On the Uniform Control of the Vapnik-Chervonenkis Dimension in Subgroup Discovery using Formal Concept Analysis

Georg Schollmeyer

30.03.2022

Subgroup Discovery

The basic task of subgroup discovery can be stated as:

"In subgroup discovery, we assume we are given a so-called population of individuals (objects, customer,...) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically "most interesting" i.e. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest." [Wrobel, 2001]

Subgroup Discovery

- Piatetsky-Shapiro quality function q_{PS} is a quantity for measuring the statistical interestingness of a subgroup.
- It can be shown that

$$q_{PS} = C \cdot \hat{D}_{\mathcal{S},n}.$$

- Here, C is a fixed constant and $\hat{D}_{\mathcal{S},n}$ is a Kolmogorov-Smirnov type supremum statistic.
- Therefore, we would like to take a statistical look at Subgroup Discovery.
- We use insights from Vapnik-Chervonenkis theory (statistical learning theory).

Overview i

- Given family $\mathcal{S}$ of sets (subgroups) over some ground space (population) $\mathscr{G}$.
- Given two probability laws P and P' over $\mathscr{G}$ that represent the distribution of a binary target attribute of interest within $\mathscr{G}$, e.g., gender:
  $P(A)\ldots$ proportion of male persons in subgroup $A \in \mathcal{S}$ (in relation to all male persons),
  $P'(A)\ldots$ proportion of female persons in subgroup $A \in \mathcal{S}$ (in relation to all female persons).

Overview ii

- Quantity of interest:

$$D_{\mathcal{S}} := \sup_{A \in \mathcal{S}} |P(A) - P'(A)| :$$

  Maximal (absolute) difference in proportions of the attribute of interest in some subgroup A.

- Or the argmax: That subgroup for which the difference in proportions is maximal.
- Problem: Inference: We have only samples of the entire population $\mathscr{G}$ (of size n, both for P and P') and replace the laws P and P' with its empirical analogues ν and ν', respectively.

Overview iii

- If the family $\mathcal{S}$ is too large (or too complex), then the estimator

$$\hat{D}_{\mathcal{S},n} := \sup_{A \in \mathcal{S}} |\nu(A) - \nu'(A)|$$

  is a very poor estimator of

$$D_{\mathcal{S}} = \sup_{A \in \mathcal{S}} |P(A) - P'(A)|$$

  (the same holds for the argmax).
- Idea: Regularization: a) Reduce $\mathcal{S}$, i.e., make the ('effective') size of $\mathcal{S}$ smaller by looking only at a subfamily $\mathcal{T} \subseteq \mathcal{S}$.

Overview iv

- Or (here): b) Restructure $\mathcal{S}$ in a way that reduces its complexity, i.e., replacing $\mathcal{S}$ by $\mathcal{U}$ which is similar to $\mathcal{S}$ w.r.t. the substance matter problem at hand, but statistically more tamely than $\mathcal{S}$.
- Tools for guiding this in a statistically sophisticated manner: Statistical learning theory.
  Important quantities:
  - Vapnik-Chervonenkis dimension (VC dimension): h
  - Growth function: $m^{\mathcal{S}}$.
- Both quantities control both the effective size, as well as the complexity of $\mathcal{S}$.

Overview v

- There are interrelations between h and $m^{\mathcal{S}}$, as well as between the size and the complexity of $\mathcal{S}$.
- But the later are somehow not very clear w.r.t. the statistical behaviour of the statistic $\hat{D}_{\mathcal{S},n}$.
- In the sequel, we would like to regularize $\hat{D}_{\mathcal{S},n}$
  a) by making $\mathcal{S}$ smaller, i.e., working with a subset $\mathcal{T} \subseteq \mathcal{S}$ or
  b) by making $\mathcal{S}$ less complex in terms of the VC dimension by 'restructuring' $\mathcal{S}$ to $\mathcal{U}$ with $h(\mathcal{U}) \leq h(\mathcal{S})$ (note that generally $\mathcal{U} \not\subseteq \mathcal{S}$).
- Then we use the regularized statistic

$$\hat{D}_{\mathcal{T},n} = \sup_{A \in \mathcal{T}} |\nu(A) - \nu'(A)| \quad \text{or} \quad \hat{D}_{\mathcal{U},n} = \sup_{A \in \mathcal{U}} |\nu(A) - \nu'(A)|,$$

respectively, e.g. for a statistical test of equality of P and P'.

