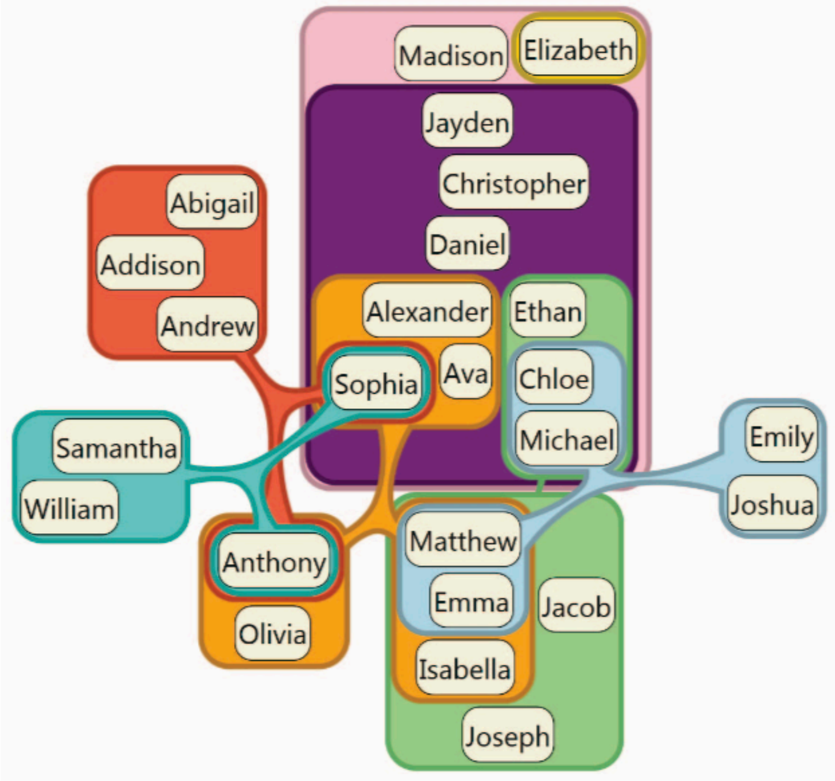
Cons

Doesn’t work well for more than 4 sets

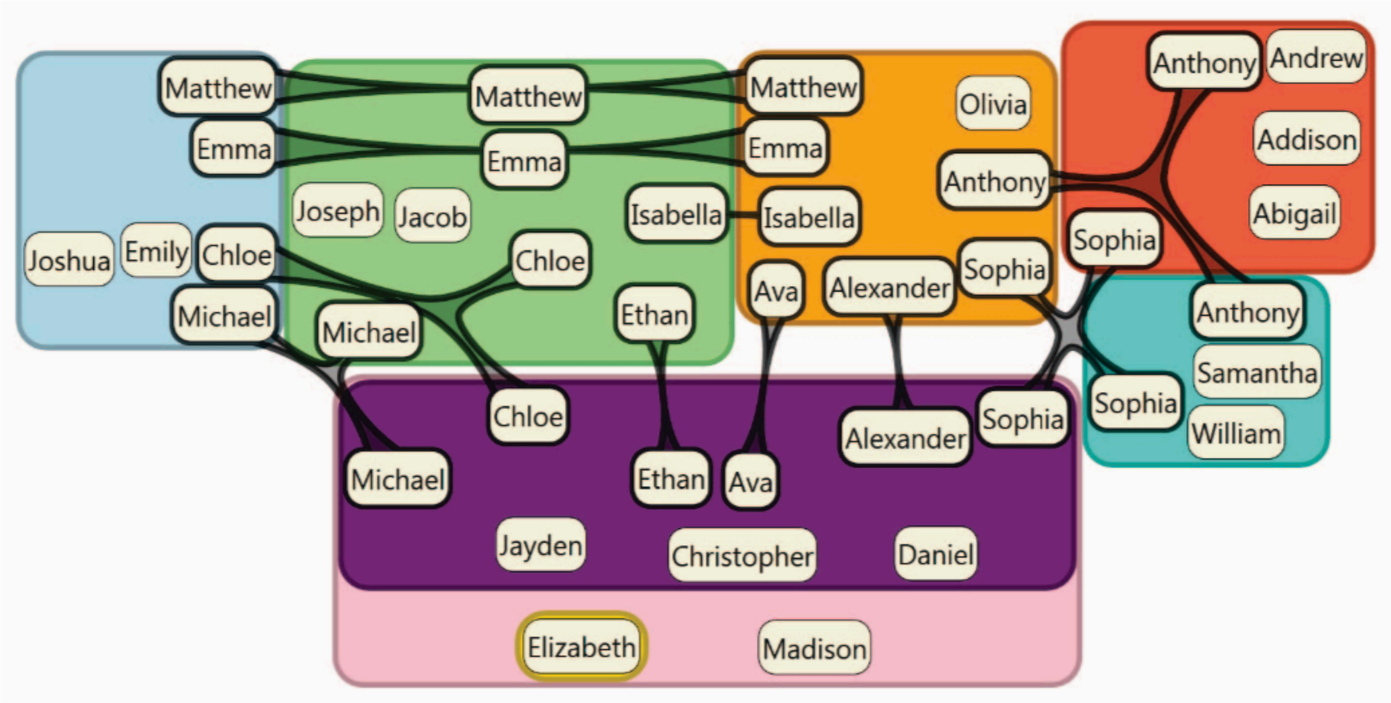
Area proportionality hard to do

Not well suited to show attributes

Relationships for specific Items

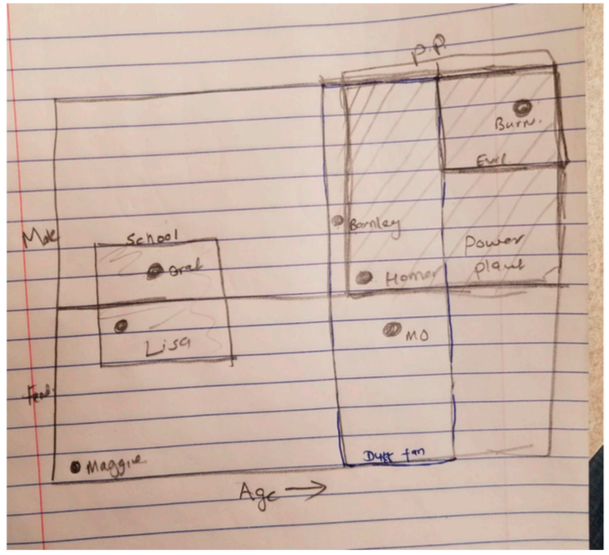


No Duplicate Nodes
Complex Shapes
Notice the Nesting

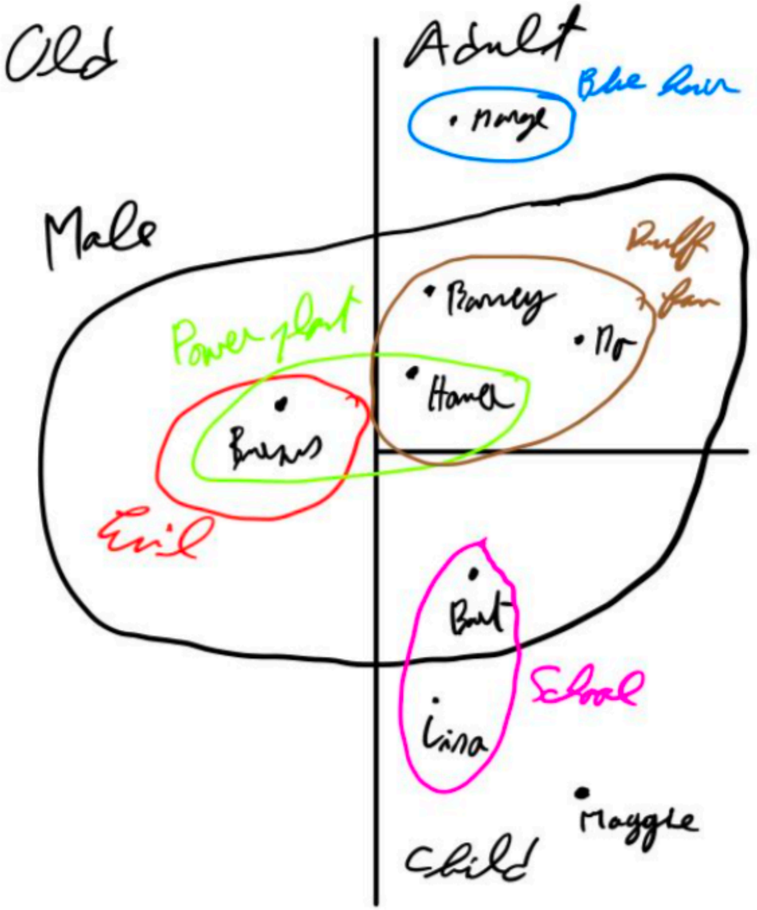


Duplicate Nodes
Simple Shapes

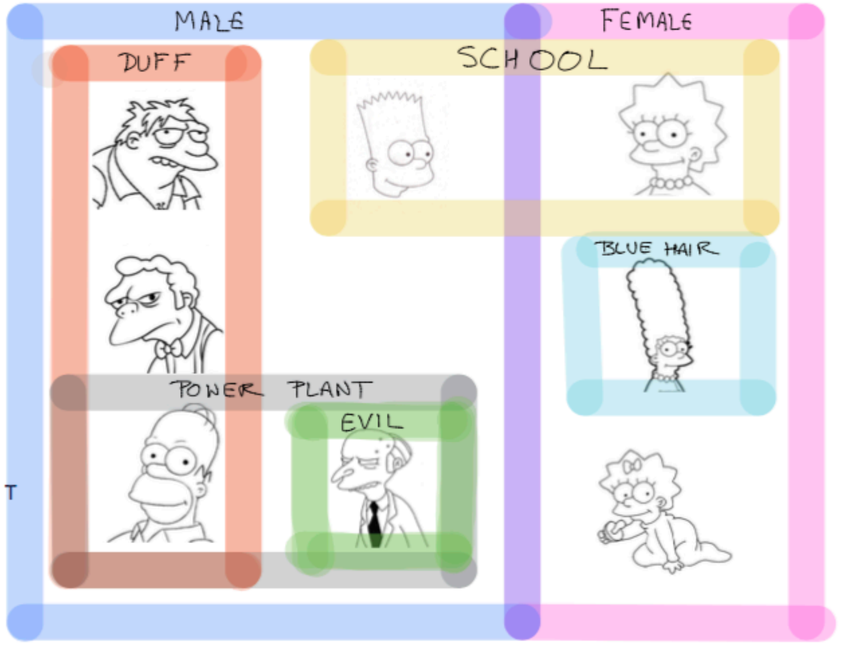
[Riche 2010]



Burn
Evil
Male
School
Bornley
Power plant
Homer
Lisa
MO
Maggie
Age →

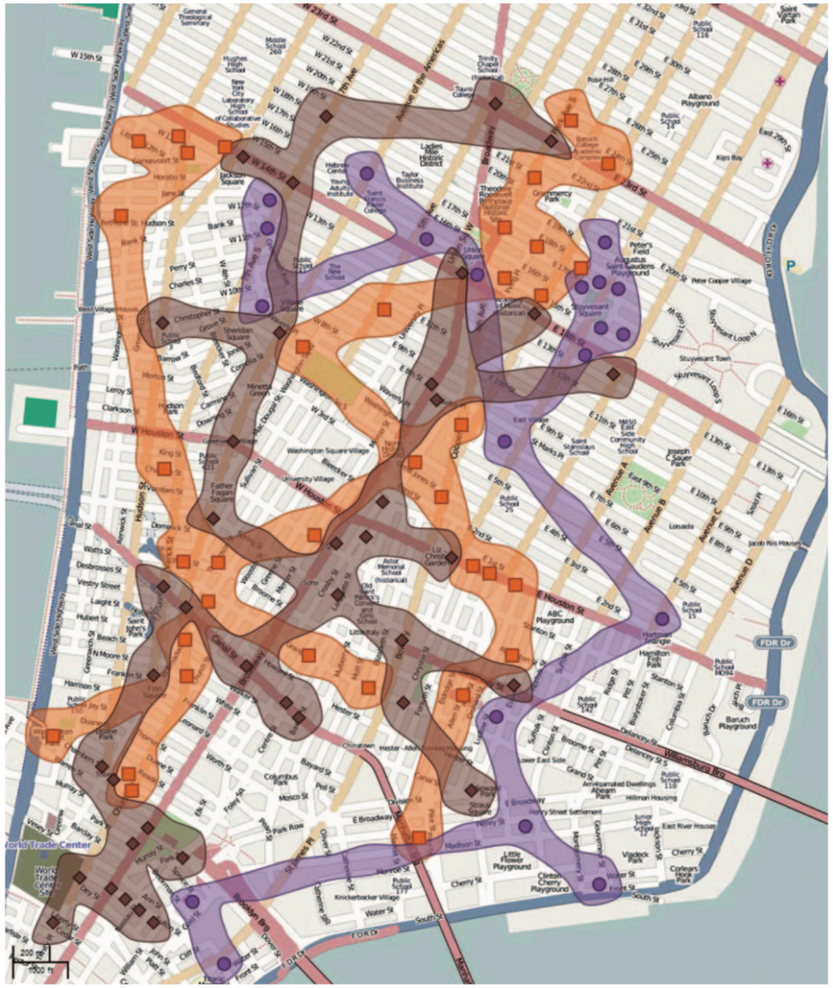


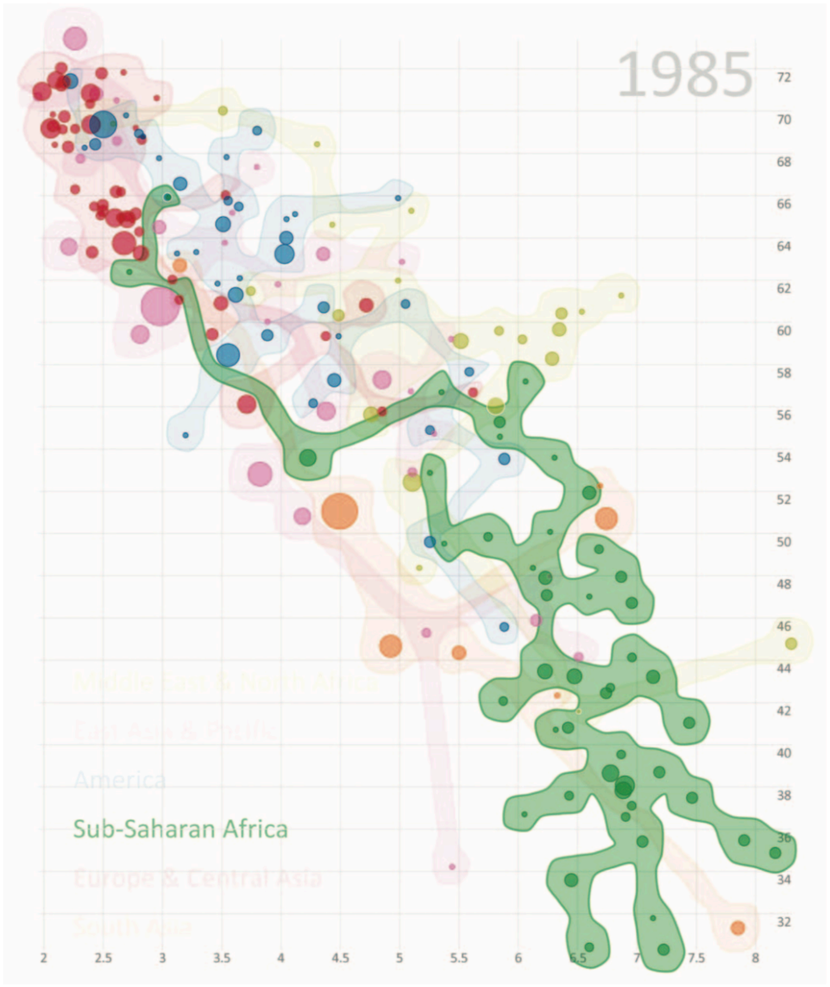
Old
Adult
Blue hair
Male
Power plant
Bornley
Duff fan
Homer
Evil
School
Lisa
Maggie
Child



MALE
FEMALE
DUFF
SCHOOL
BLUE HAIR
POWER PLANT
EVIL

Sets on top of a fixed layout

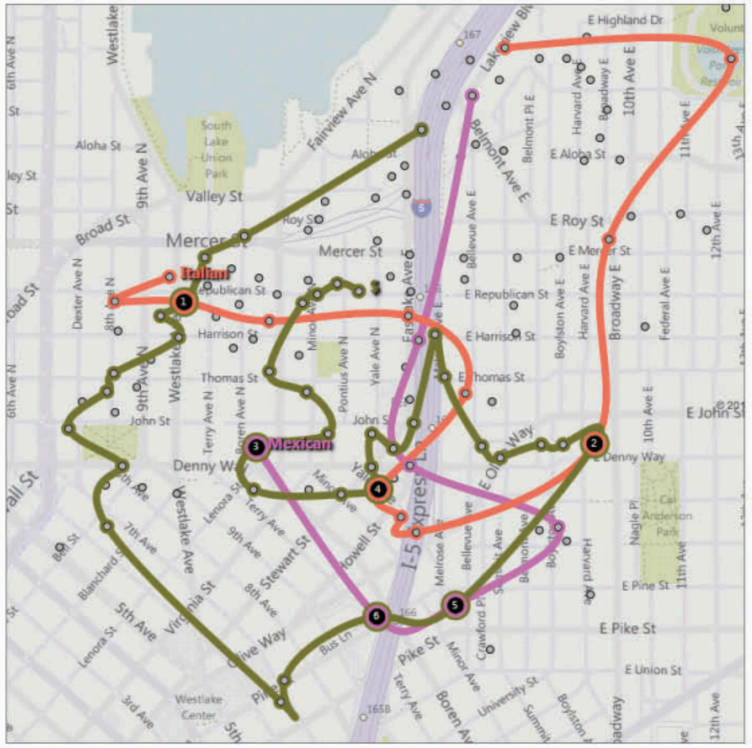




https://www.youtube.com/watch?v=Ju2hSThmPWA

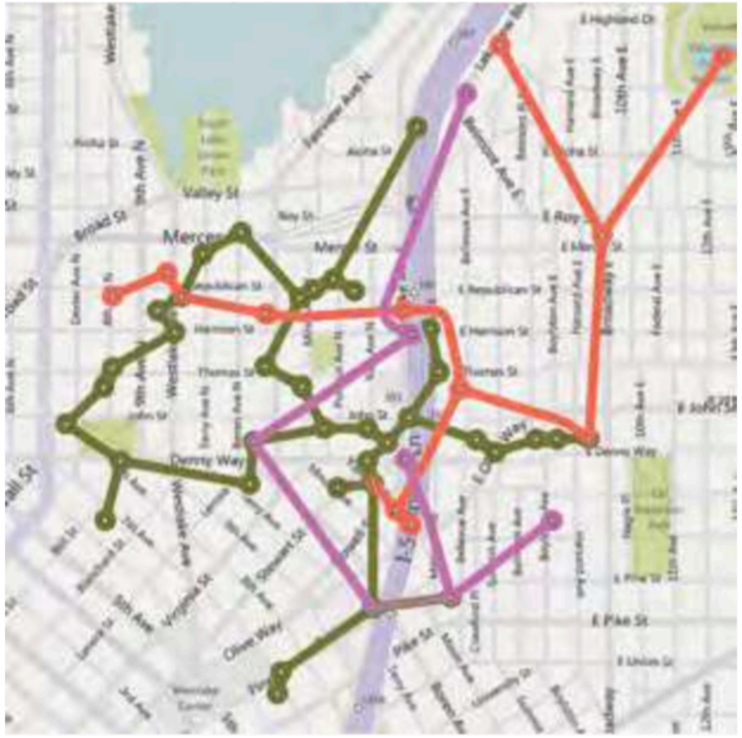
Sets on top of a fixed layout

LineSets



[Alper 2011]

