Martin Hahmann, Markus Dumat, Dirk Habich and Wolfgang Lehner

Explorative Multi-View Clustering Using Frequent-Groupings

MultiClust 2012
28.04.2012

> Two sides of Multiple Clusterings

Out of many, one:

- integrate multiple solution into a single result
- ensemble clustering
- multi-source clustering

Out of one, many:

- generate different results from the same source
- alternative clustering
- subspace clustering

Our goal:

- bring both sides together using frequent groupings

Core Concept

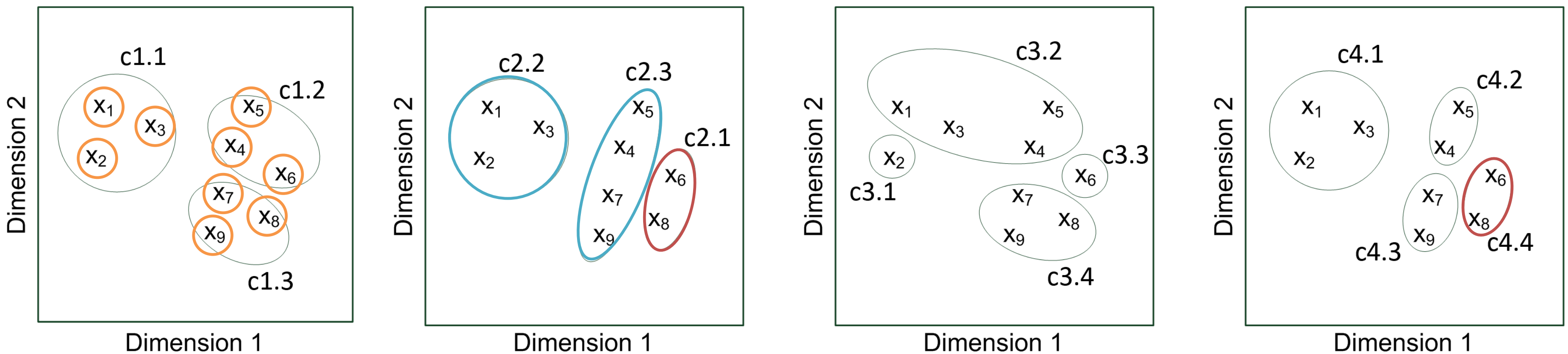
> Best of both worlds

Frequent Groupings Idea (MultiClust2011)

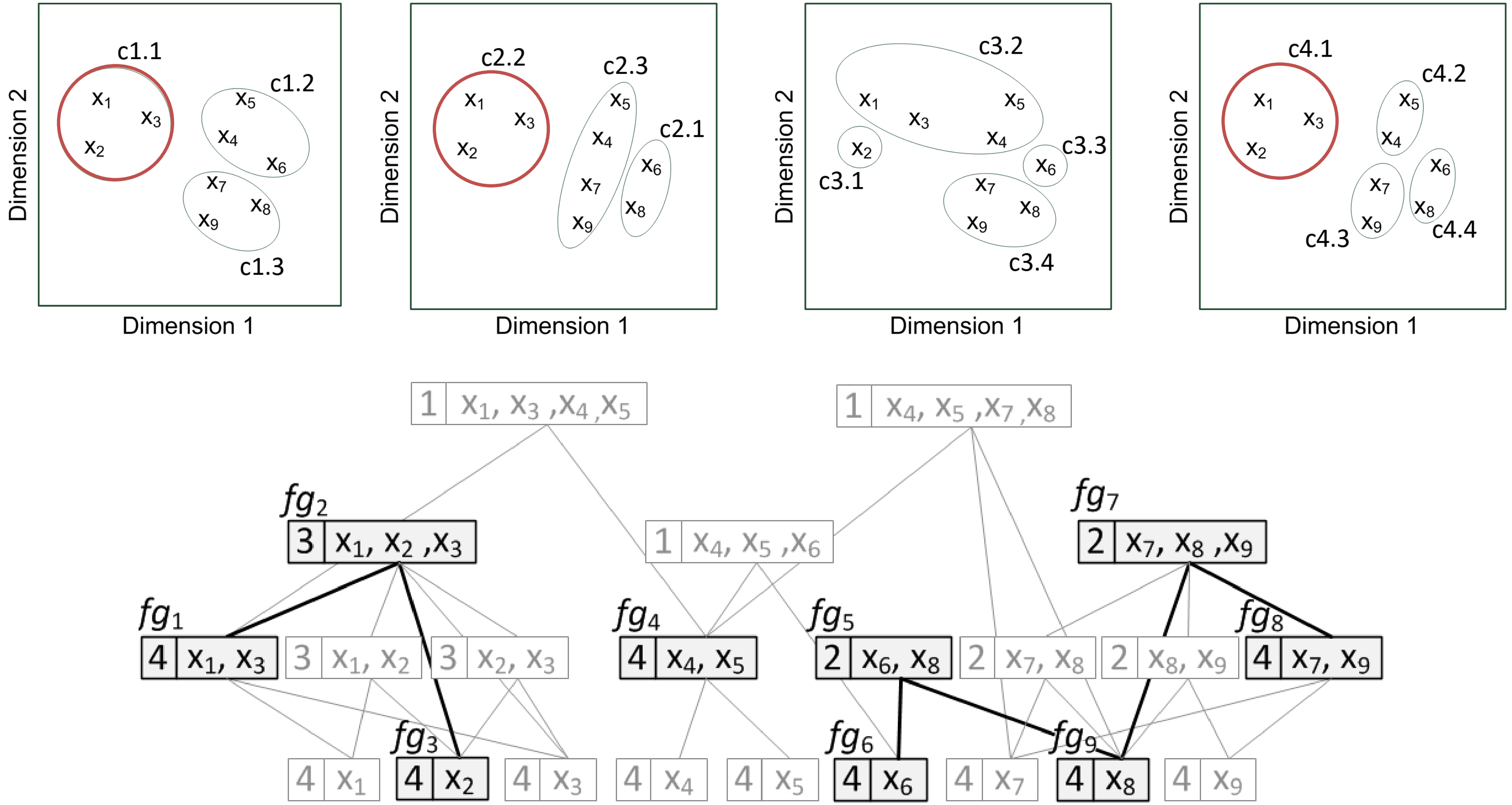
- starting point: ensemble clustering
- goal: find cluster assignment that agrees with majority of ensemble
- find objects frequently assigned to the same cluster