Relation to Formal Concept Analysis

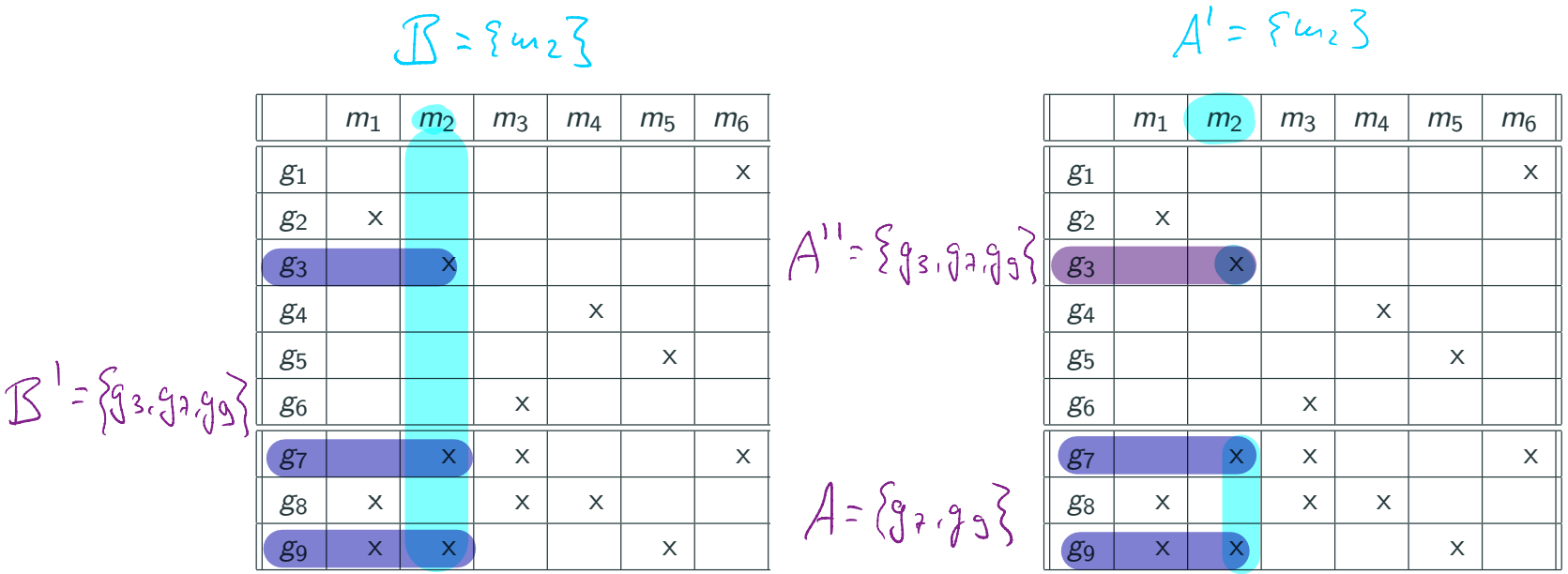
Starting point: Formal context $\mathbb{K} = (G, M, I)$ ('crosstable') with

- $G \subseteq \mathscr{G}$ ... set of objects (here: surveyed statistical units, e.g., respondents in a social survey).
- M ... set of (binary) attributes (here, covariates).
- $I \subseteq G \times M$ ... binary relation with $(g, m) \in I$ if person g has attribute m.

	x			y
	income $\leq$ 1000	occupied	$\cdots$	sex=male
person 1	$\times$	$\times$		0
person 2		$\times$		1
person 3				0
$\vdots$				1
person m		$\times$		0

- Subgroups in Subgroup Discovery are usually described by attribute descriptions, i.e. by attribute sets: For $B \in M$ define $B' := \{g \in G \mid \forall m \in B : gIm\}$.
- In Formal Concept Analysis (FCA): Family of all subgroups can be also seen as all subgroups generated by arbitrary sets of objects: For arbitrary $A \subseteq G$: $A \mapsto A' := \{m \in M \mid \forall g \in A : gIm\}$ $A' \mapsto A'' := \{g \in G \mid \forall m \in A' : gIm\}$.
- The map $A \mapsto A' \mapsto A''$ is a closure operator which is studied in FCA. All images of the operator $''$ are called closed sets, hulls or **extents** and are exactly the subgroups in Subgroup Discovery.
- The images of the operator $'$ are called closed item sets or **intents**.