Kelp Diagrams

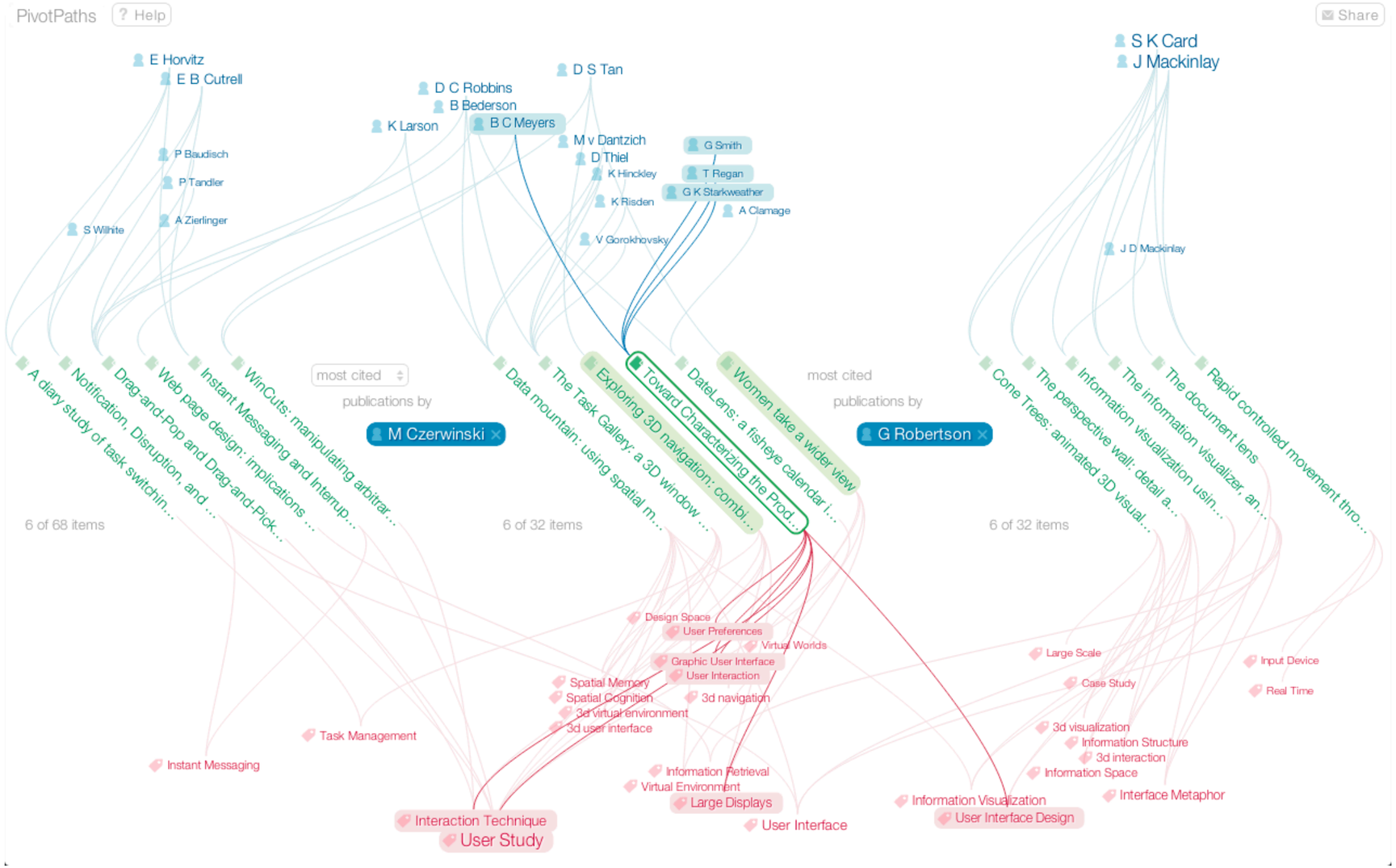


[Dinkla 2012]

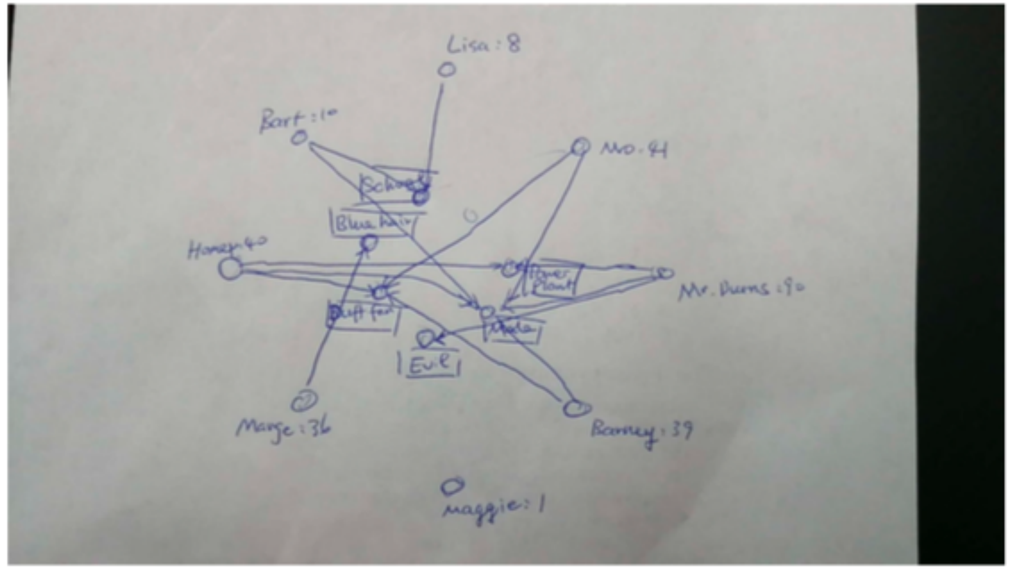
Node-Link Techniques

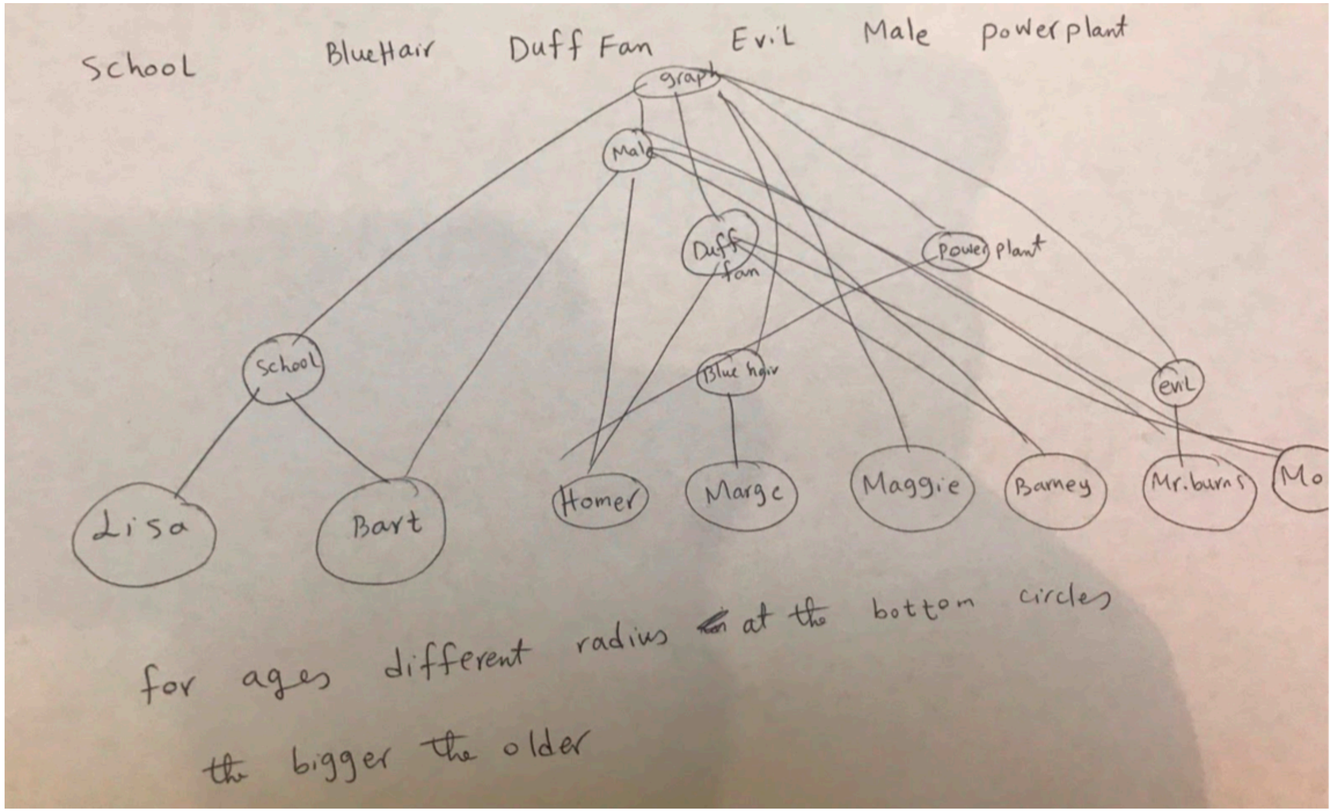
Treat sets as nodes

Connect to elements that are in set



http://mariandoerk.de/pivotpaths/demo/#/1:0_497686


Lisa : 8
Bart : 10
No. 44
Blue hair
Homer 40
Mr. Burns : 90
Duff fan
Evil
Marge : 36
Barney : 39
Maggie : 1


School
BlueHair
Duff Fan
Evil
Male
powerplant
graph
Male
Duff fan
Power Plant
School
Blue hair
evil
Lisa
Bart
Homer
Marge
Maggie
Barney
Mr.burns
Mo
for ages different radius at the bottom circles
the bigger the older

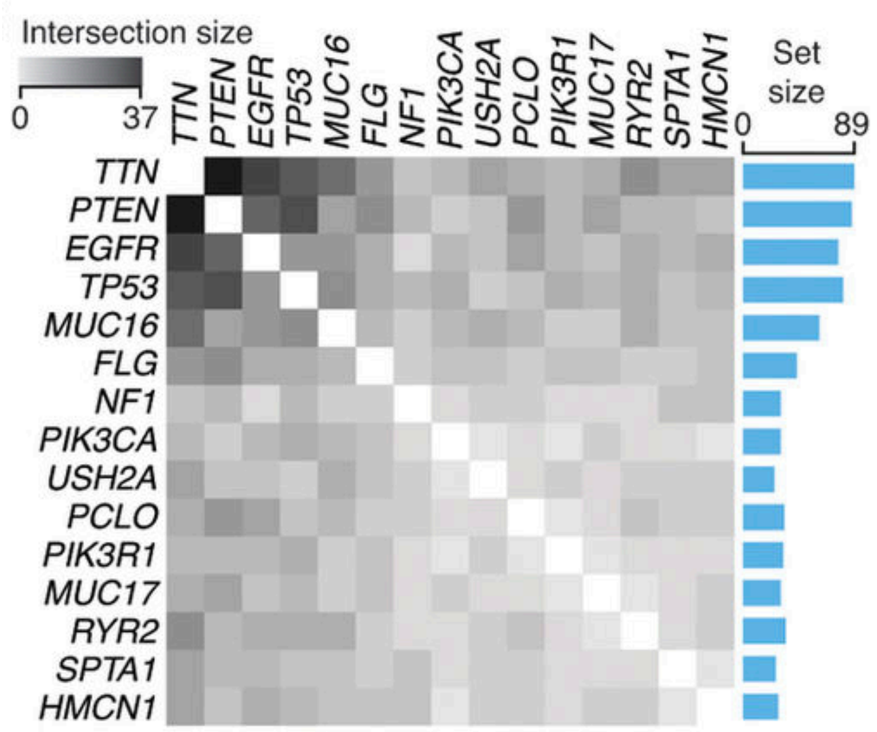
Showing Pairwise Overlap

Doesn’t show higher-order overlaps

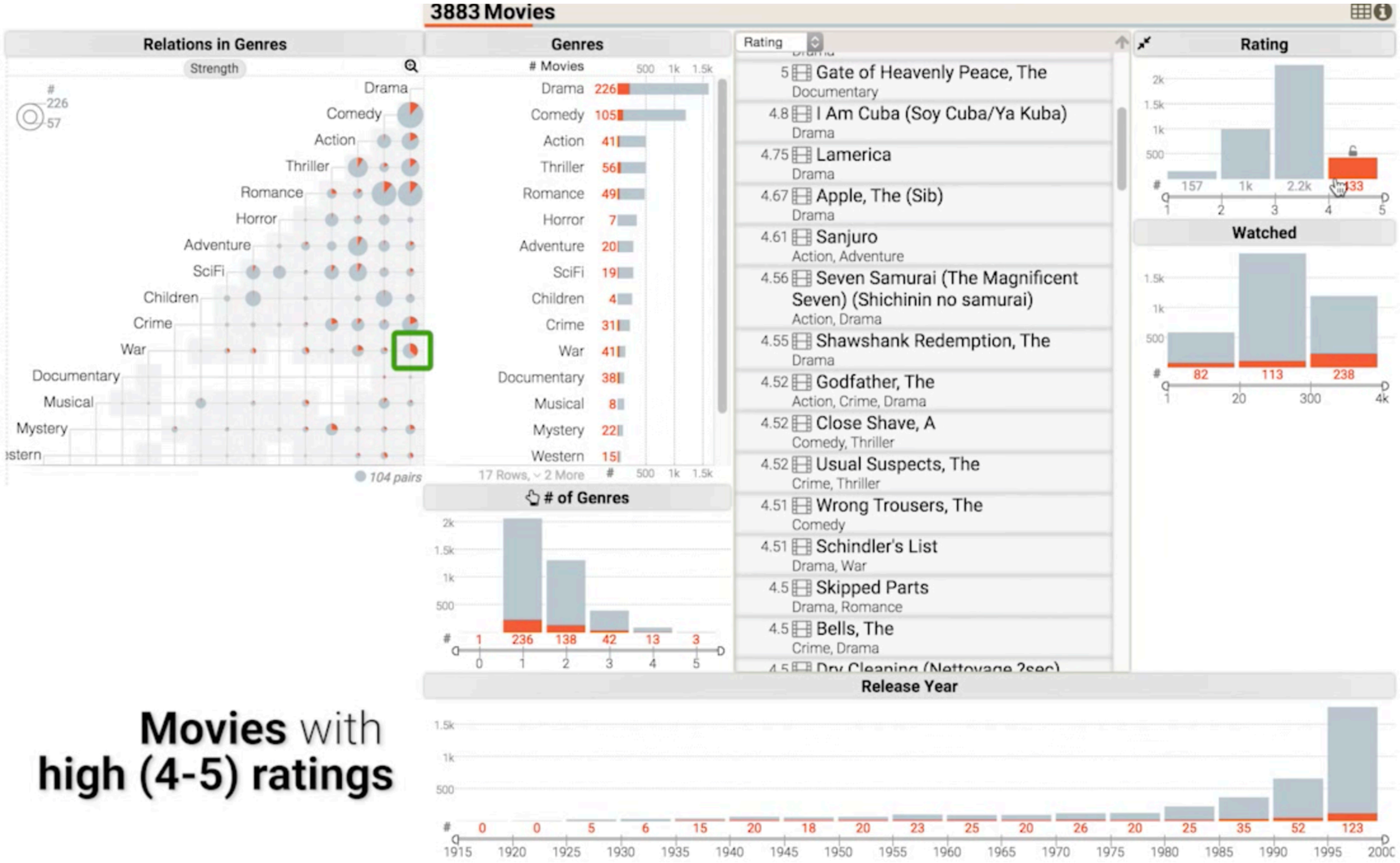
Very scalable

Can’t show attributes

Co-Mutations of genes



Pairwise + Interaction



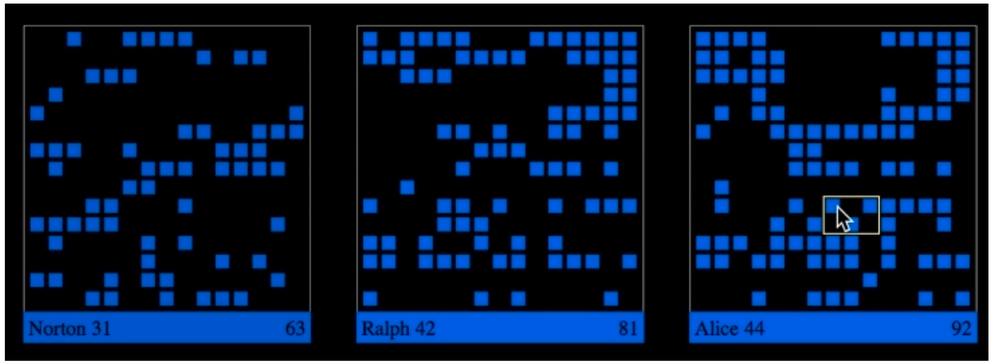
Movies with
high (4-5) ratings

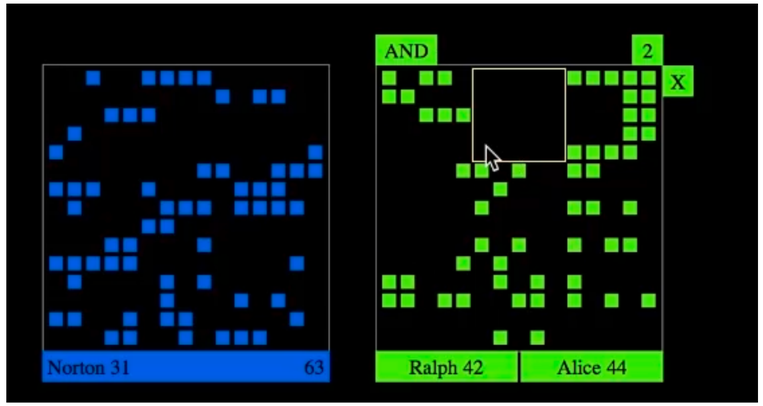
Set Matrices: OnSet

Set membership for each item shown in matrix

Comparisons can be made using AND or OR operations

Good for many sets and few items





https://vimeo.com/213029678#at=0

Linear Diagrams

Fig. 1. Visualizing sets: linear diagrams.