- → parallel to the concept of frequent itemsets
- mapping to ensemble clustering:

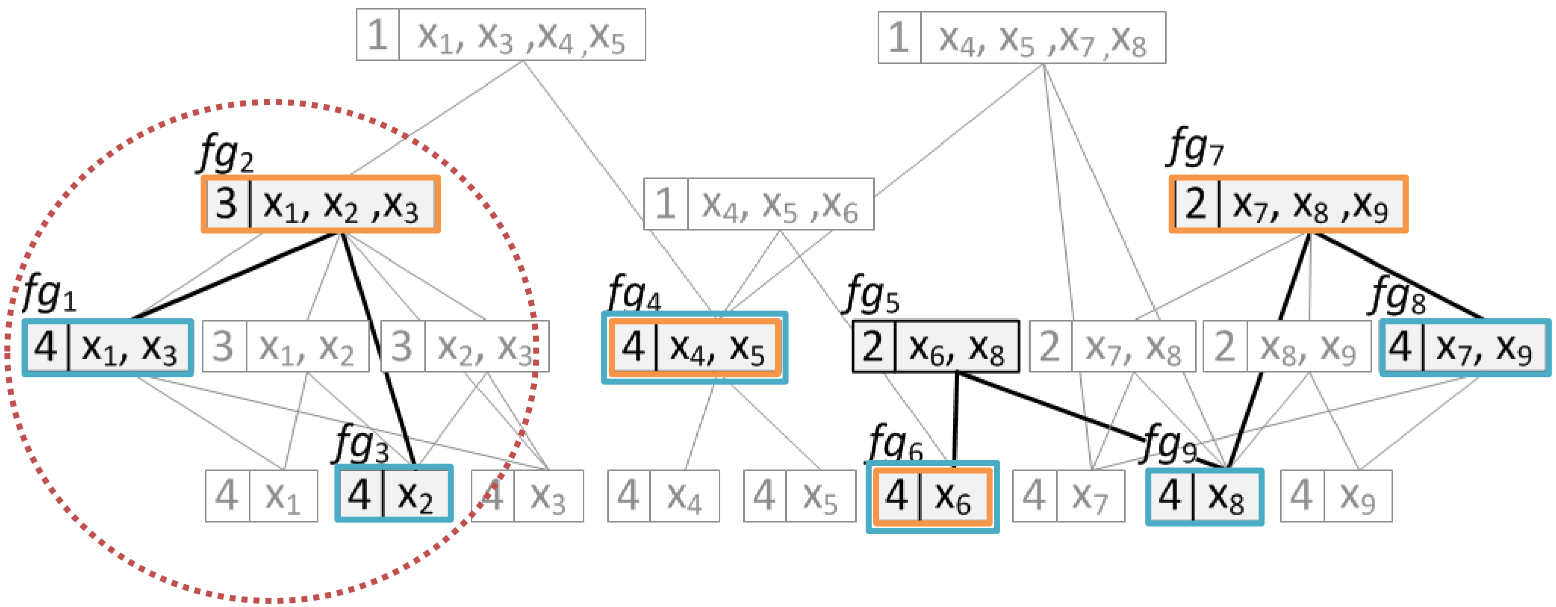
items *transaction* *support*



Frequent Grouping Graph

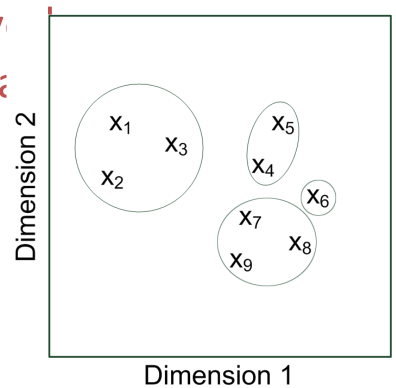


> Result Extraction

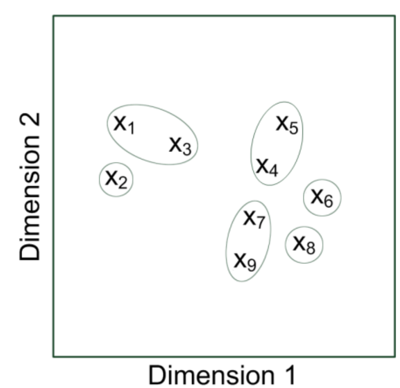


frequent groupings ov

→ alternative combina



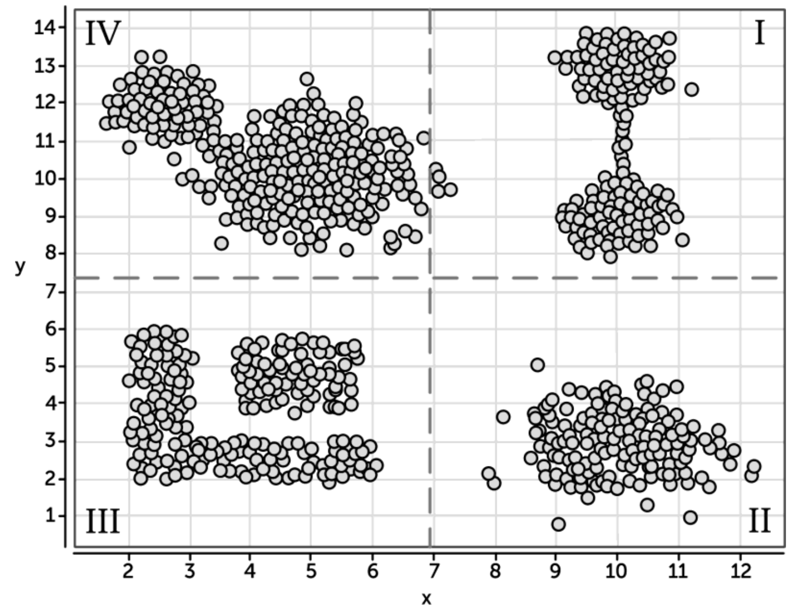
maximize cluster size



maximize support

Scaled Up Scenario

Manual processing not feasible



- 1500 objects, 10 clusterings, 69 clusters
- apply existing frequent itemset mining algorithms

- CARPENTER with minsupp= 0.5 → 72 frequent groupings

> Result Extraction

How to create alternative Solutions?