Important here: FCA allows for looking at subgroups as generated by object sets, not only as generated by attribute sets.

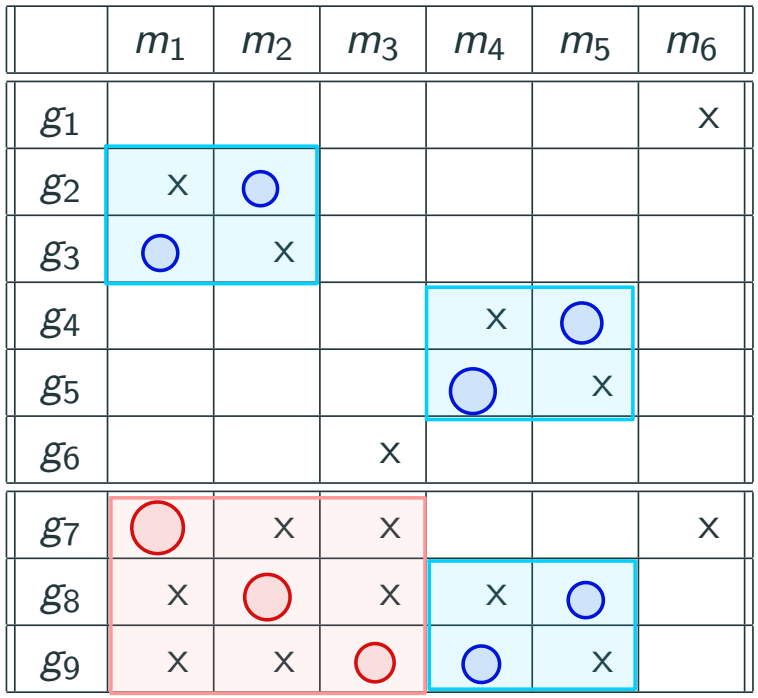


	m_1	m_2	m_3	m_4	m_5	m_6
g_1						x
g_2	x					
g_3		x				
g_4				x		
g_5					x	
g_6			x			
g_7		x	x			x
g_8	x		x	x		
g_9	x	x			x	

	m_1	m_2	m_3	m_4	m_5	m_6
g_1						x
g_2	x					
g_3		x				
g_4				x		
g_5					x	
g_6			x			
g_7		x	x			x
g_8	x		x	x		
g_9	x	x			x	

Vapnik-Chervonenkis theory in Formal Concept Analysis

In FCA, the VC dimension of the space $\mathcal{S} = \{B' \mid B \subseteq M\}$ is given by the largest size of a contranominal scale:



VC dimension = largest size of a contranominal scale (here) = 3

Regularization: A): Reducing $\mathcal{S}$ to a subfamily $\mathcal{T} \subseteq \mathcal{S}$:

don't use, e.g. $\{m_1, m_2\}$

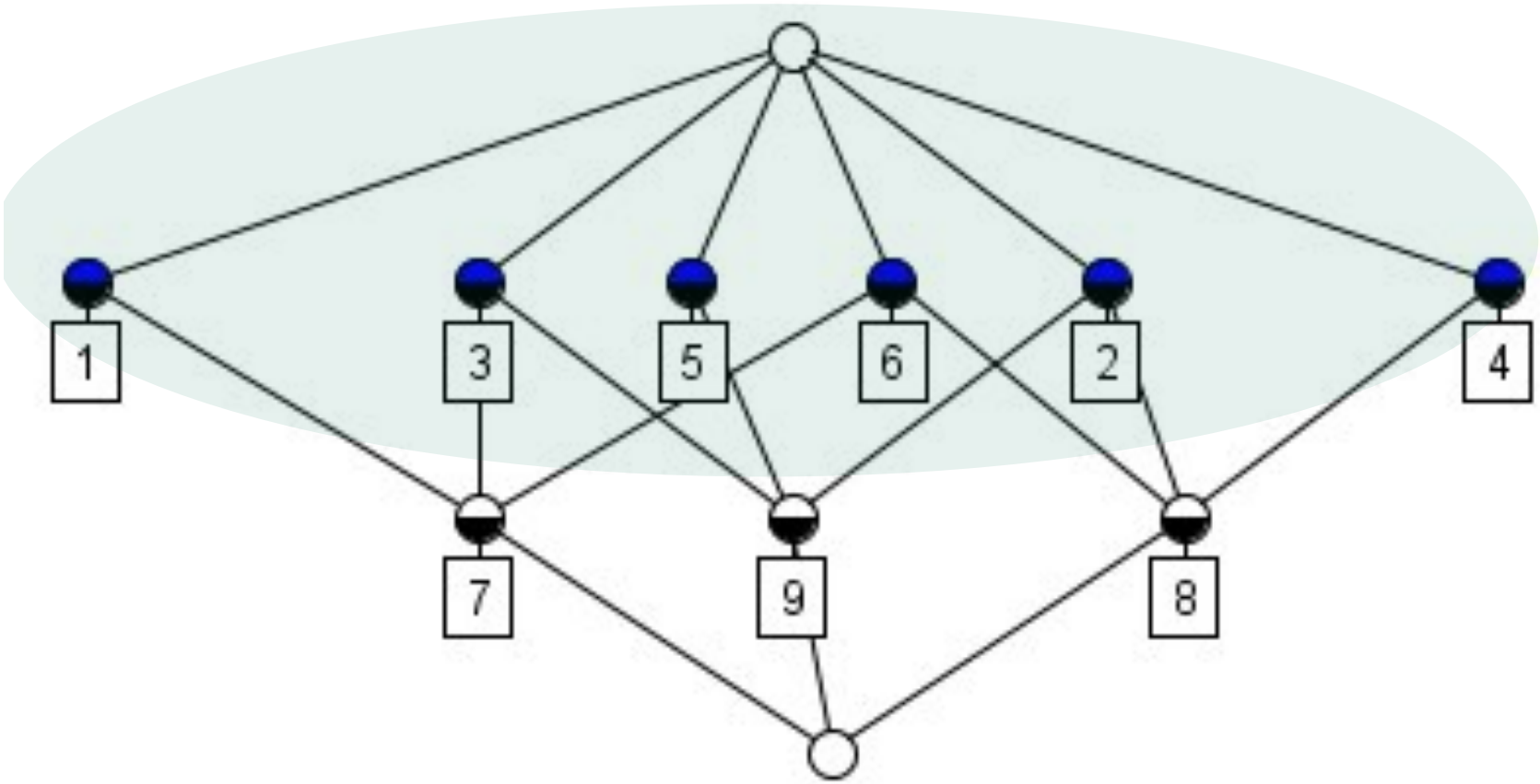
	m_1	m_2	m_3	m_4	m_5	m_6
g_1						×
g_2	×	○				
g_3	○	×				
g_4				×	○	
g_5				○	×	
g_6			×			
g_7	○	×	×			×
g_8	×	○	×	×	○	
g_9	×	×	○	○	×	

Subcontext with VC dimension k!