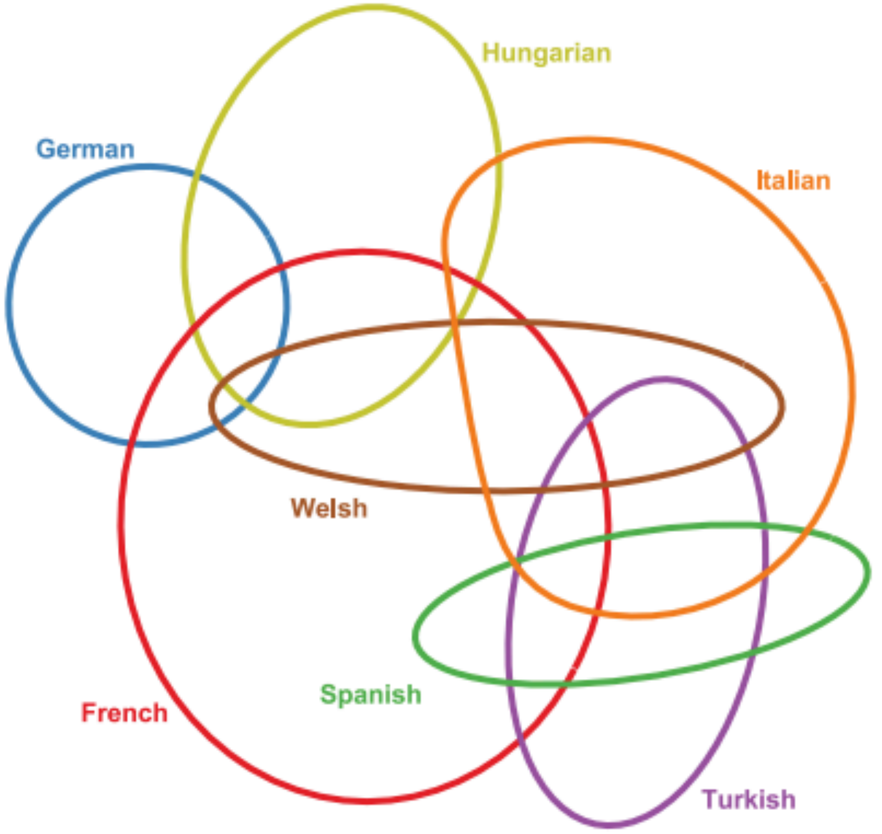
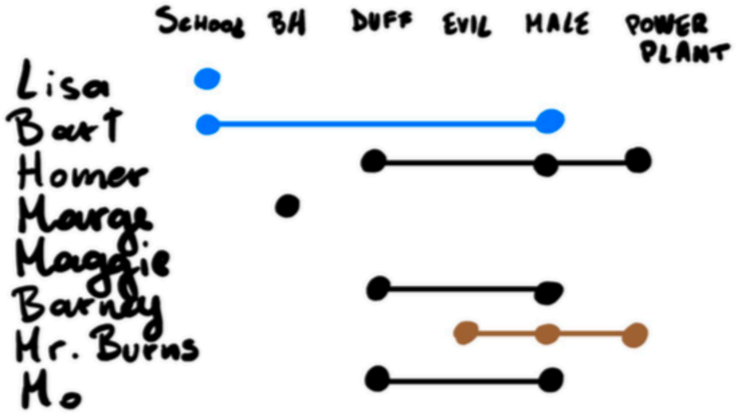
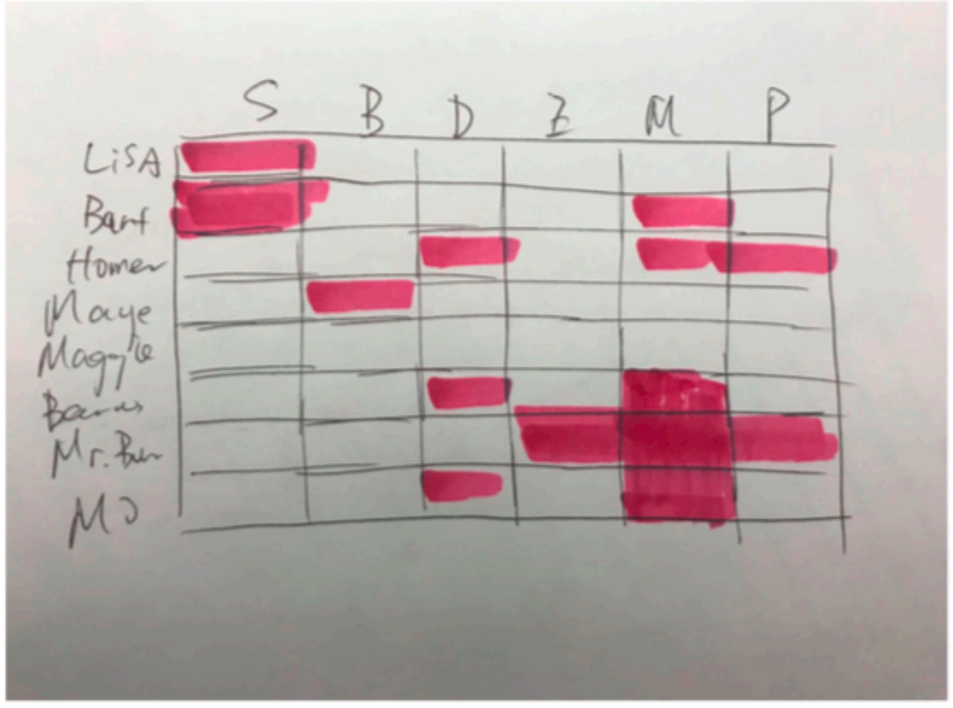
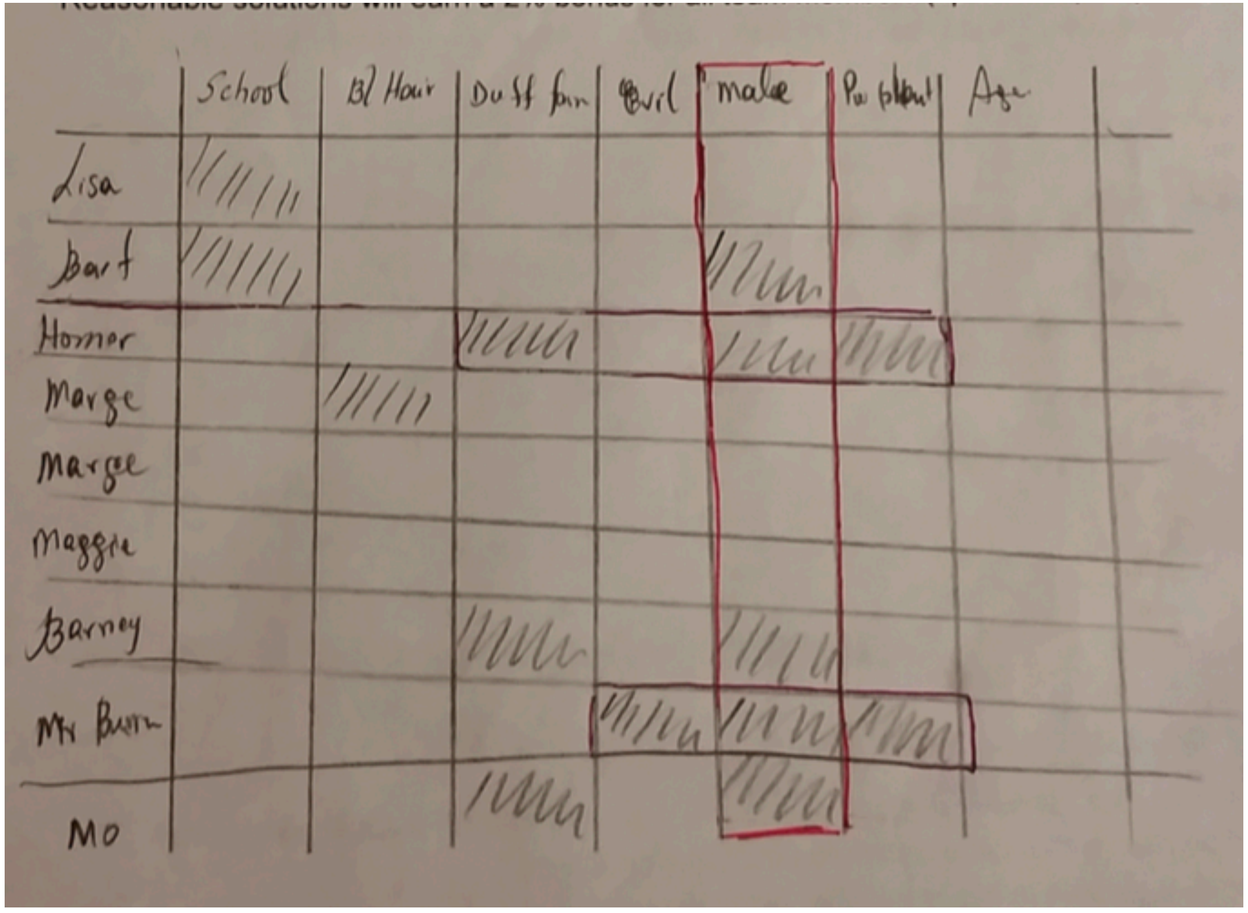


Fig. 2. Visualizing sets: Euler diagrams.

[RODGERS 2015]

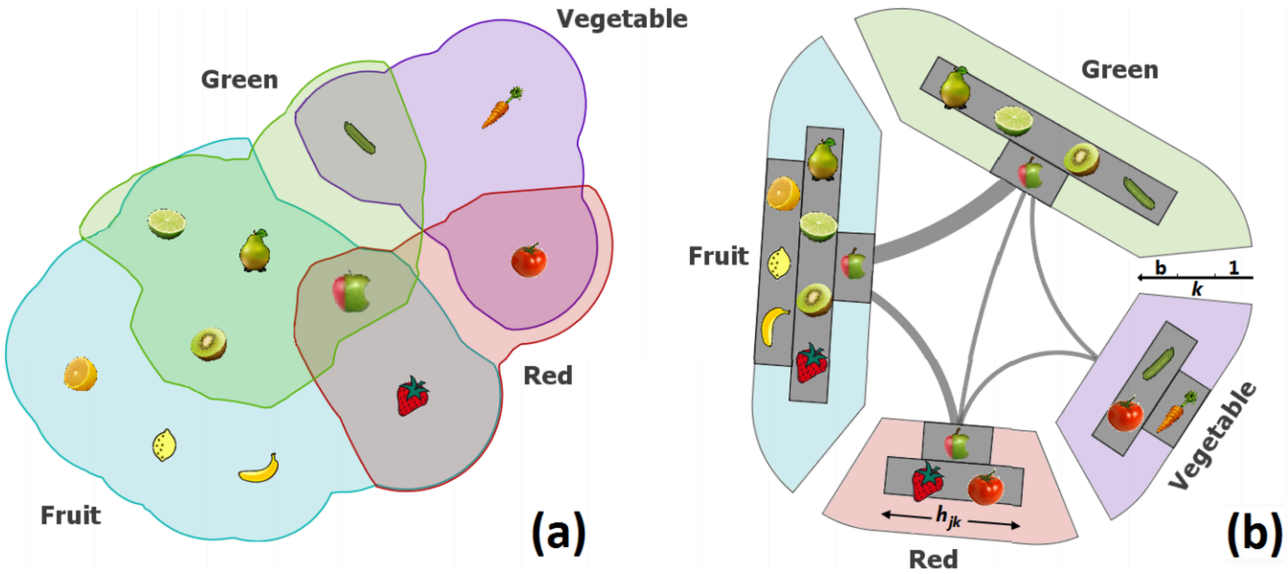

School
BH
DUFF
EVIL
MALE
POWER PLANT
Lisa
Bart
Homer
Marge
Maggie
Barney
Mr. Burns
Mo


S
B
D
E
M
P
LiSA
Bart
Homer
Maye
Maggie
Barns
Mr. Bur
MO


School
Bl Hair
Duff fan
Evil
make
Pow blant
Age
Lisa
Bart
Homer
Marge
Margee
Maggie
Barney
Mr Burn
MO

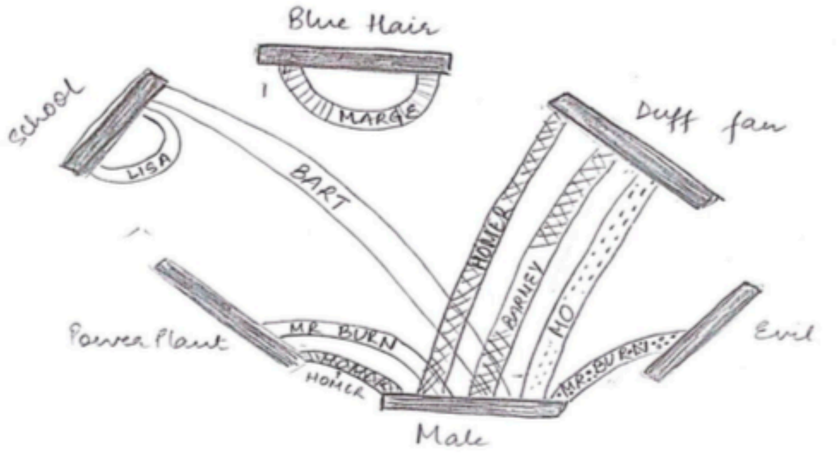
Radial Sets

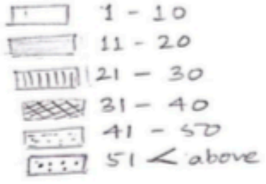
- Sets are segments on a “circle”
- Relationships are encoded as ribbons
- Size of segments encodes size of sets
- Histograms in segments show degrees

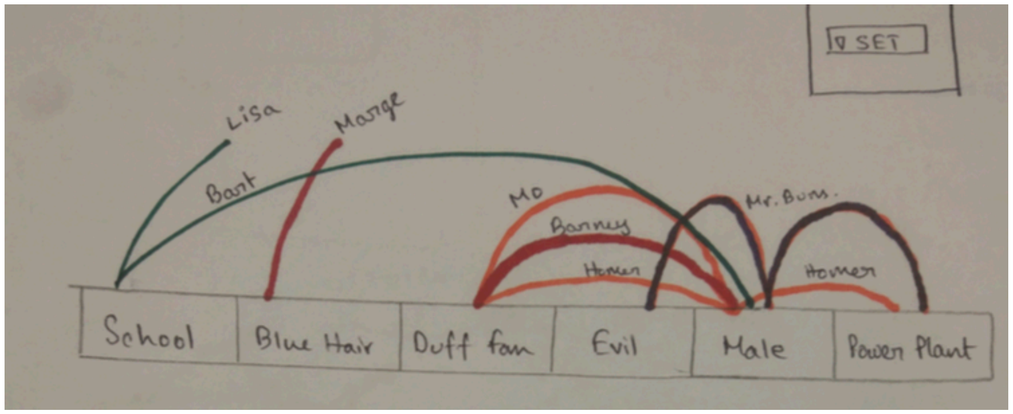


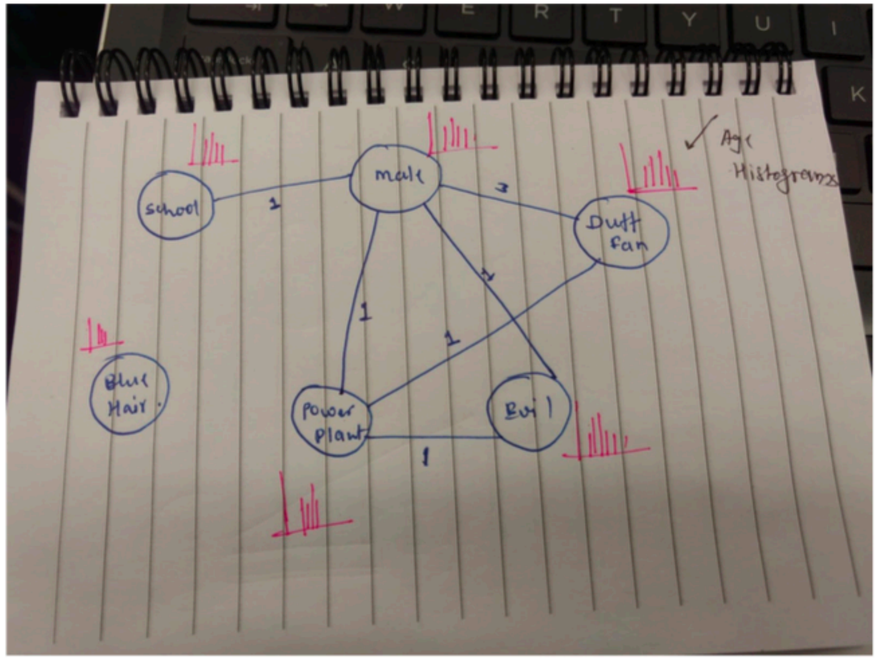
https://www.youtube.com/watch?v=UcYRrPqC5A8

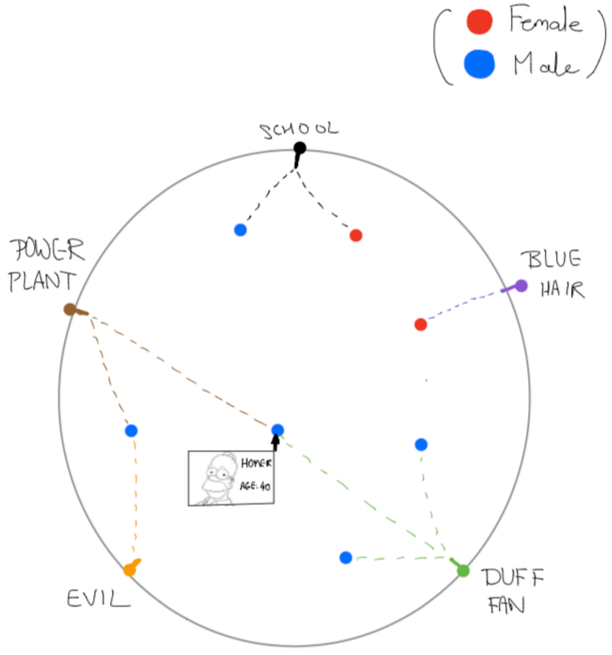
[Alsallakh 2013]


Blue Hair
MARGE
School
LISA
BART
Duff fan
HOMER
BARNEY
MO
Power Plant
MR BURN
HOMER
MR BURN
Evil
Male


1 - 10
11 - 20
21 - 30
31 - 40
41 - 50
51 < above


SET
Lisa
Marge
Bart
MO
Barney
Homer
Mr. Burn.
Homer
School
Blue Hair
Duff fan
Evil
Male
Power Plant