- assumption: every object is assigned to a cluster
- resembles exact-cover problem (np-hard)

First Approach

- Algorithm X by Donald E. Knuth
- implementation: Dancing Links

570.000 alternative clusterings

Result Extraction

Problems

- too many results to work with
- high similarity between alternatives
- smallest frequent groupings define minimum difference

Options for result limitation:

- filter small groupings
- compute similarity matrix

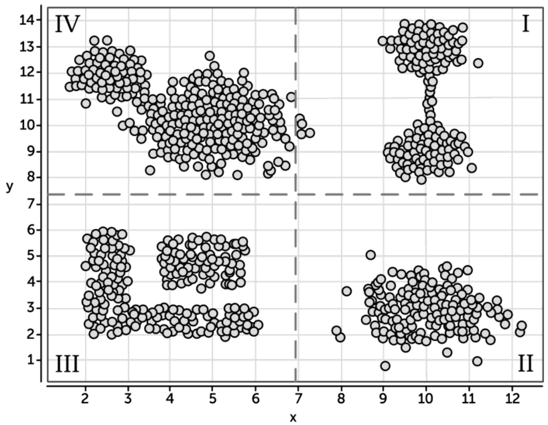
- top-k Dancing Links:

Top-k: 5

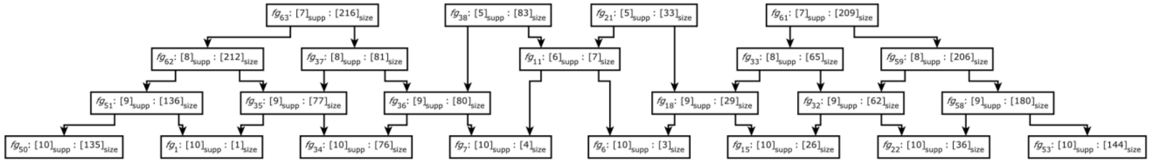
fg14	fg35	fg10	fg3	fg7	fg9	fg6	fg17	fg23	...