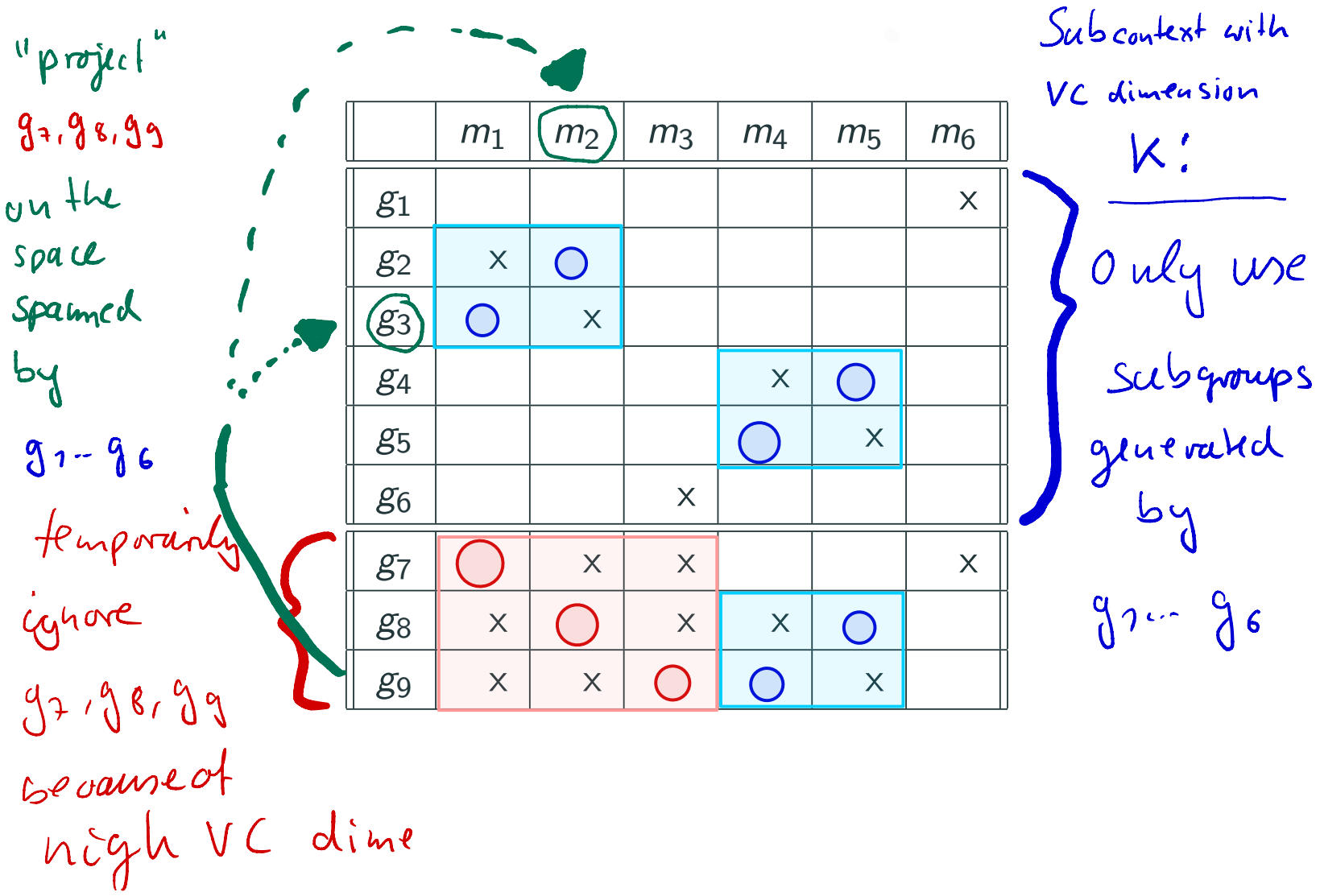
Only use subgroups generated by $g_1 \dots g_6$