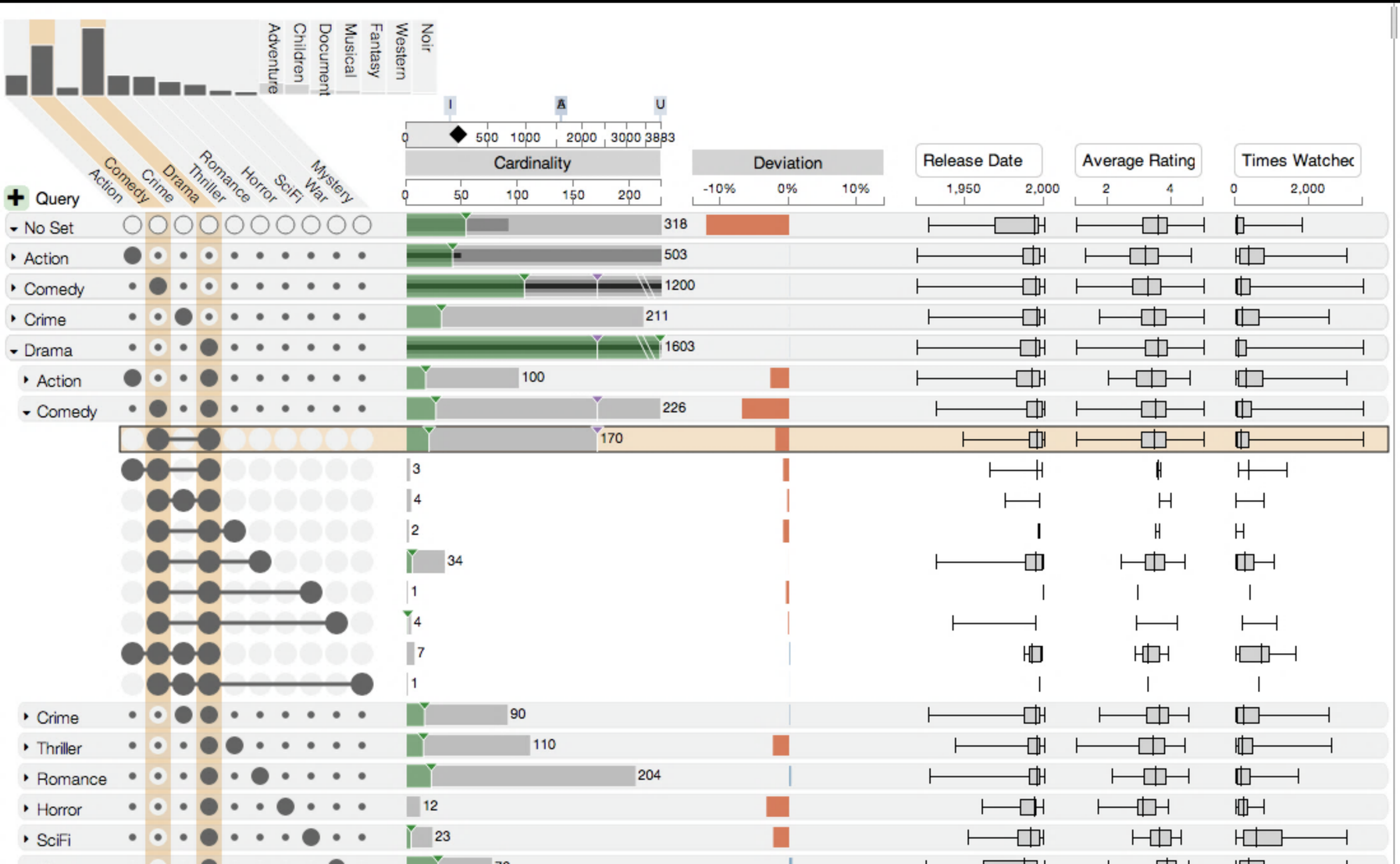

school
male
Duff fan
Blue Hair
power plant
Evil
1
1
1
1
1
1
Age Histograms

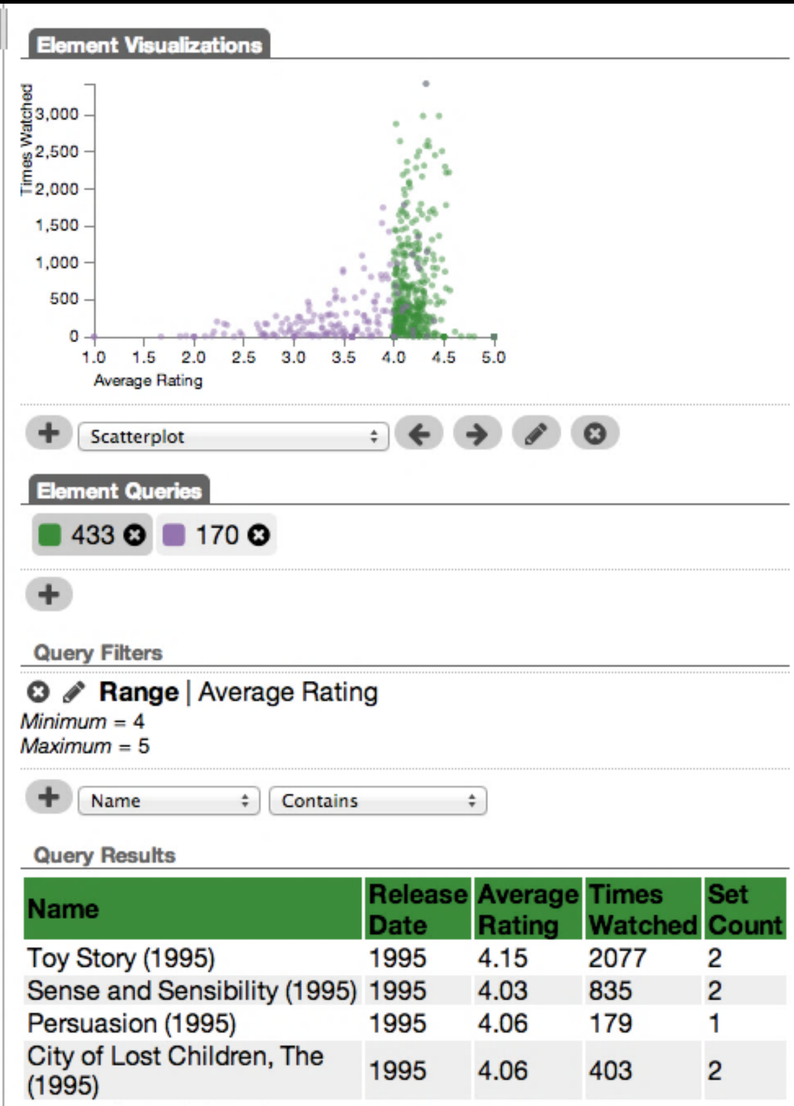

Female
Male
SCHOOL
POWER PLANT
BLUE HAIR
HOMER AGE: 40
EVIL
DUFF FAN

UpSet

[InfoVis'14]

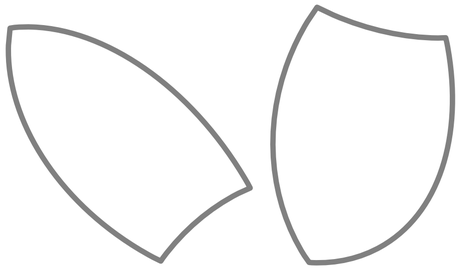
Visualizing Intersecting Sets



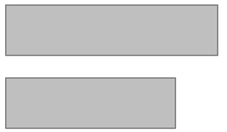


Set Vis Goals

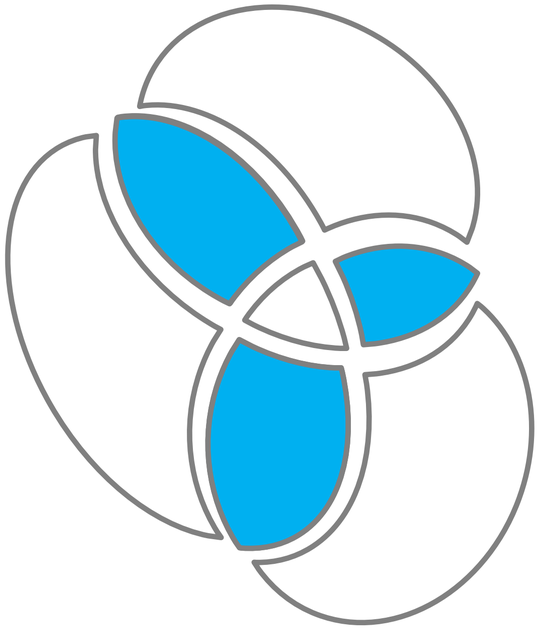
1. Efficient visual encoding



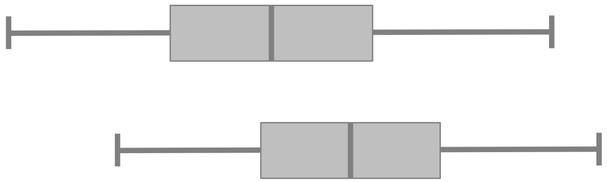
vs.



2. Creating complex slices of a dataset



3. Visualize attributes



[Movie Lens Dataset]