Clusterings: {(fg14, fg 35, fg7), (fg10, fg7), (fg3, fg7)}

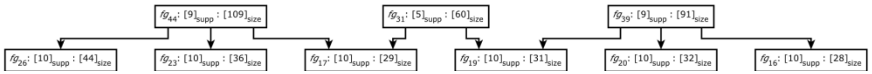
Second Approach: Exploring the graph



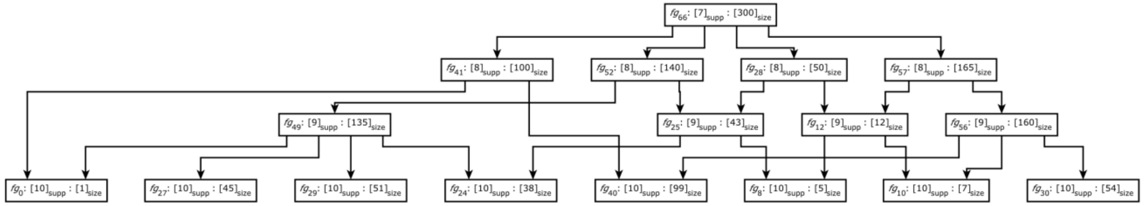
I



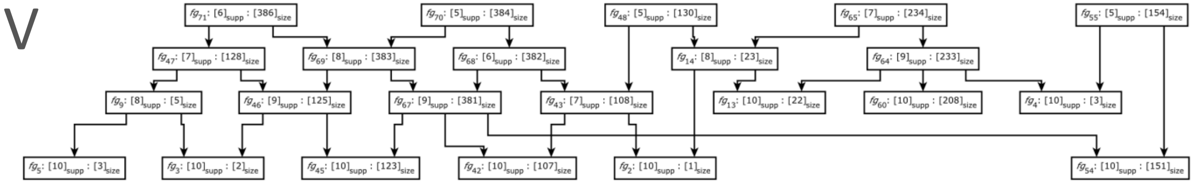
II



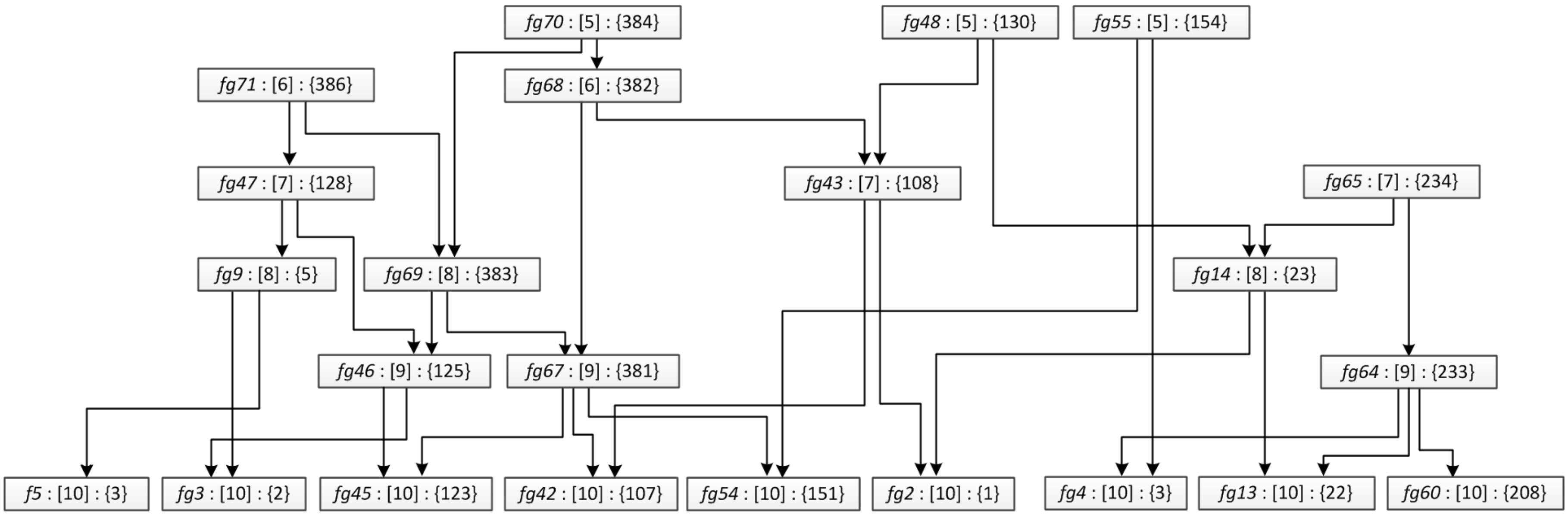
III

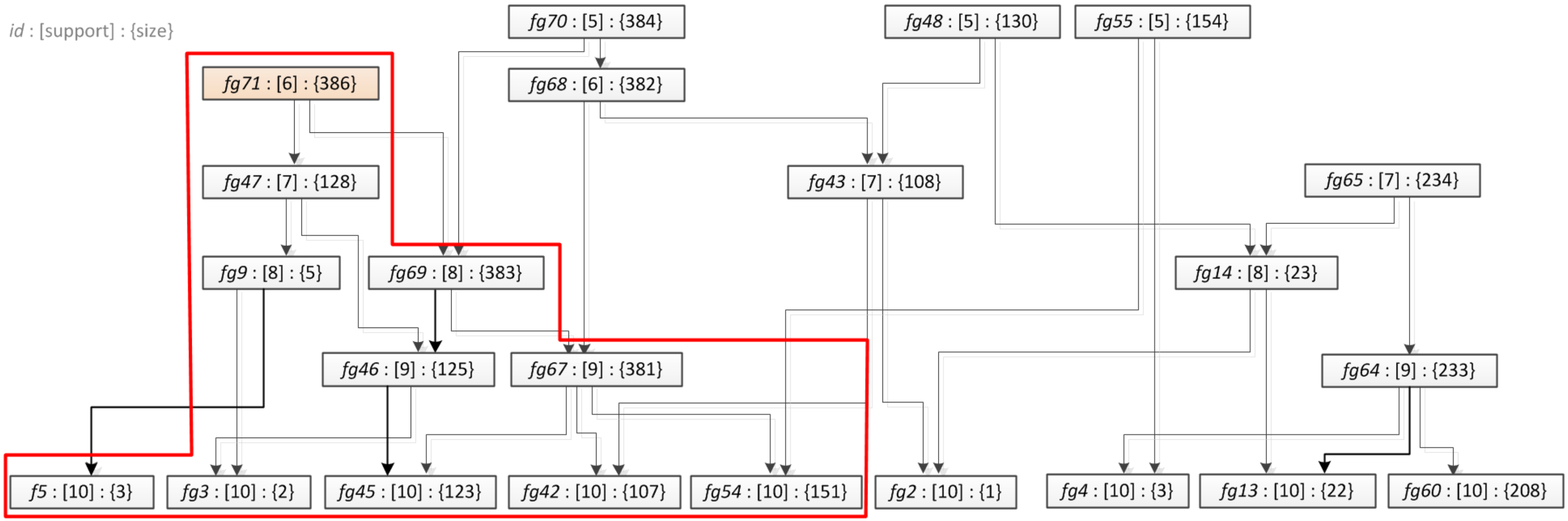


IV

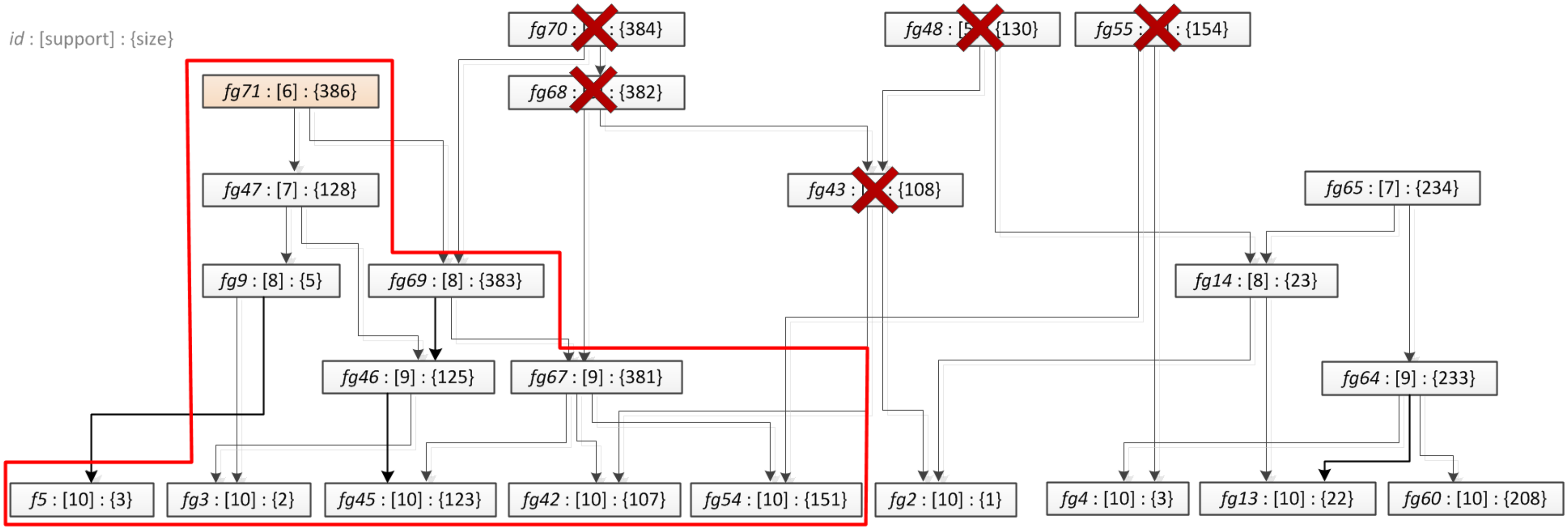


> Result Extraction

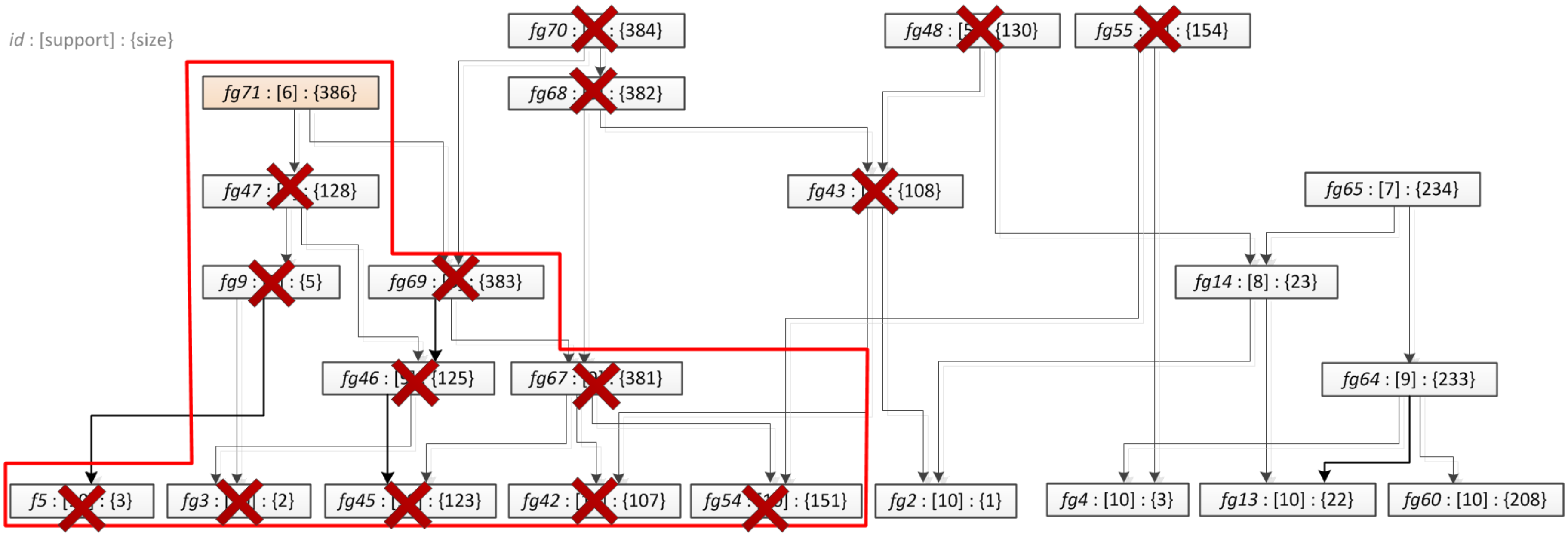


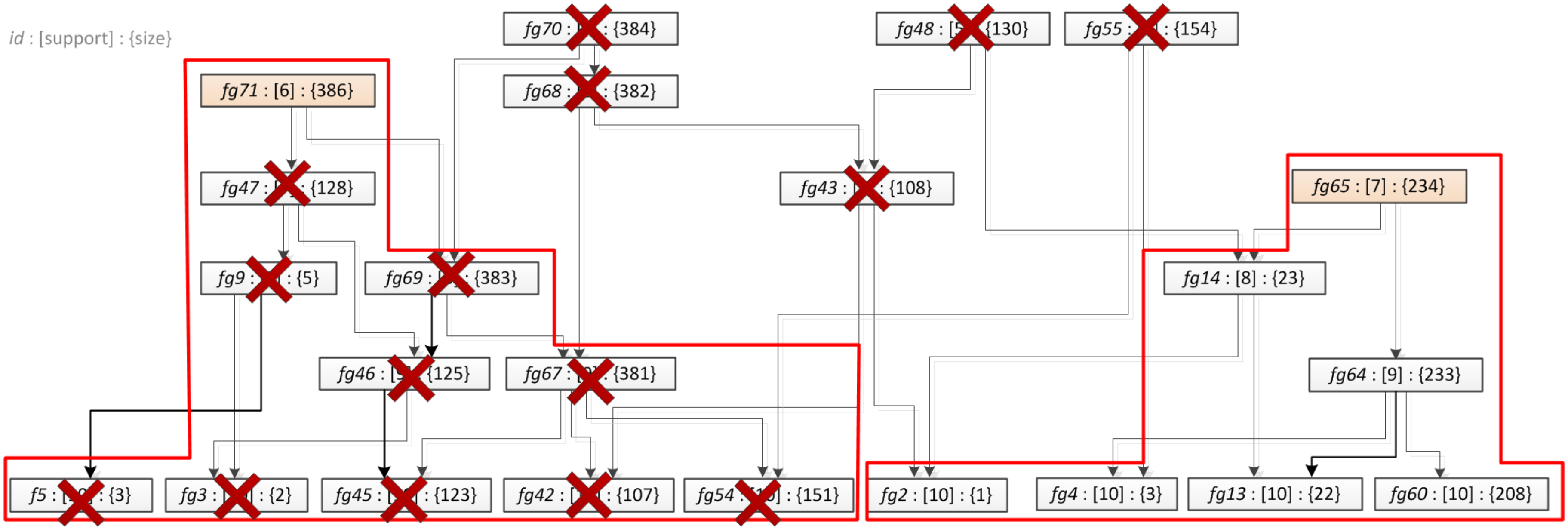

id : [support] : {size}
fg70 : [5] : {384}
fg48 : [5] : {130}
fg55 : [5] : {154}
fg71 : [6] : {386}
fg68 : [6] : {382}
fg47 : [7] : {128}
fg43 : [7] : {108}
fg65 : [7] : {234}
fg9 : [8] : {5}
fg69 : [8] : {383}
fg14 : [8] : {23}
fg46 : [9] : {125}
fg67 : [9] : {381}
fg64 : [9] : {233}
f5 : [10] : {3}
fg3 : [10] : {2}
fg45 : [10] : {123}
fg42 : [10] : {107}
fg54 : [10] : {151}
fg2 : [10] : {1}
fg4 : [10] : {3}
fg13 : [10] : {22}
fg60 : [10] : {208}