temporarily ignore g_7, g_8, g_9 because of high VC dimension

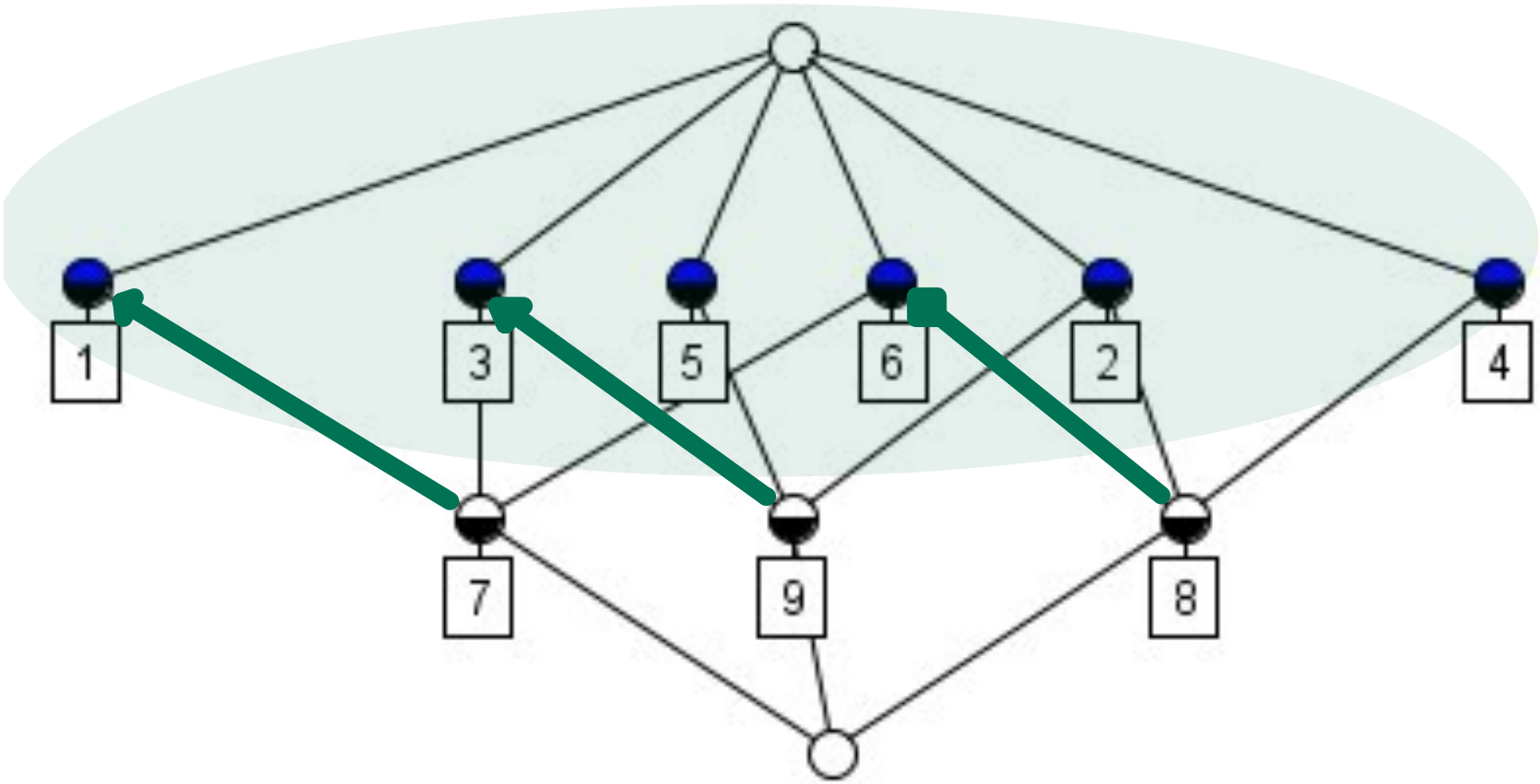
Regularization: A): Reducing $\mathcal{S}$ to a subfamily $\mathcal{T} \subseteq \mathcal{S}$:



Regularization: B): Restructuring $\mathcal{S}$ to $\mathcal{U}$ with $h(\mathcal{U}) \leq h(\mathcal{S})$:



Regularization: B): Restructuring $\mathcal{S}$ to $\mathcal{U}$ with $h(\mathcal{U}) \leq h(\mathcal{S})$:



Theorem (Restructuring $\mathcal{S}$)

i) If a domain $A \subseteq G$ does not suffer from a high VC dimension in the sense that no $g \in A \subseteq G$ belongs to a large contranominal scale, then

$$\mathcal{S}_A = \mathcal{U}_A, \text{ i.e.:}$$

$$\{B \cap A \mid B \in \mathcal{S}\} = \{B \cap A \mid B \in \mathcal{U}\}.$$

ii) In general, it does not hold that $\mathcal{U} \subseteq \mathcal{S}$.

iv) The VC dimension of $\mathcal{U}$ is at most K. (where K is the VC dimension of the subcontext).

iii) Therefore we have the generalization inequality (under $H_0 : P = P'$):

$$P^{\otimes n}\left(\hat{D}_{\mathcal{U},n} \geq \varepsilon\right) \leq 6 \cdot \left[(2n)^K + 1\right] e^{-\frac{n\varepsilon^2}{4}}.$$