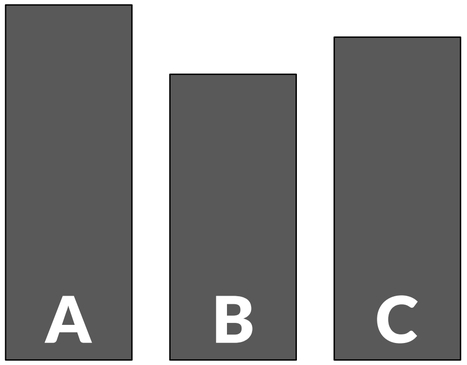
Attribute Details

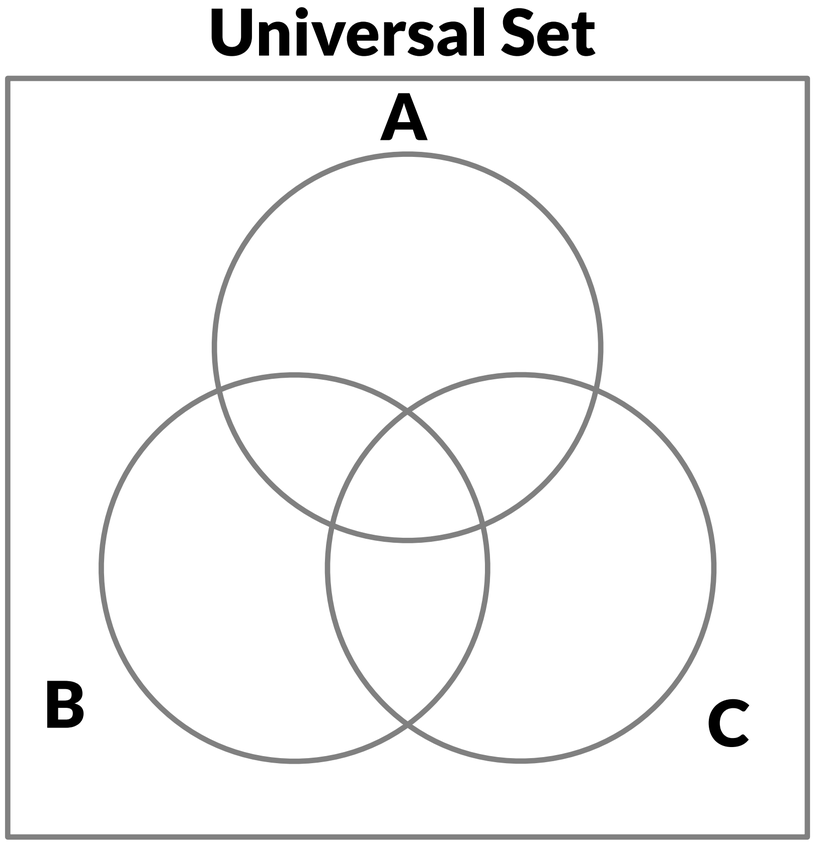
Visualizing Intersections

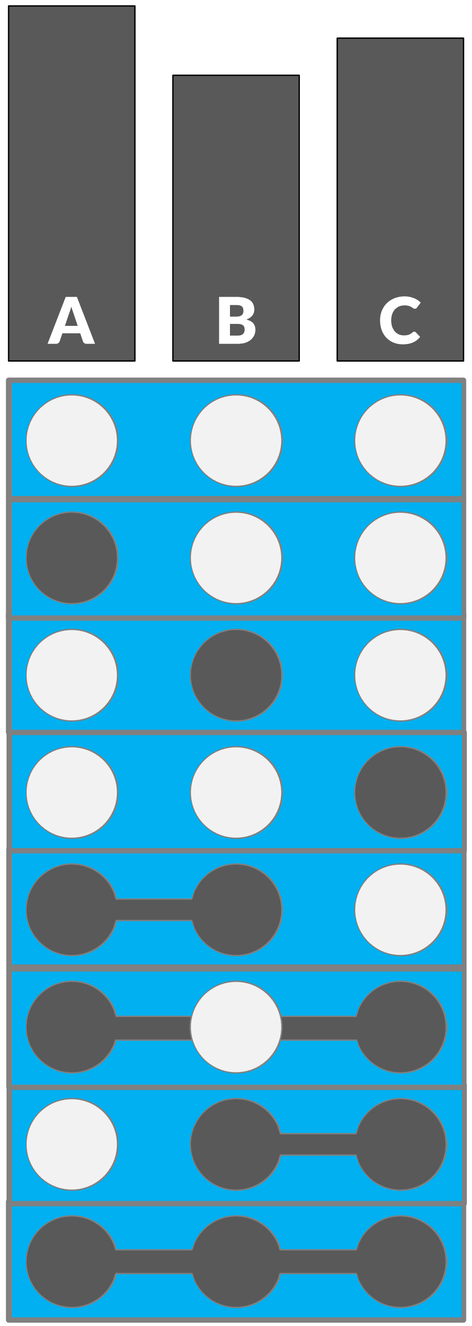
Visualizing Properties

Element List & Queries

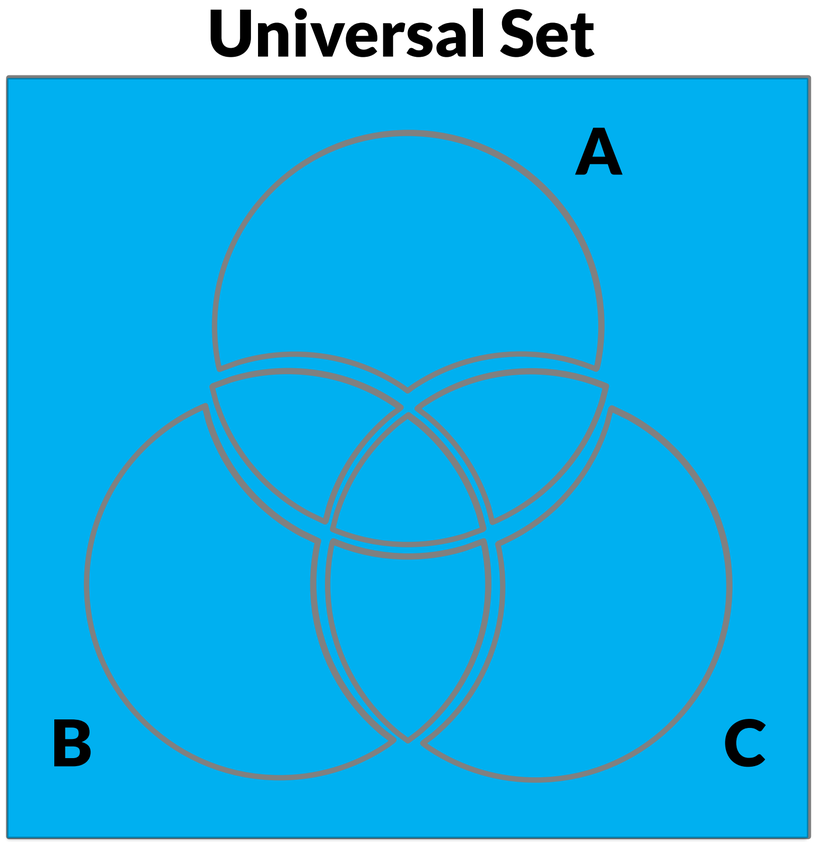
Visualizing Intersections

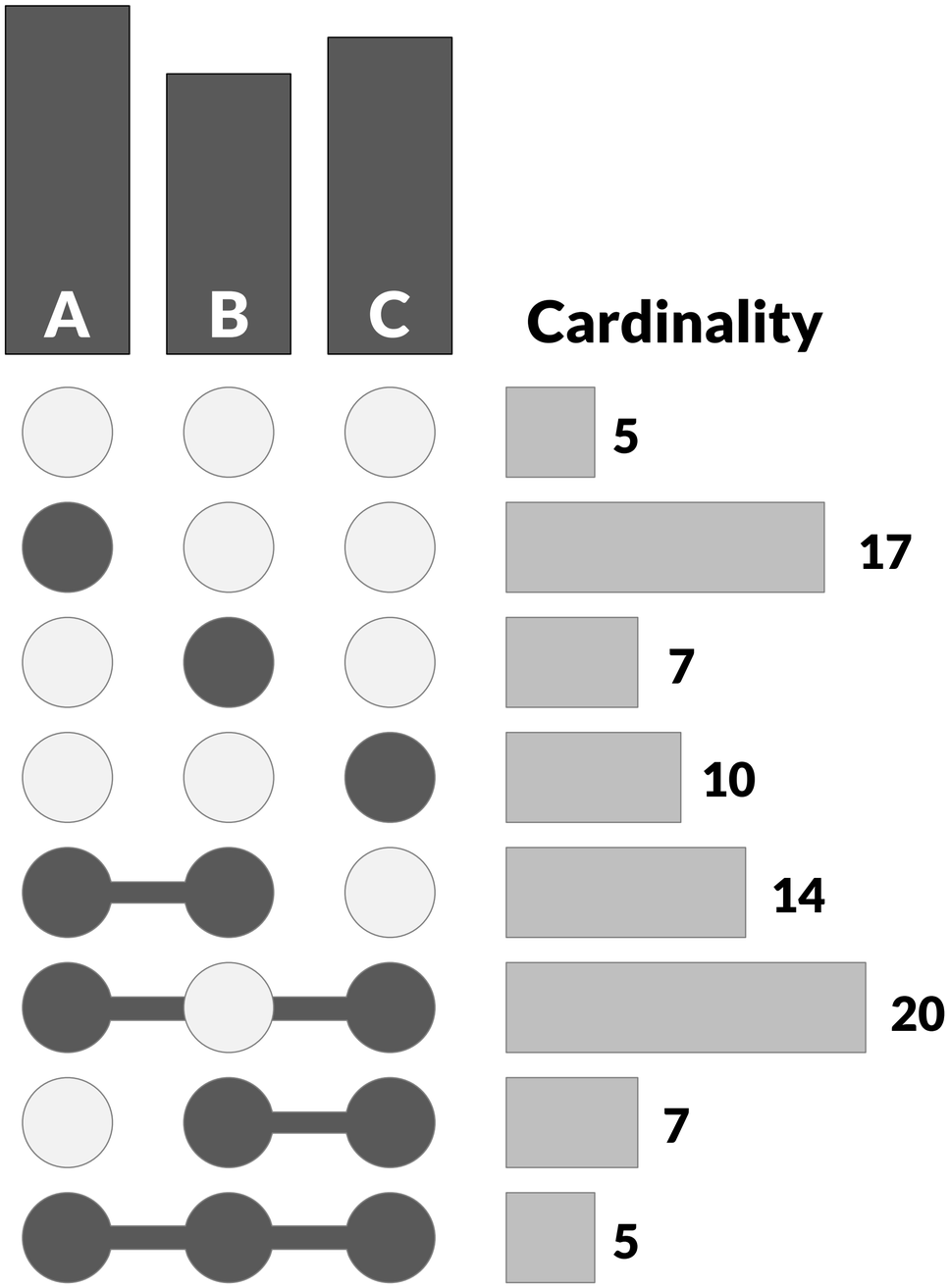

A
B
C

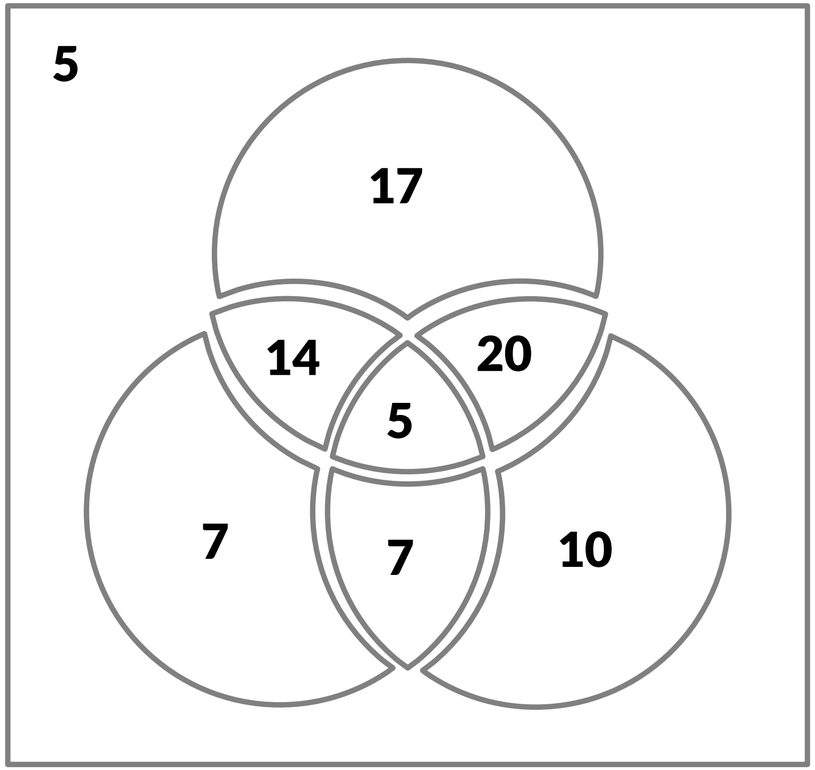

Universal Set
A
B
C


A
B
C

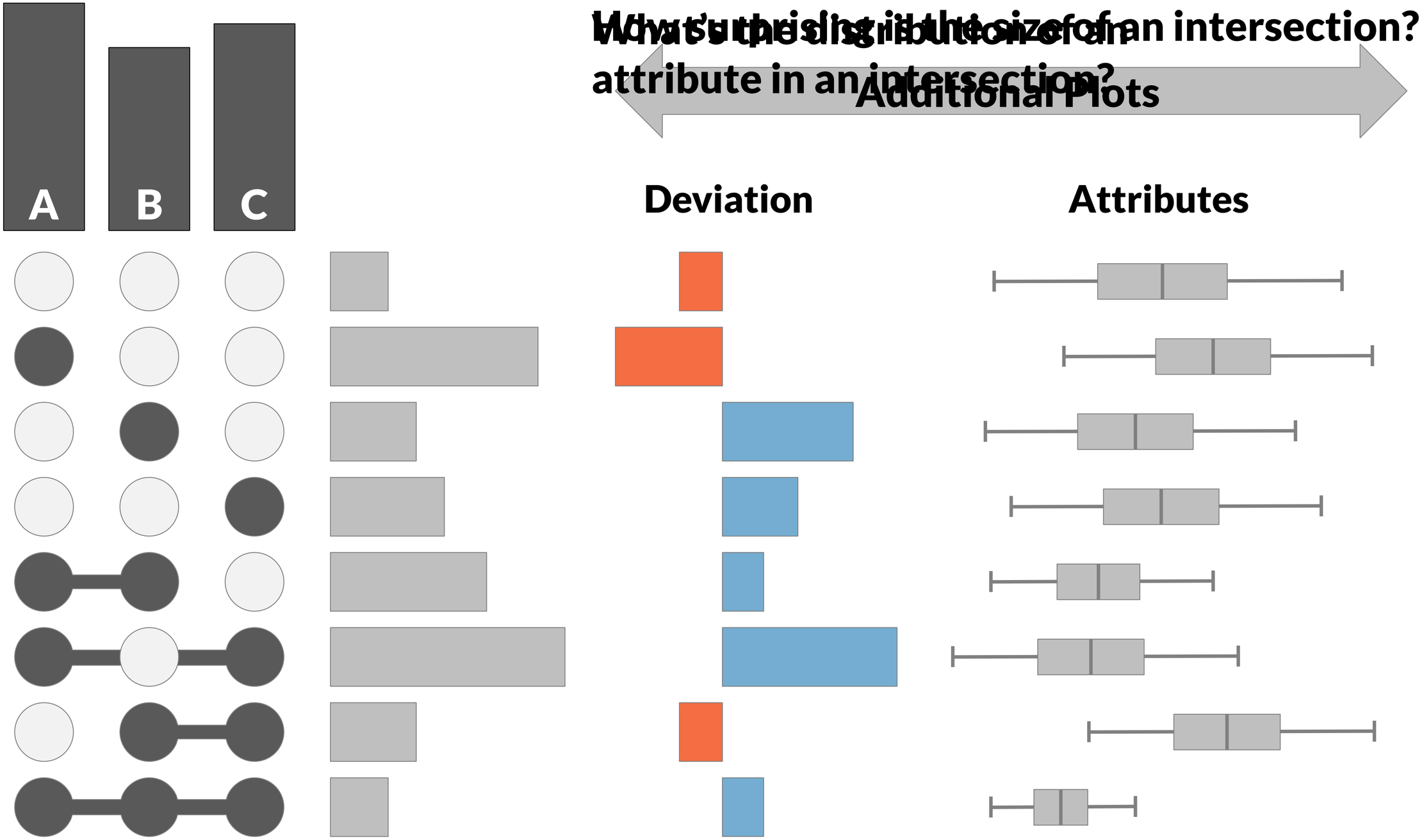
Must
Must Not

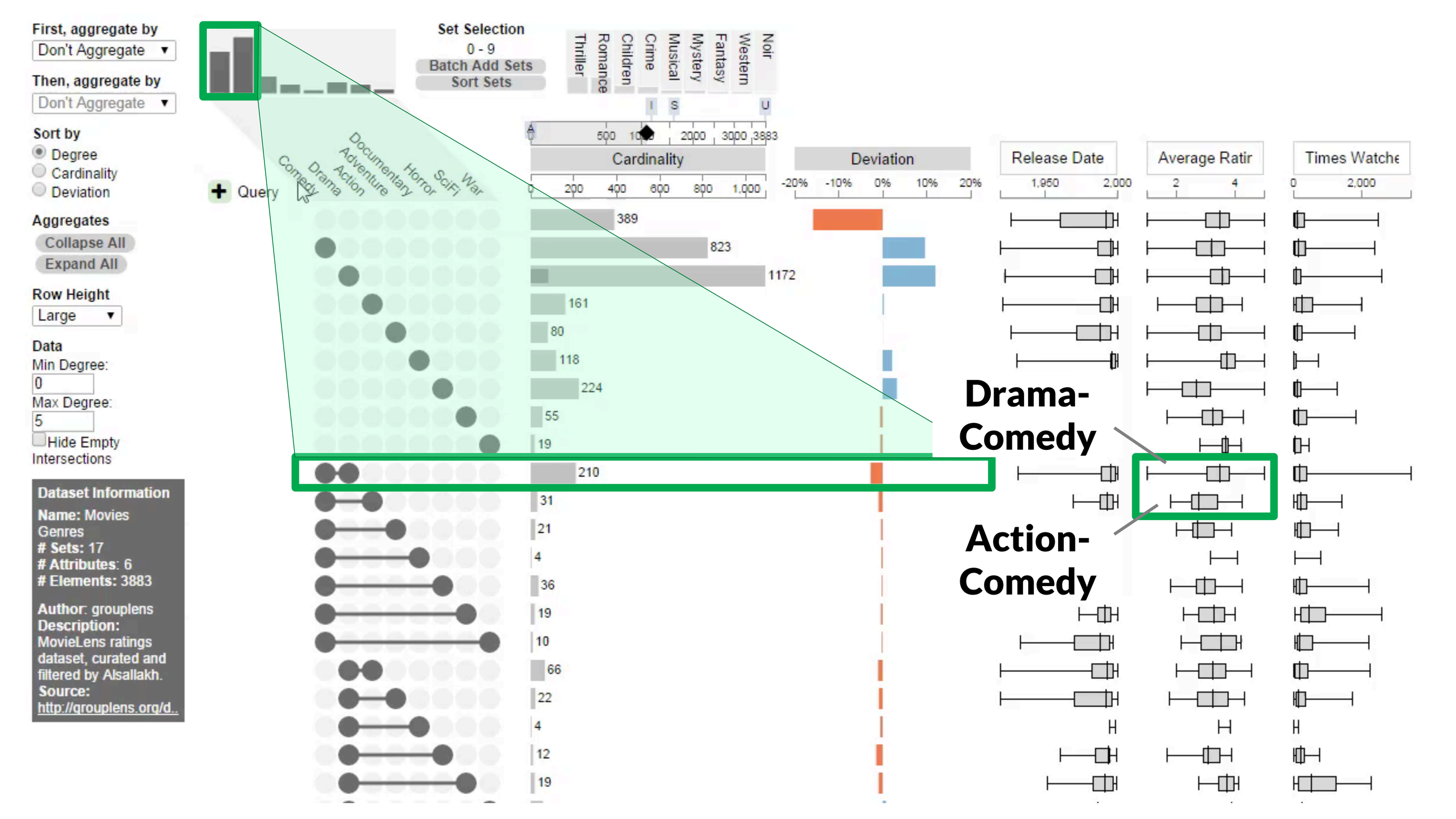

Universal Set
A
B
C


A
B
C
Cardinality
5
17
7
10
14
20
7
5

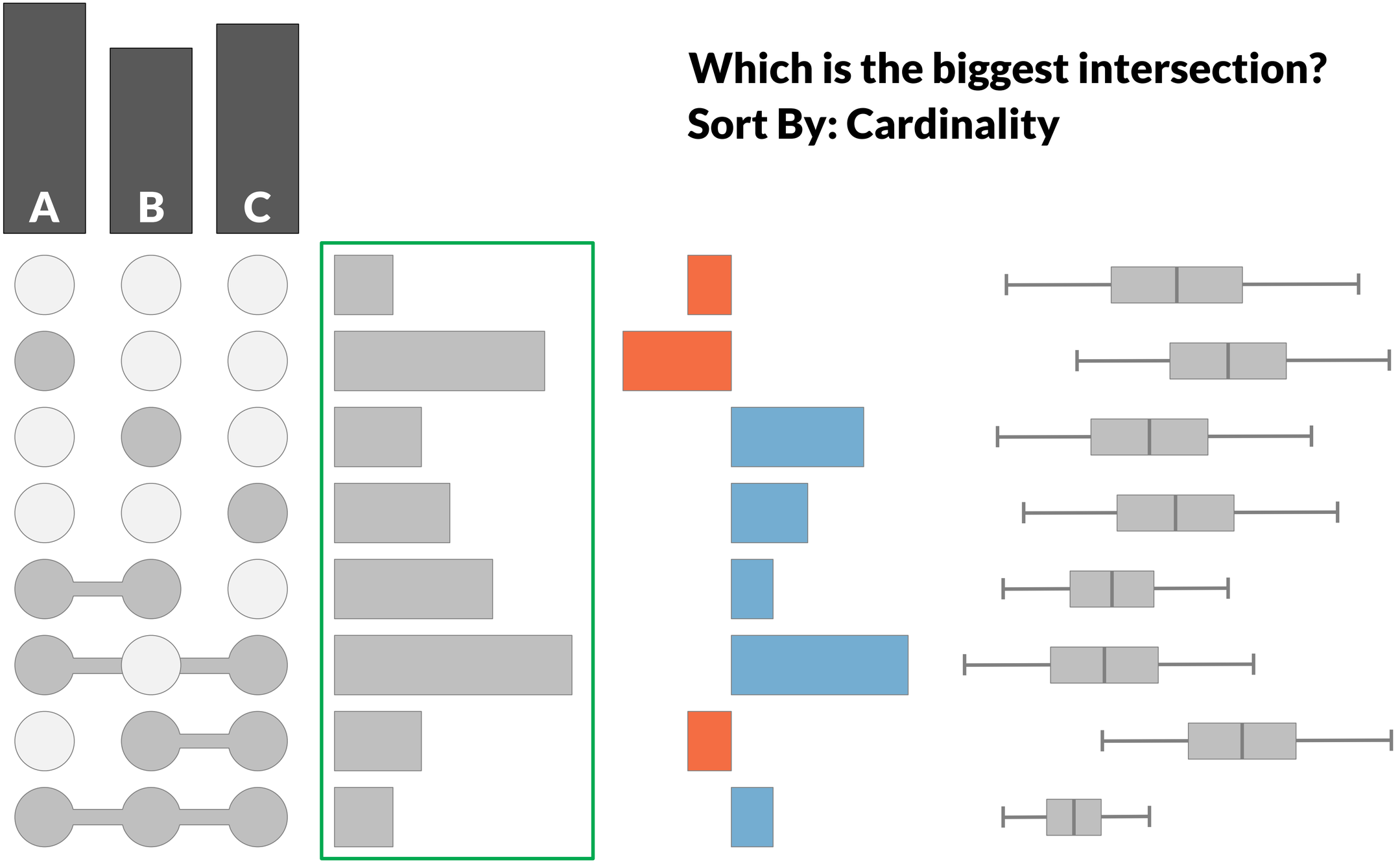

5
17
14
20
5
7
7
10

Plotting Attributes

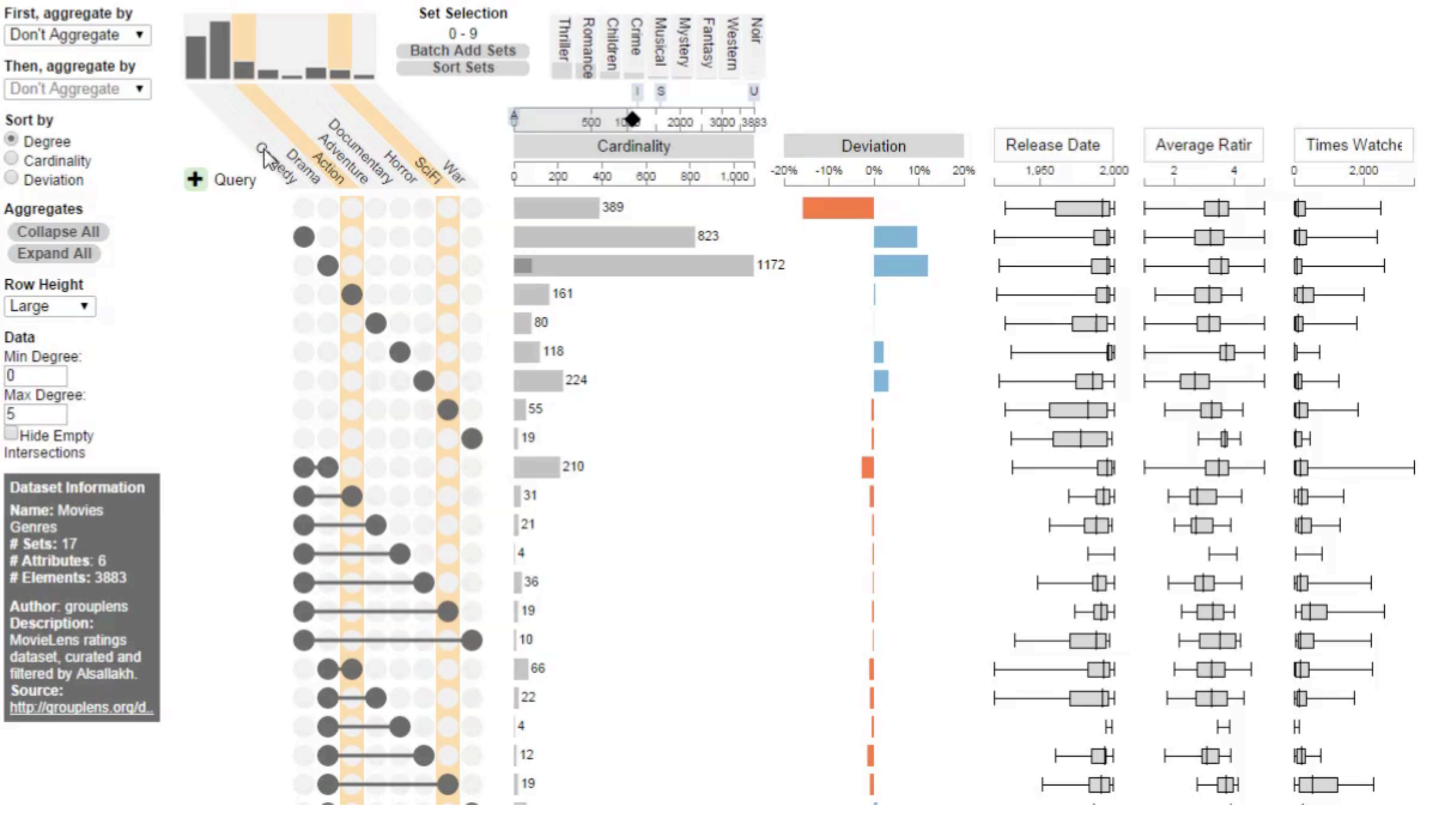

A
B
C
What's the size of an intersection?
What's the distribution of an attribute in an intersection?
Additional Plots
Deviation
Attributes


First, aggregate by
Don't Aggregate
Then, aggregate by
Don't Aggregate
Sort by
Degree
Cardinality
Deviation
Aggregates
Collapse All
Expand All
Row Height
Large
Data
Min Degree:
0
Max Degree:
5
Hide Empty Intersections
Dataset Information
Name: Movies Genres
Sets: 17
Attributes: 6
Elements: 3883
Author: grouplens
Description: MovieLens ratings dataset, curated and filtered by Alsallakh.
Source: http://grouplens.org/d..
Set Selection
0 - 9
Batch Add Sets
Sort Sets
Thriller
Romance
Children
Crime
Musical
Mystery
Fantasy
Western
Noir
Comedy
Drama
Action
Adventure
Documentary
Horror
SciFi
War
Query
Cardinality
Deviation
Release Date
Average Ratir
Times Watche
389
823
1172
161
80
118
224
55
19
210
31
21
4
36
19
10
66
22
4
12
19
Drama-Comedy
Action-Comedy

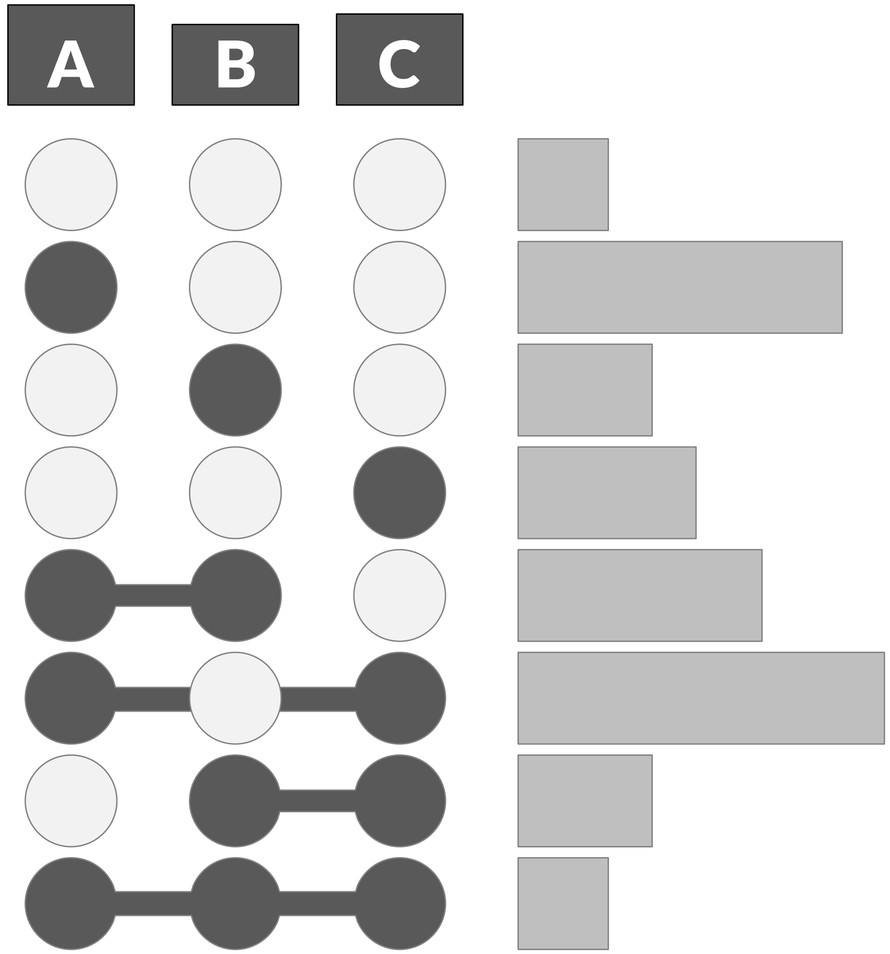
Sorting



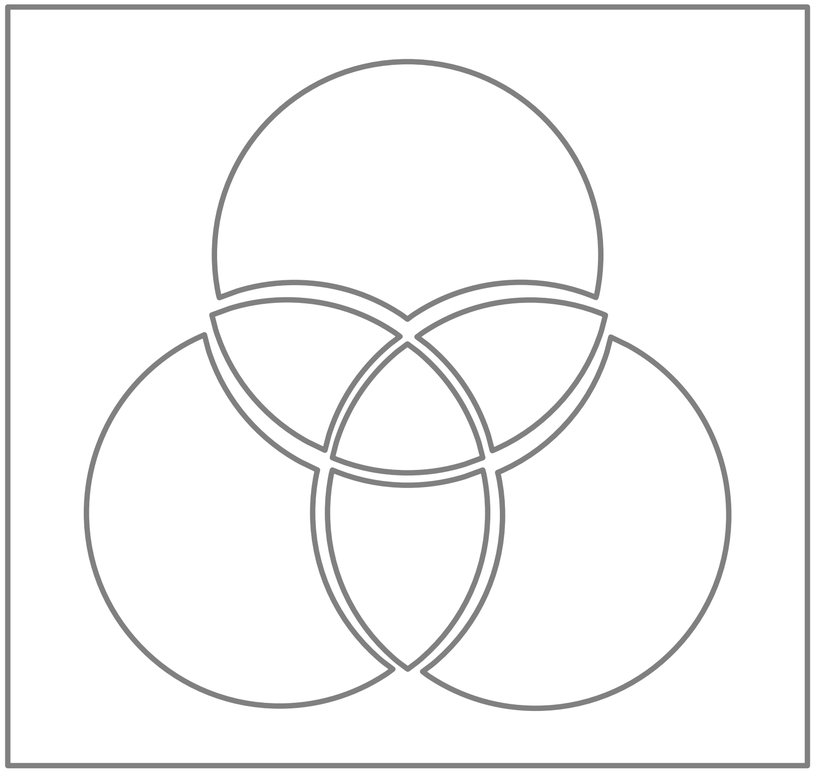
Which is the biggest intersection?
Sort By: Cardinality