> Result Extraction

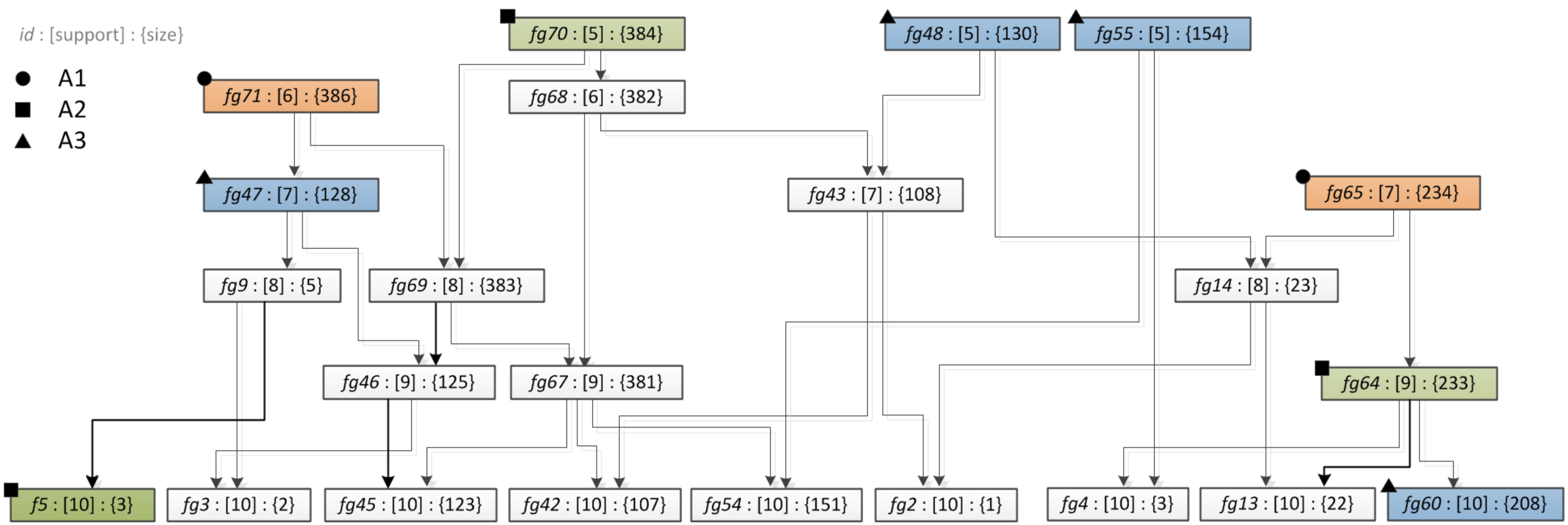


> Result Extraction



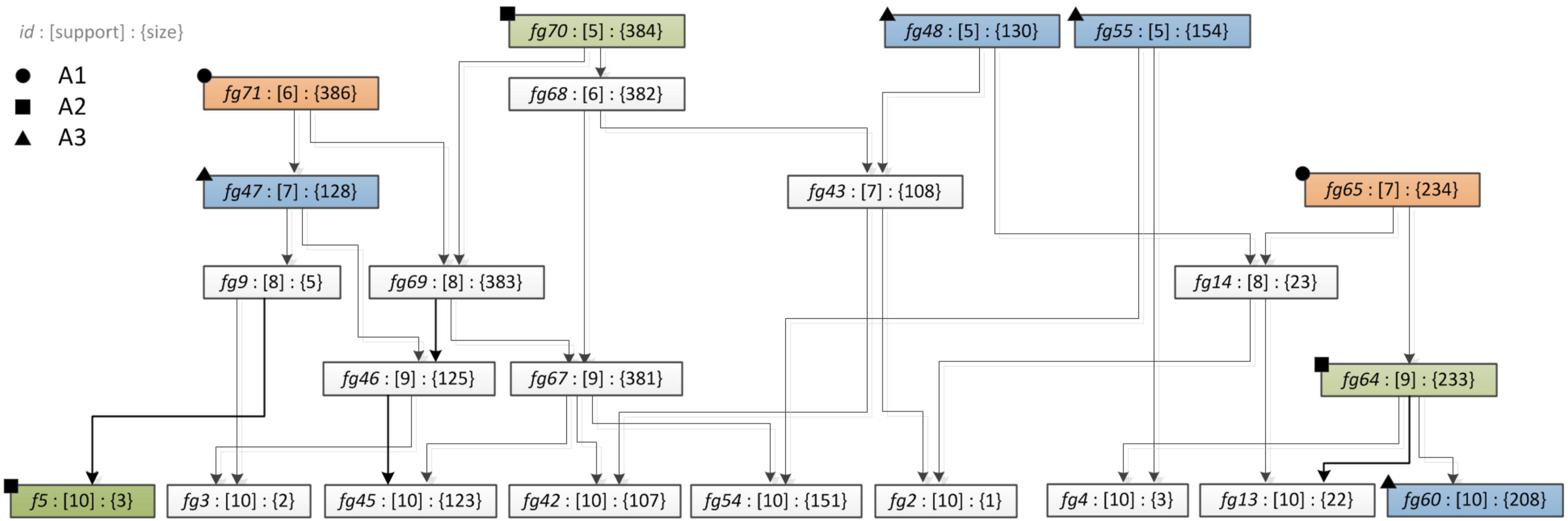

id : [support] : {size}
fg70 : : {384}
fg48 : [5 {130}
fg55 : : {154}
fg71 : [6] : {386}
fg68 : : {382}
fg47 : : {128}
fg43 : : {108}
fg65 : [7] : {234}
fg9 : : {5}
fg69 : : {383}
fg14 : [8] : {23}
fg46 : [{125}
fg67 : : {381}
fg64 : [9] : {233}
f5 : : {3}
fg3 : : {2}
fg45 : : {123}
fg42 : [: {107}
fg54 :] : {151}
fg2 : [10] : {1}
fg4 : [10] : {3}
fg13 : [10] : {22}
fg60 : [10] : {208}