Application i

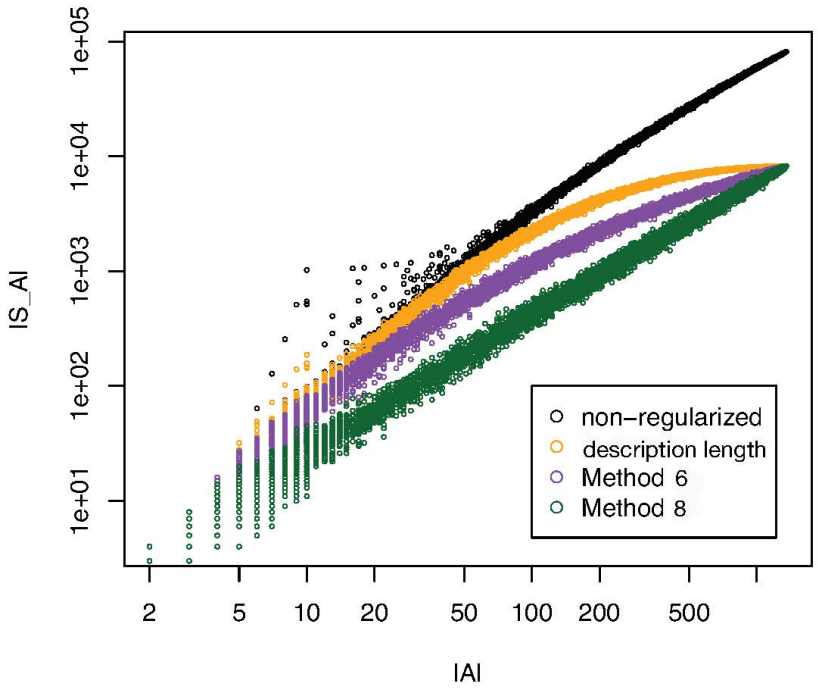
- German General Social Survey (GGSS): ISSP 2017 module on social networks and social resources.
- Covariates: 10 questions asking if the respondents know several people who have an occupation from a list of 10 occupational groups (e.g., bus/lorry driver). (4 answer categories)
- Target variable: Answer to the question 'At these occasions [going out to eat or drink with three or more friends or acquaintances who are not family members], how often do you make new friends or acquaintances?' (dichotomized here as 1 if 'often' or 'very often' and 0 otherwise)

Application ii

- Test:

  $H_0 : P = P'$: Making new friends is not related to the people one knows vs

  $H_1 : P \neq P'$: There is a subgroup (described by which people one knows) for which making new friends is different compared to the whole population (or compared to persons who know/do not know certain other people).



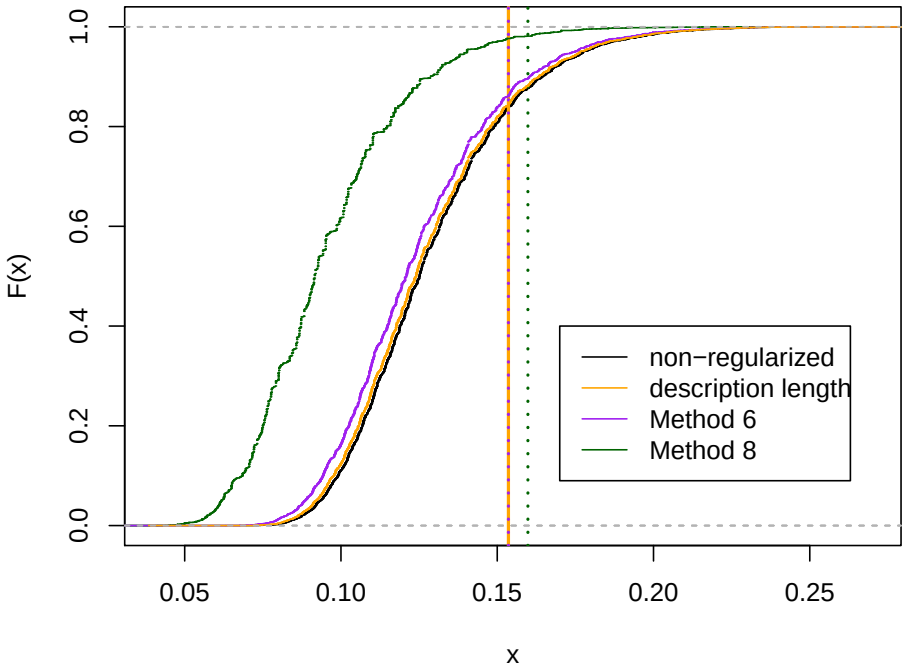


Figure 1: Simulation-based estimate of the growth function (left) and the distribution of the statistic D under H_0 based on a permutation scheme (right). The vertical lines represent the value of the statistic for the actually observed data.

References

S. Wrobel. Inductive logic programming for knowledge discovery in databases. In *Relational data mining*, pages 74–101. Springer, 2001.