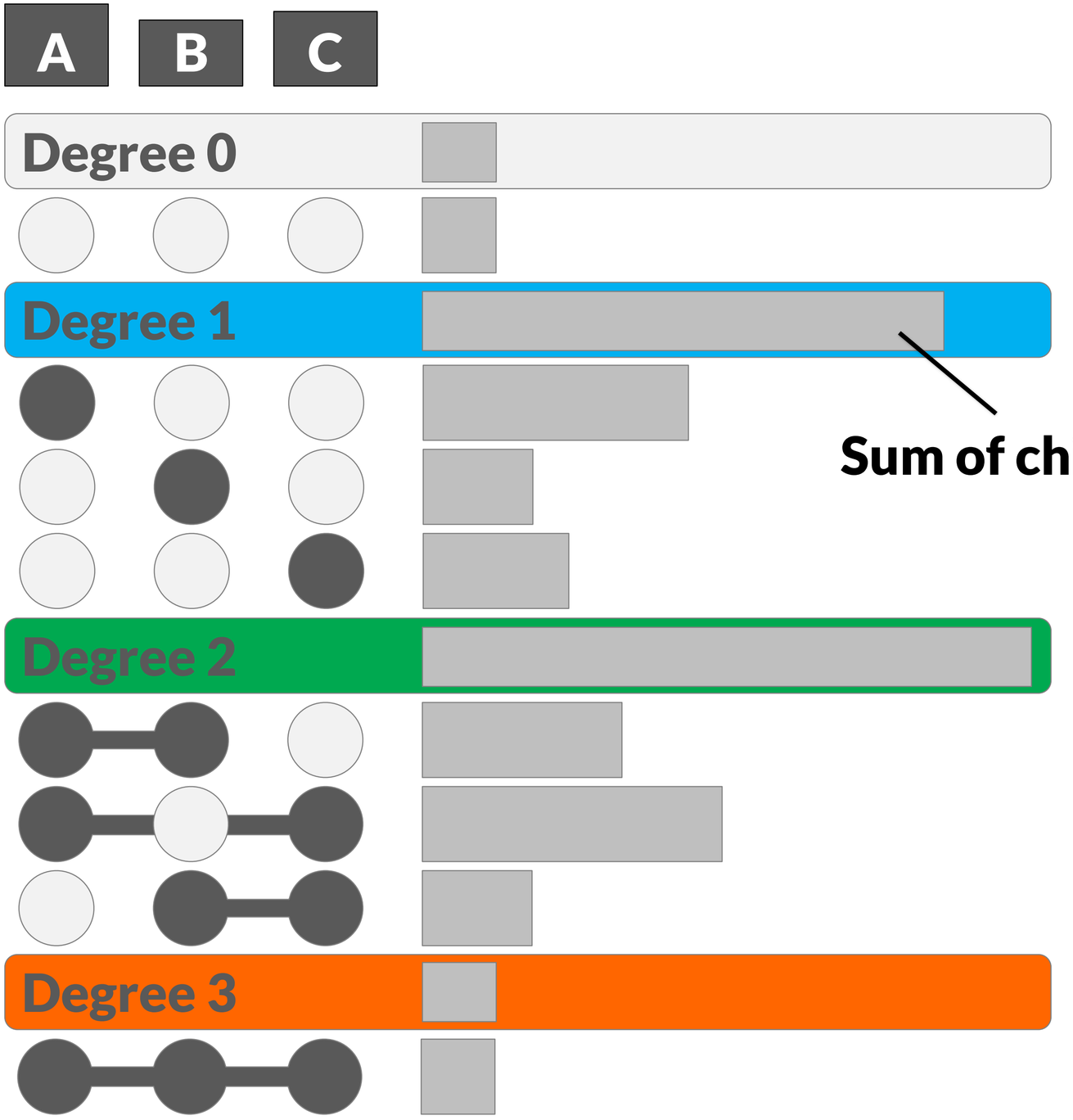

First, aggregate by
Don't Aggregate
Then, aggregate by
Don't Aggregate
Sort by
Degree
Cardinality
Deviation
Aggregates
Collapse All
Expand All
Row Height
Large
Data
Min Degree:
0
Max Degree:
5
Hide Empty Intersections
Dataset Information
Name: Movies Genres
Sets: 17
Attributes: 6
Elements: 3883
Author: grouplens
Description: MovieLens ratings dataset, curated and filtered by Alsallakh.
Source: http://grouplens.org/d..
Set Selection
0 - 9
Batch Add Sets
Sort Sets
Thriller
Romance
Children
Crime
Musical
Mystery
Fantasy
Western
Noir
Comedy
Drama
Action
Adventure
Documentary
Horror
SciFi
War
Query
Cardinality
0
200
400
600
800
1,000
389
823
1172
161
80
118
224
55
19
210
31
21
4
36
19
10
66
22
4
12
19
Deviation
-20%
-10%
0%
10%
20%
Release Date
1,950
2,000
Average Ratir
2
4
Times Watche
0
2,000

Aggregation



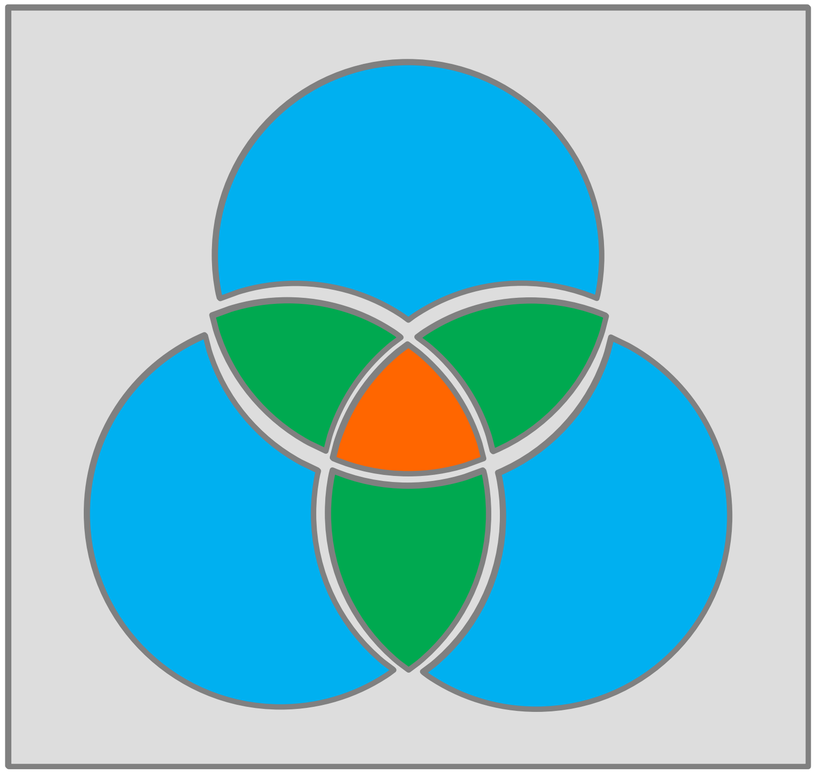
Are many items shared between two sets?
Aggregate By: Degree



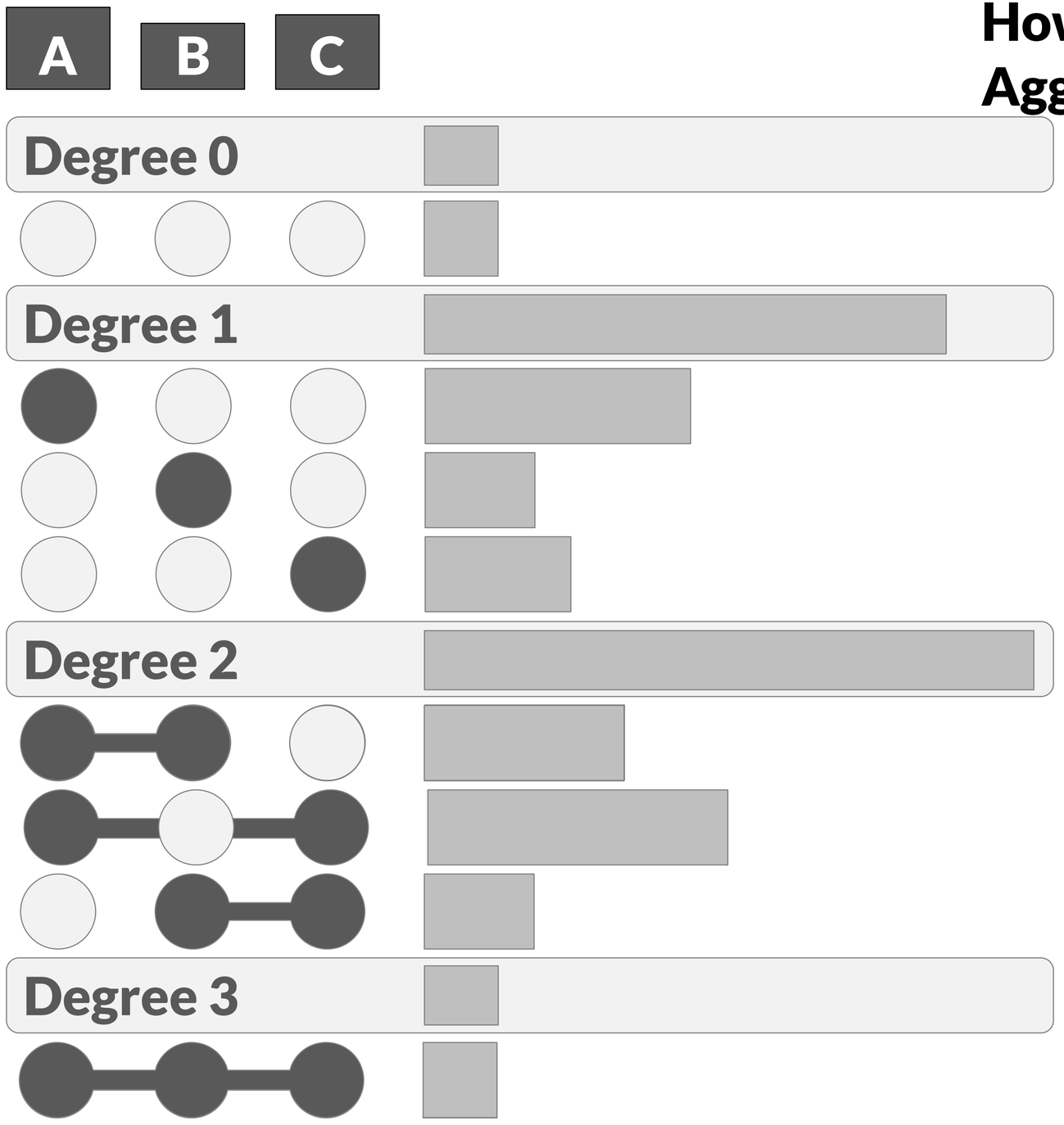


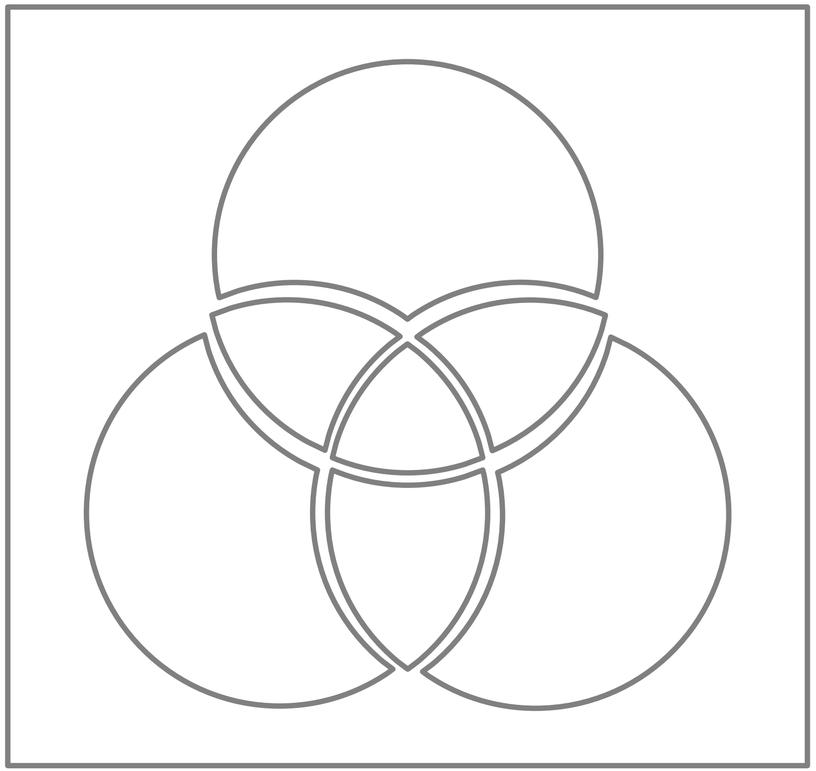
Are many items shared between two sets?
Aggregate By: Degree

Sum of children

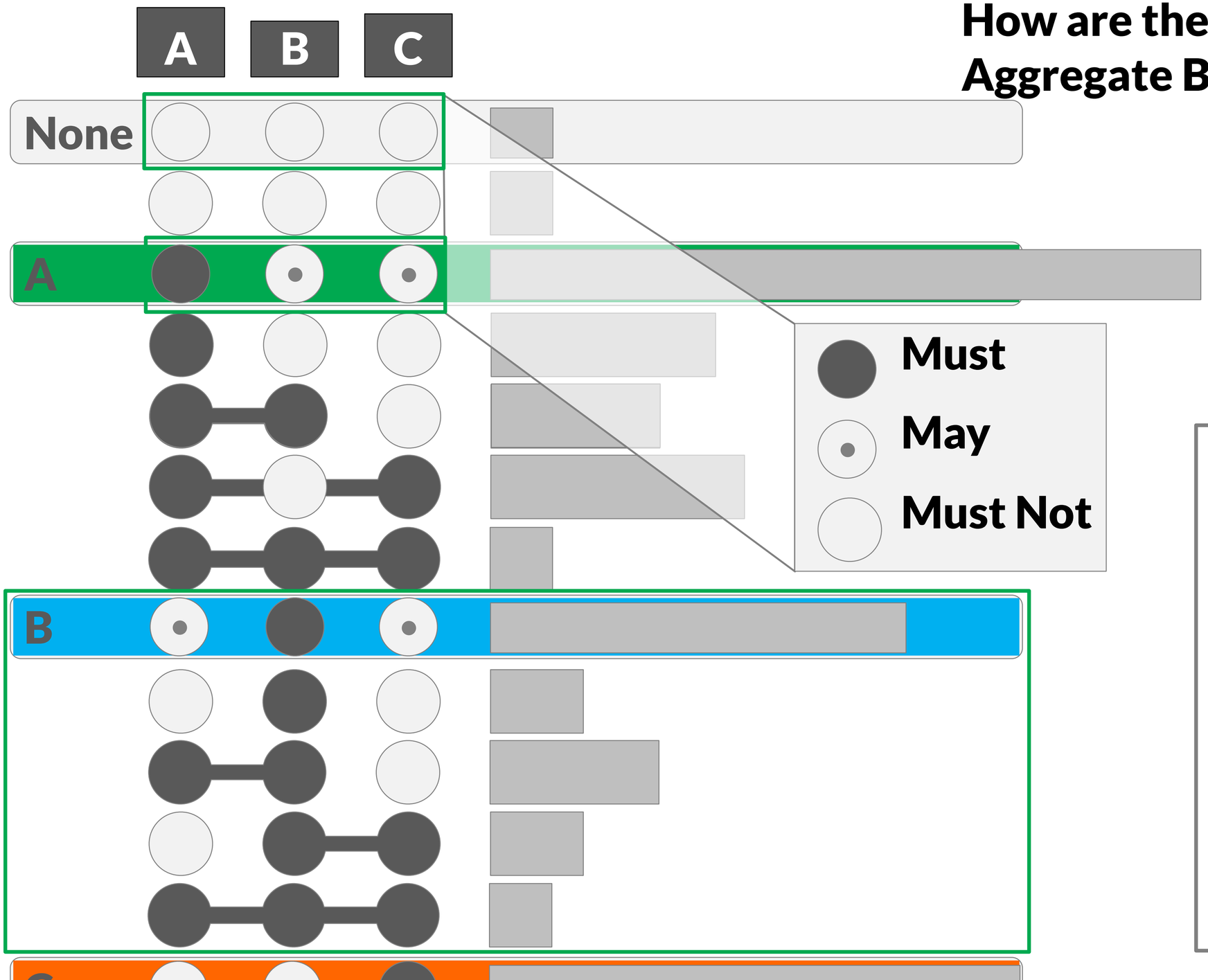


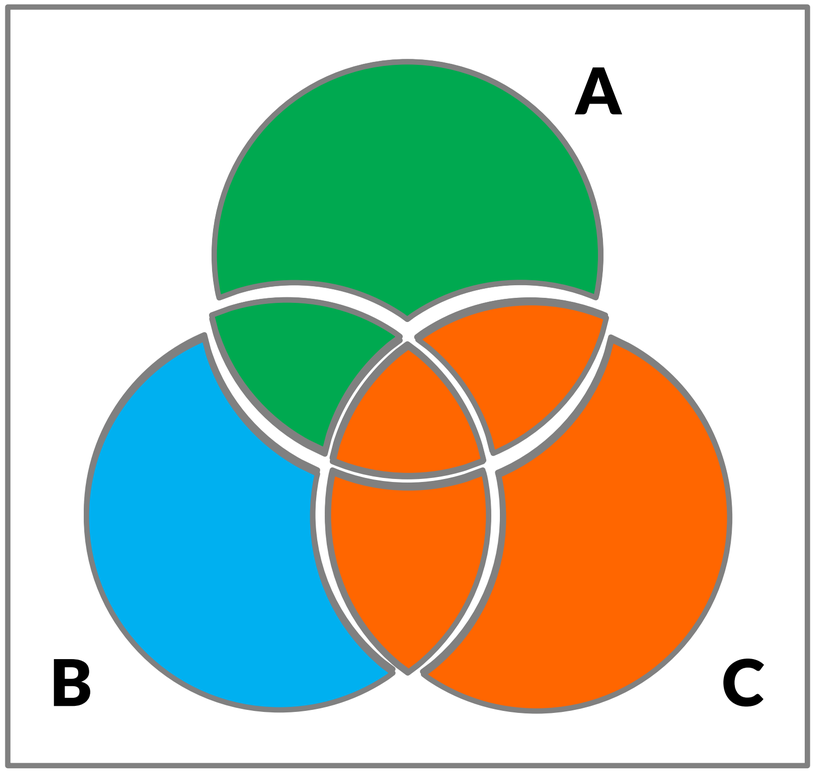
How are the elements of ‘B’ distributed?
Aggregate By: Set