> Result Extraction

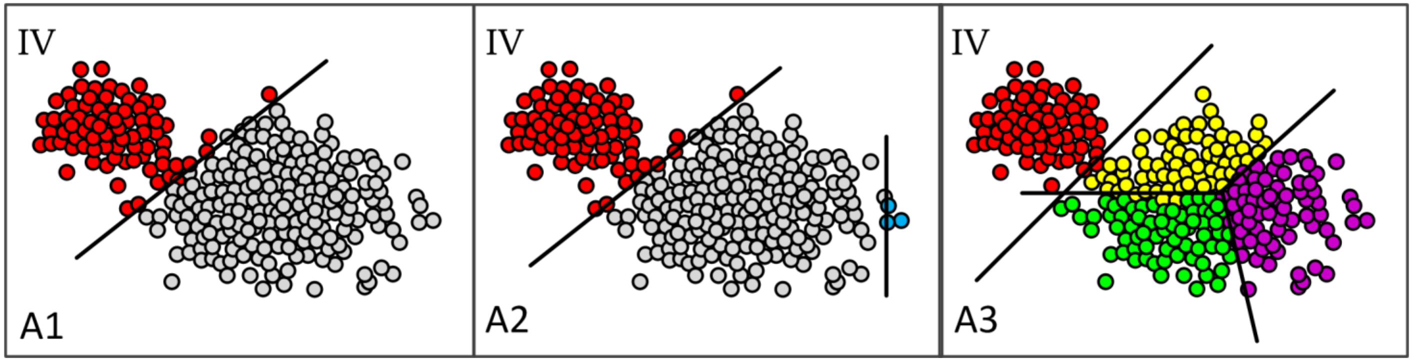


- number of alternatives is predictable
- less or even the number of root nodes

- additional alternatives via root removal & recursion

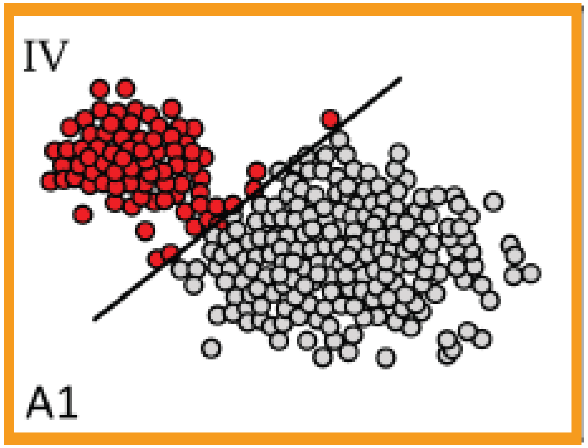


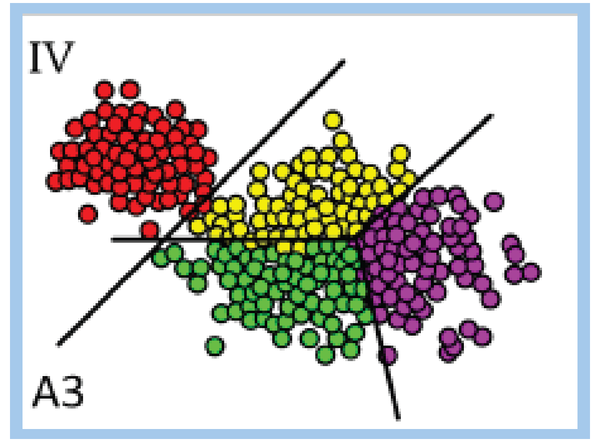
id : [support] : {size}
A1
A2
A3
fg70 : [5] : {384}
fg48 : [5] : {130}
fg55 : [5] : {154}
fg71 : [6] : {386}
fg68 : [6] : {382}
fg47 : [7] : {128}
fg43 : [7] : {108}
fg65 : [7] : {234}
fg9 : [8] : {5}
fg69 : [8] : {383}
fg14 : [8] : {23}
fg46 : [9] : {125}
fg67 : [9] : {381}
fg64 : [9] : {233}
f5 : [10] : {3}
fg3 : [10] : {2}
fg45 : [10] : {123}
fg42 : [10] : {107}
fg54 : [10] : {151}
fg2 : [10] : {1}
fg4 : [10] : {3}
fg13 : [10] : {22}
fg60 : [10] : {208}



IV
A1
IV
A2
IV
A3

Similarity measurement via intersection





$\lvert fg_{71} \cap fg_{48} \rvert$	107
$\lvert fg_{71} \cap fg_{55} \rvert$	**151**
$\lvert fg_{71} \cap fg_{47} \rvert$	128
$\lvert fg_{71} \cap fg_{60} \rvert$	0
$\lvert fg_{65} \cap fg_{48} \rvert$	23
$\lvert fg_{65} \cap fg_{55} \rvert$	3
$\lvert fg_{65} \cap fg_{47} \rvert$	0
$\lvert fg_{65} \cap fg_{60} \rvert$	**208**

$\lvert fg_{48} \cap fg_{71} \rvert$	**107**
$\lvert fg_{48} \cap fg_{65} \rvert$	23
$\lvert fg_{55} \cap fg_{71} \rvert$	**151**
$\lvert fg_{55} \cap fg_{65} \rvert$	3
$\lvert fg_{47} \cap fg_{71} \rvert$	**128**
$\lvert fg_{47} \cap fg_{65} \rvert$	0
$\lvert fg_{60} \cap fg_{71} \rvert$	0
$\lvert fg_{60} \cap fg_{65} \rvert$	**208**