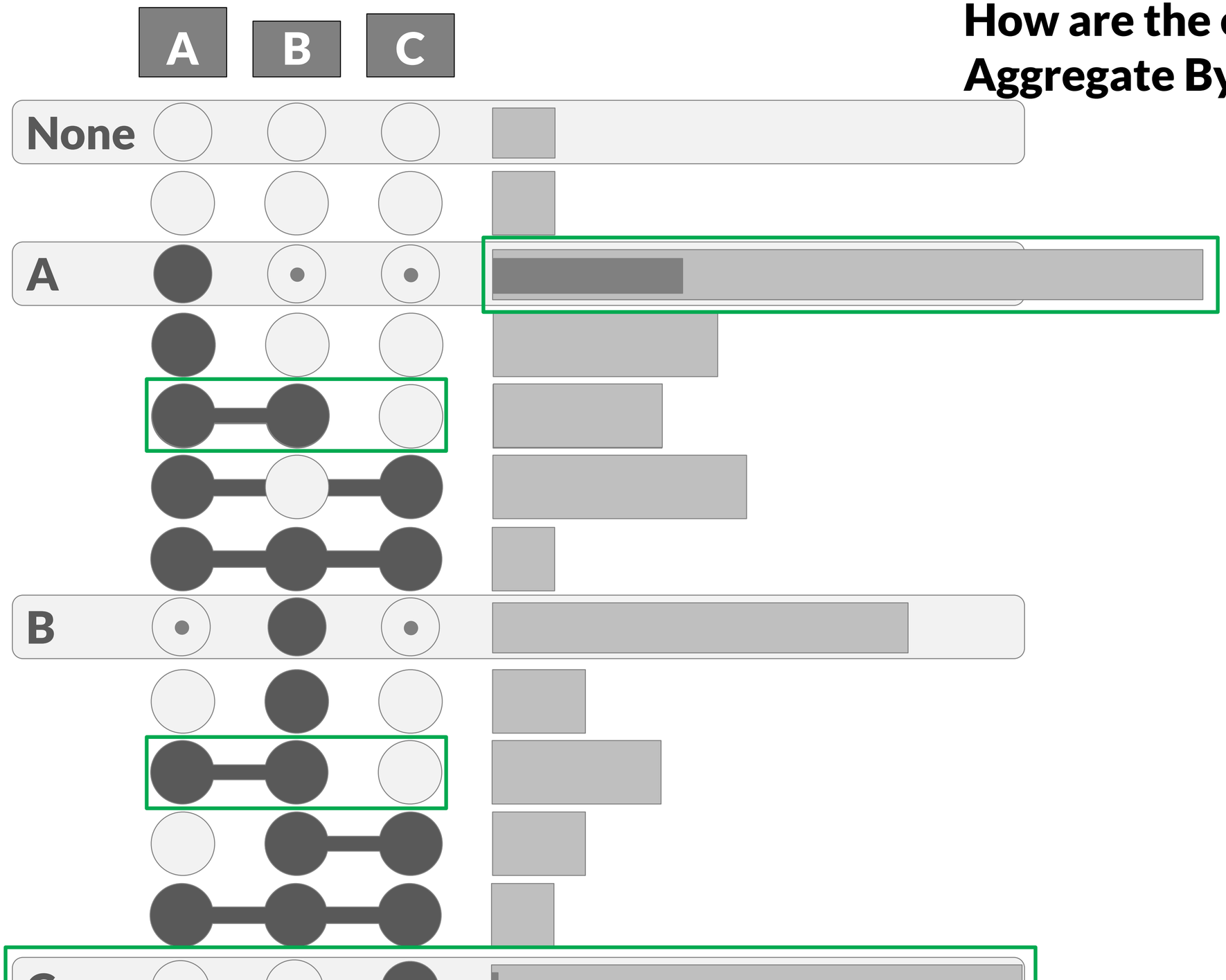


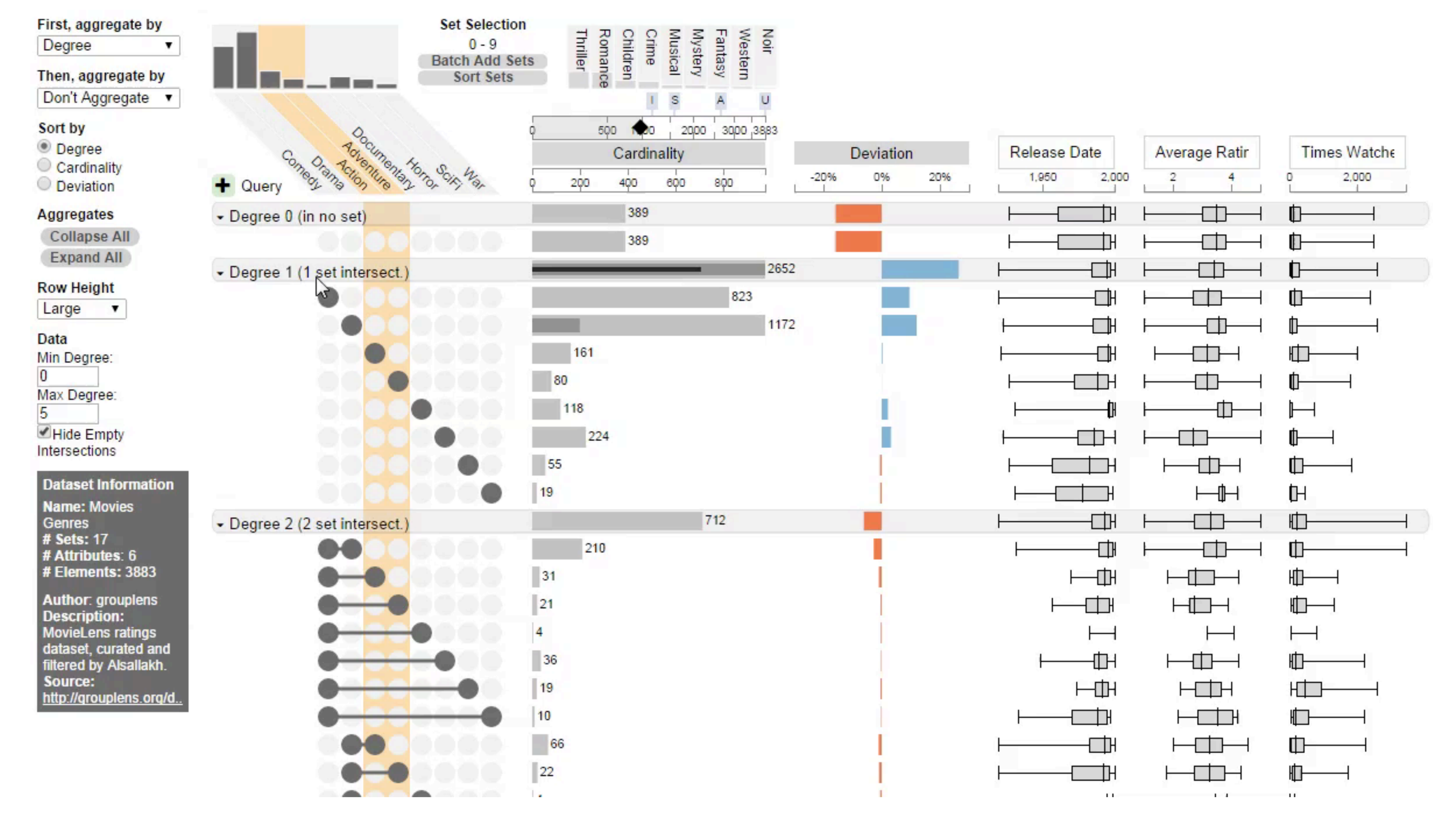
How are the elements of ‘B’ distributed?
Aggregate By: Set



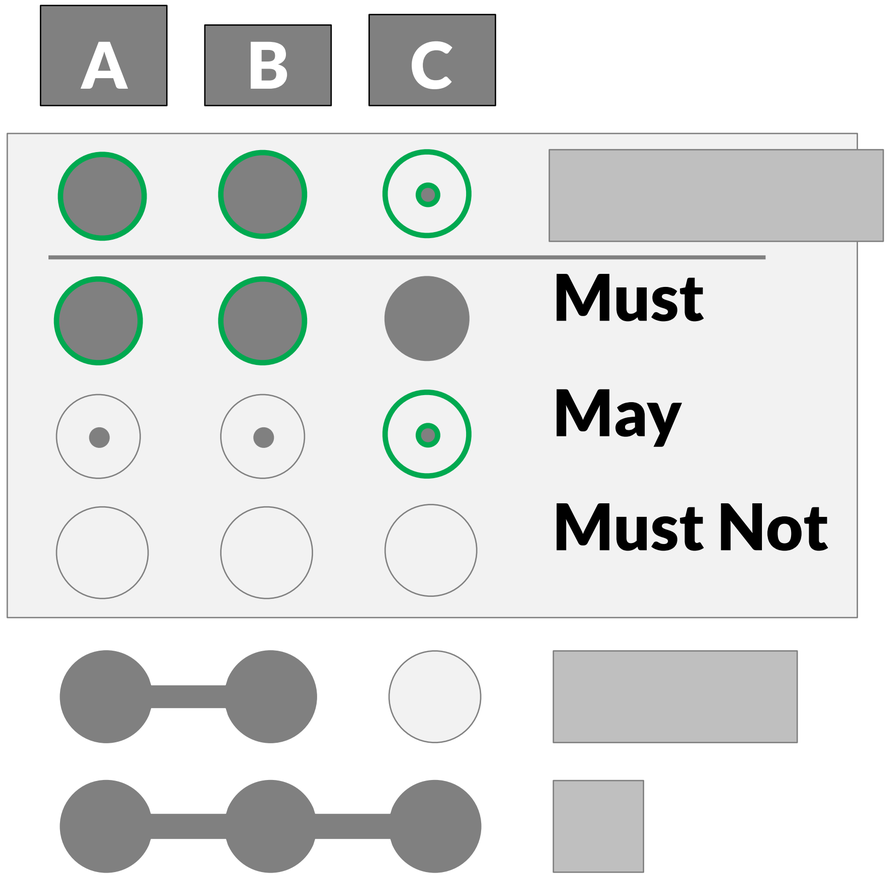


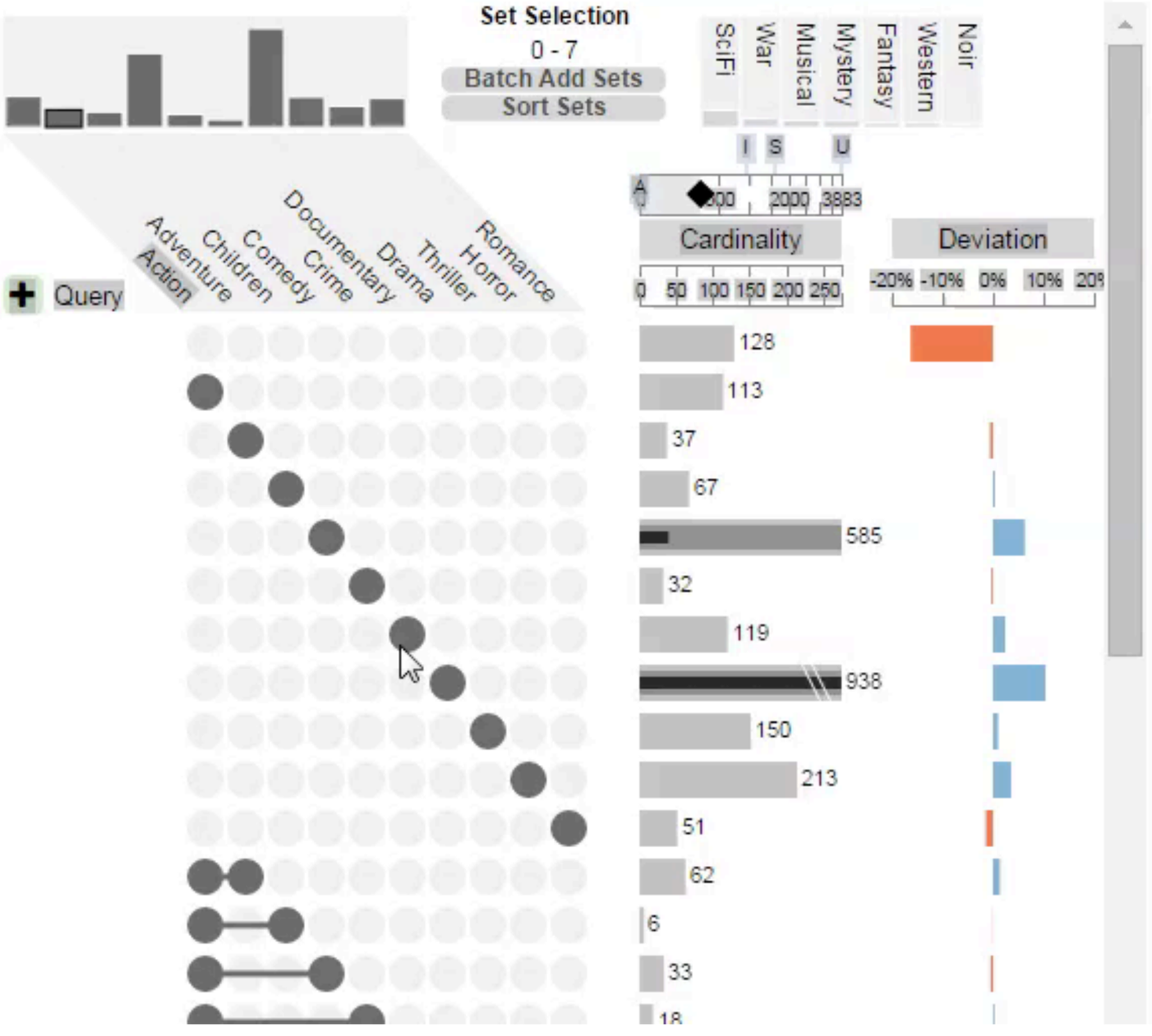
How are the elements of ‘B’ distributed?
Aggregate By: Set




First, aggregate by
Degree
Then, aggregate by
Don't Aggregate
Sort by
Degree
Cardinality
Deviation
Aggregates
Collapse All
Expand All
Row Height
Large
Data
Min Degree:
0
Max Degree:
5
Hide Empty Intersections
Dataset Information
Name: Movies Genres
Sets: 17
Attributes: 6
Elements: 3883
Author: grouplens
Description: MovieLens ratings dataset, curated and filtered by Alsallakh.
Source: http://grouplens.org/d..
Set Selection
0 - 9
Batch Add Sets
Sort Sets
Thriller
Romance
Children
Crime
Musical
Mystery
Fantasy
Western
Noir
Comedy
Drama
Action
Adventure
Documentary
Horror
SciFi
War
Query
Cardinality
Deviation
Release Date
Average Ratir
Times Watche
Degree 0 (in no set)
389
389
Degree 1 (1 set intersect.)
2652
823
1172
161
80
118
224
55
19
Degree 2 (2 set intersect.)
712
210
31
21
4
36
19
10
66
22

Queries


A
B
C
Must
May
Must Not



Element Visualizations

No visualizations configured. Click + button to add a new visua

Scatterplot

Element Queries

No queries. Click + button to add a new query.

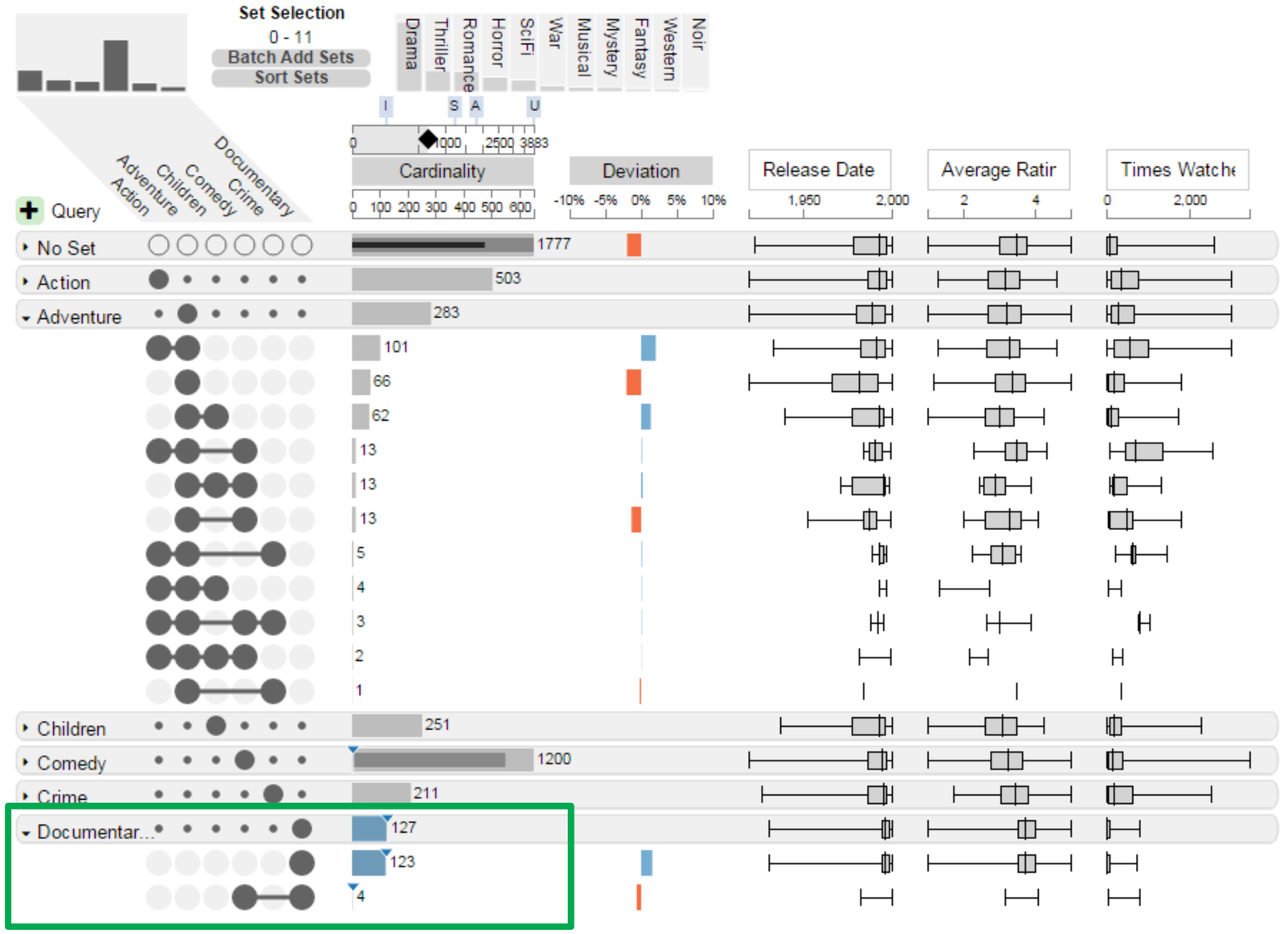
Query Filters

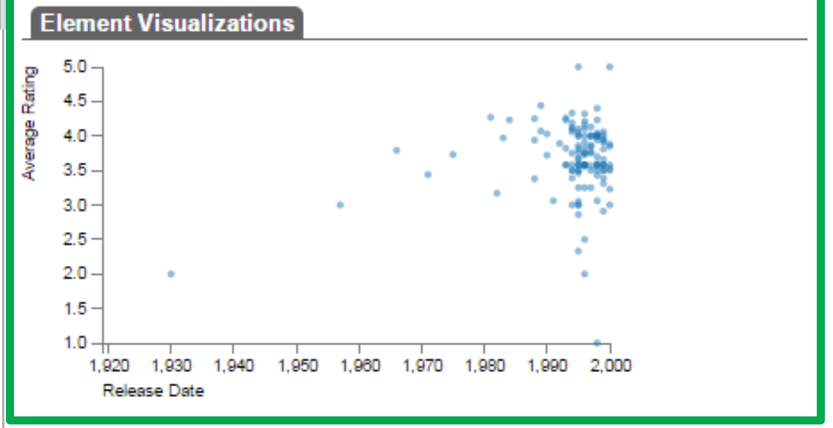
No active query.

Query Results

No active query.

Elements & Attributes

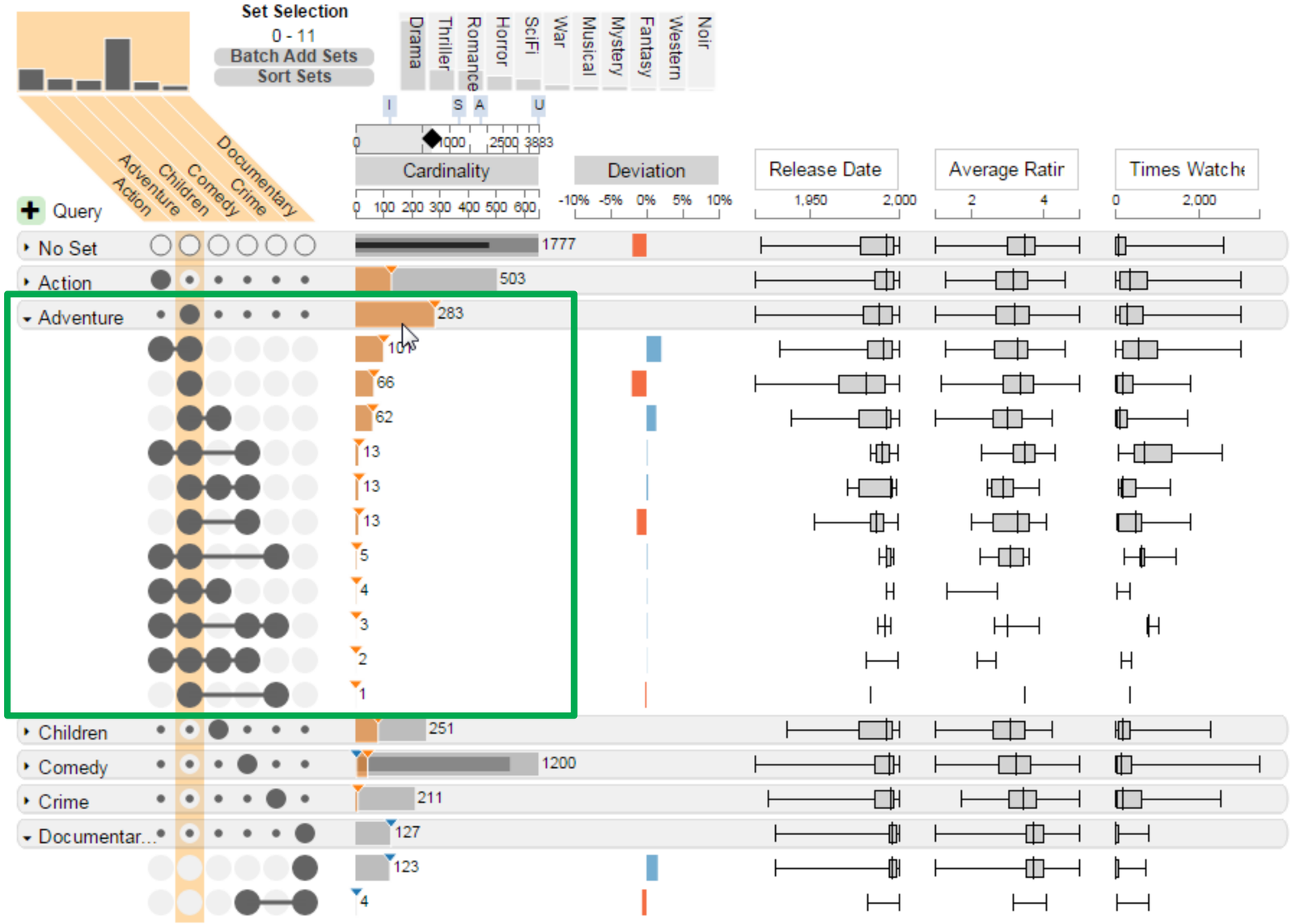




Query Results

Name	Release Date
Across the Sea of Time (1995)	1995
Nico Icon (1995)	1995
Heidi Fleiss: Hollywood Madam (1995)	1995
Catwalk (1995)	1995
Anne Frank Remembered (1995)	1995
Jupiter's Wife (1994)	1994
Sonic Outlaws (1995)	1995
From the Journals of Jean Seberg (1995)	1995
Man of the Year (1995)	1995

How do documentaries compare to adventure movies?



Element Queries

127 283

Query Filters

Subset | Sets

Name Contains

Query Results

Name	Release Date
Jumanji (1995)	1995
Tom and Huck (1995)	1995
GoldenEye (1995)	1995
Cutthroat Island (1995)	1995
City of Lost Children, The (1995)	1995
Wings of Courage (1995)	1995
Mortal Kombat (1995)	1995
Kids of the Round Table (1995)	1995
Indian in the Cupboard, The (1995)	1995
White Squall (1996)	1996
Muppet Treasure Island (1996)	1996

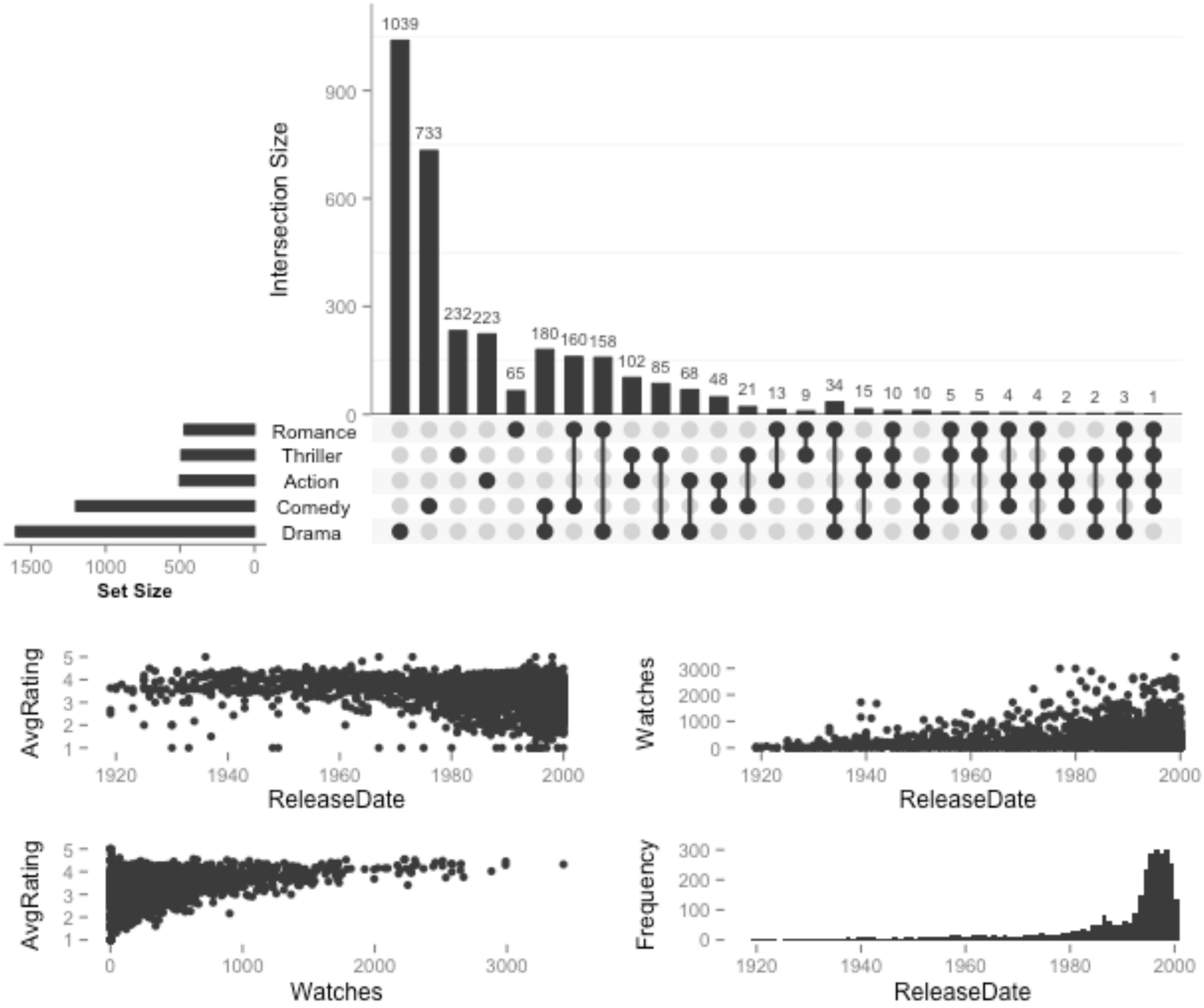
How do documentaries compare to adventure movies?

Applications

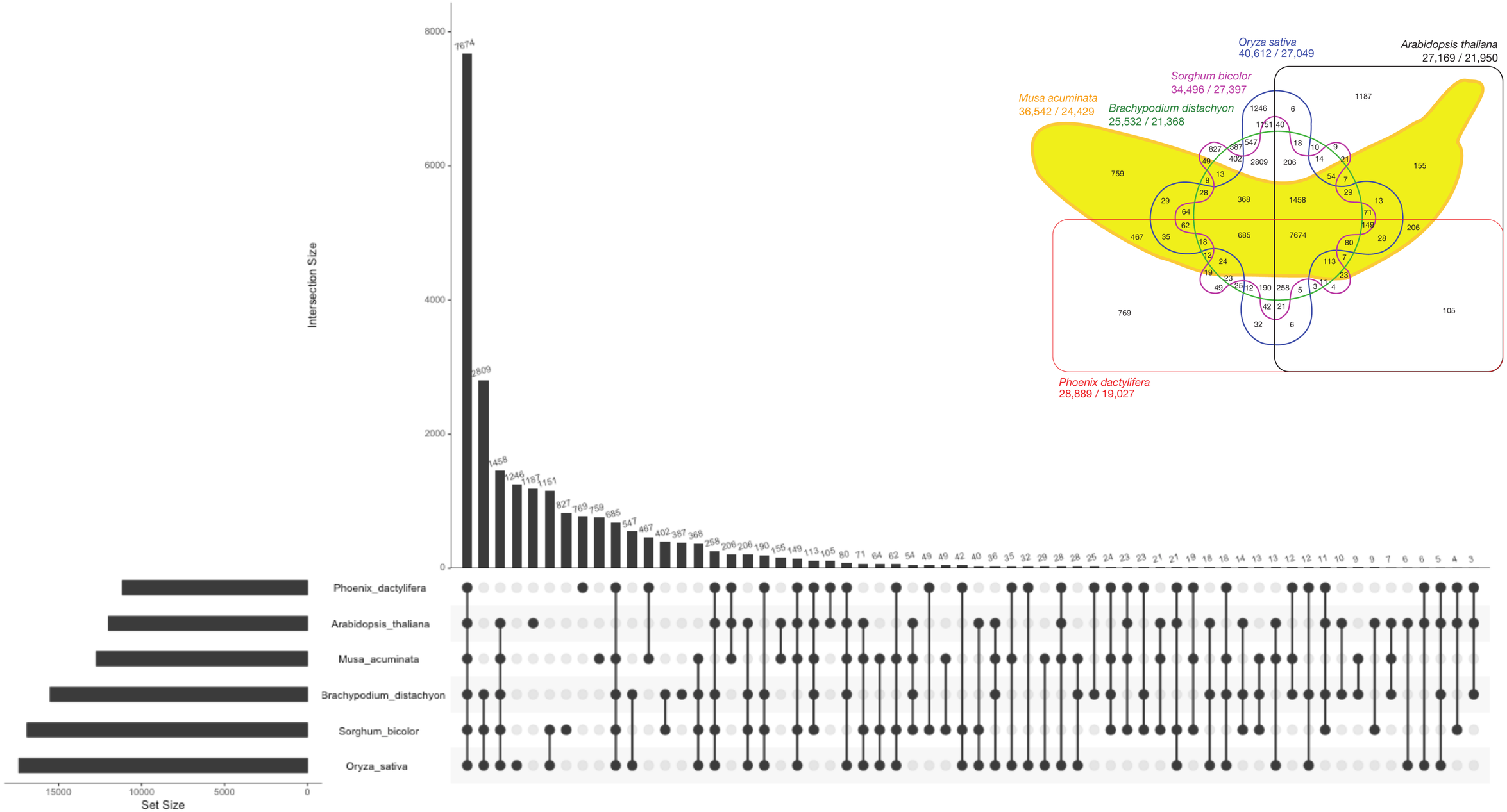
R-Version: UpSetR

Developed at HMS

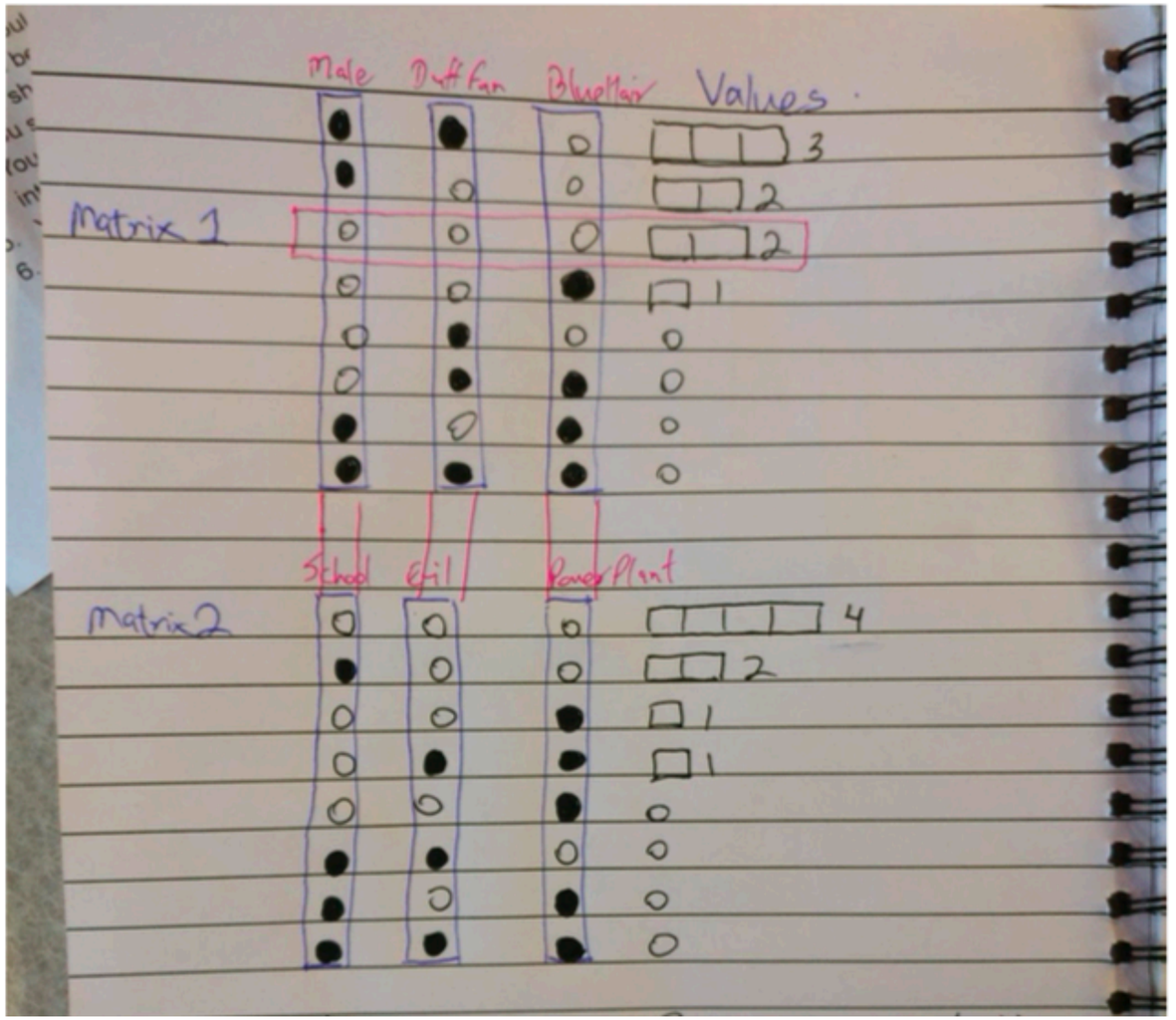
Some design adaptions



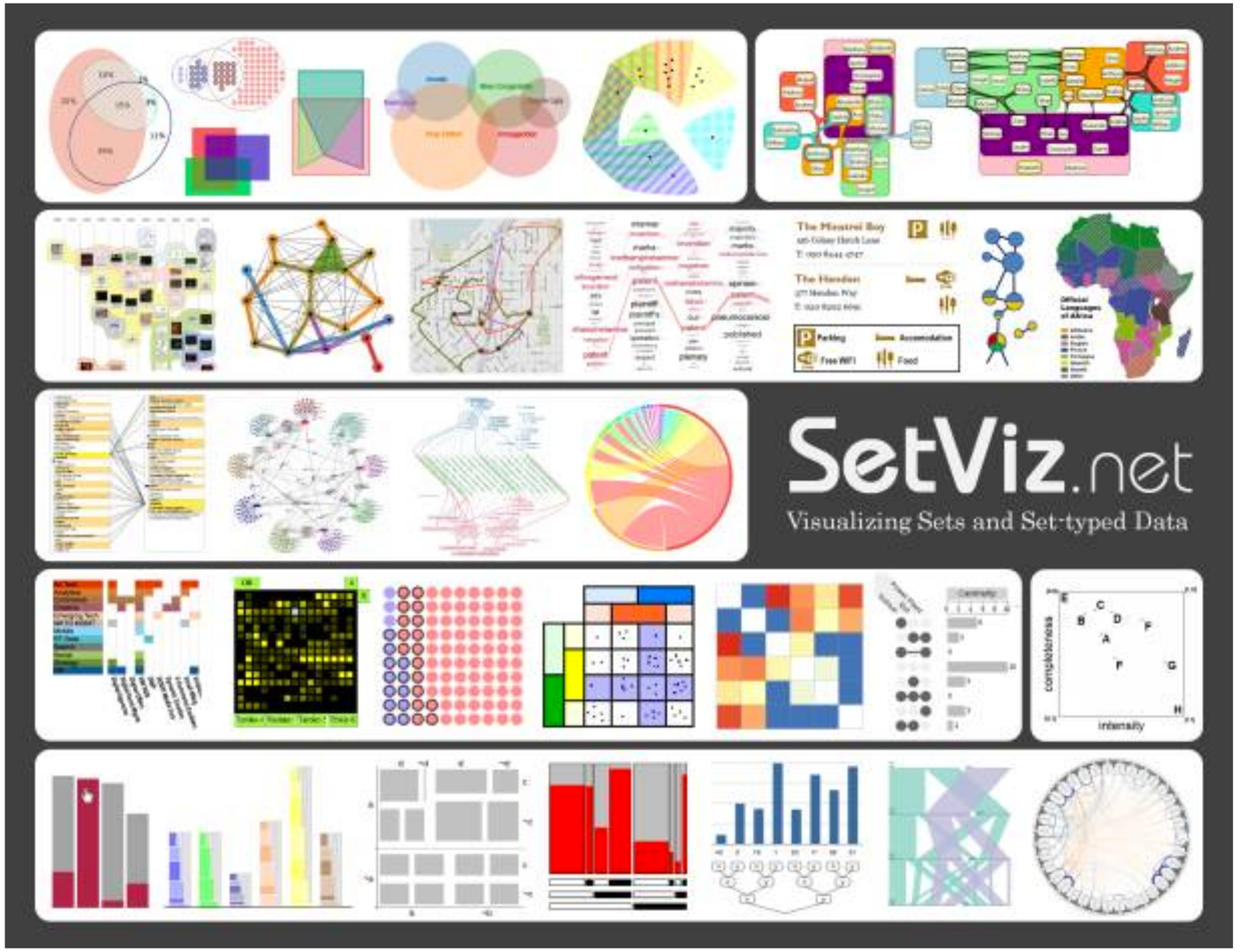
The Banana Chart Redesigned



DESIGN 2


Male
Diff Fan
BlueHair
Values
3
2
Matrix 1
2
1
School
Girl
Power Plant
Matrix 2
4
2
1
1

Other Options



http://setviz.net