(105+208)/620 = 0,58

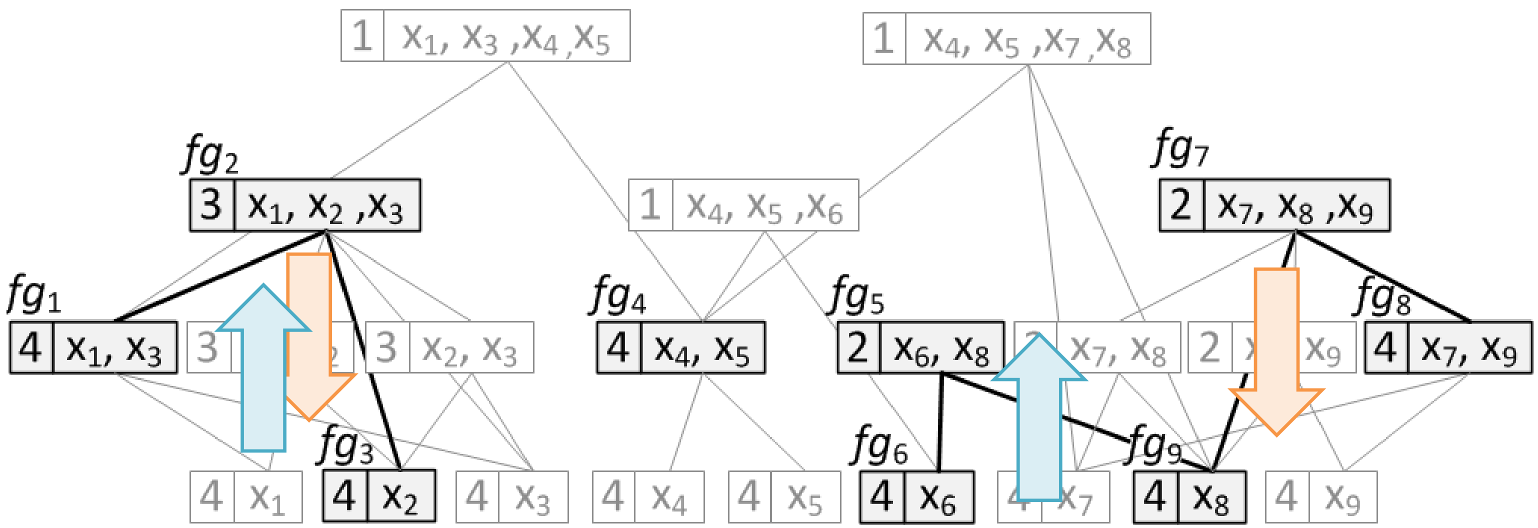
Future Work

Exploration so far

- top-down traversal from roots → alternatives through splitting clusters

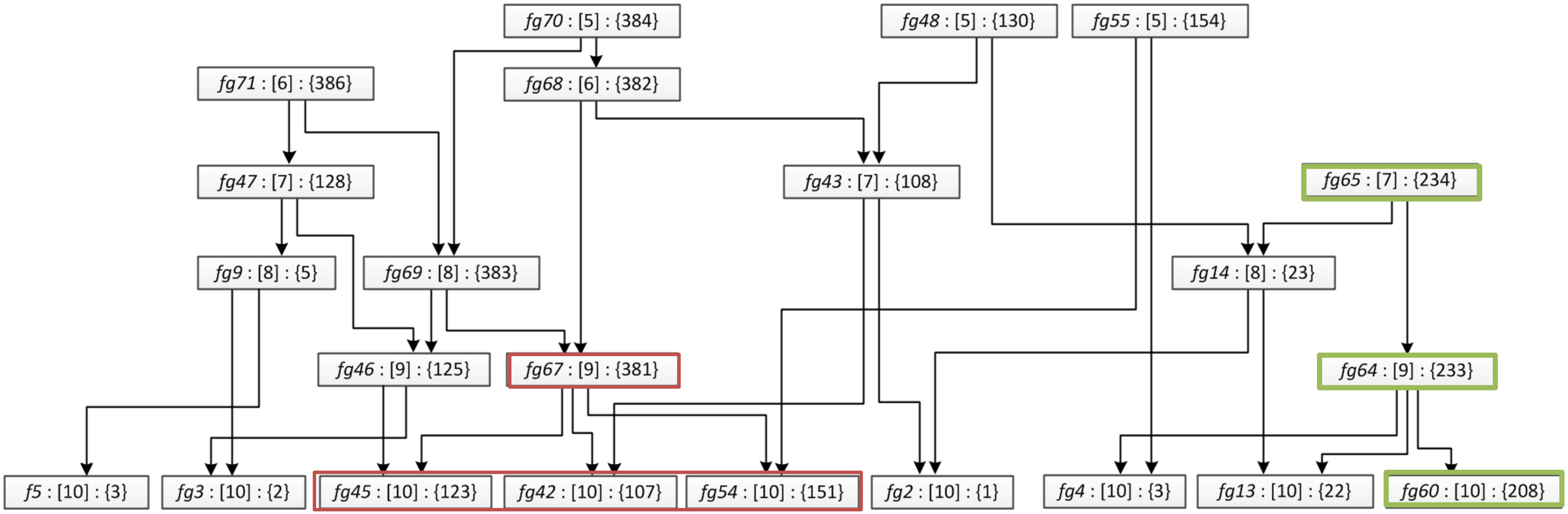
Other possibilities

- bottom-up traversal from leaves →alternatives through merging clusters
- starting level between root and leaves?
- combine bottom-up and top-down



Further Graph Utilization

- examine branching for dissimilarity of alternatives
- for feedback on ensemble diversity/homogenity
- relation between ensemble and alternatives



> Conclusion

- frequent groupings combine alternative and ensemble-clustering
- offer alternative and robust results
- automated recombination and its problems
- user-centered exploration

Questions?