Oct. 22, 2020
第41回IBISML

Inter-University Research Institute Corporation /
Research Organization of Information and Systems
National Institute of Informatics

PRESTO
SAKIGAKE

統計的に有意な相互作用探索

杉山 麿人（国立情報学研究所）

Significant Pattern Mining

- Detect patterns (feature combinations) that are statistically significantly enriched in a class of a dataset

Input:

	F1	F2	F3	F4	F5	...	Class
ID1	0	1	1	1	0	...	0
ID2	1	1	0	1	1	...	1
ID3	1	1	0	0	1	...	1
ID4	0	0	1	0	1	...	0
⋮			⋮				⋮

(Features F1, F2, ... form $\boldsymbol{x}$; Class forms y.)

Output:

{F1}, {F3}, {F2, F5}, {F2, F5, F6}, ...

Significant Pattern Mining

- Detect patterns (feature combinations) that are statistically significantly enriched in a class of a dataset

Input:

	x						y
	F1	F2	F3	F4	F5	...	Class
ID1	0	1	1	1	0	...	0
ID2	1	1	0	1	1	...	1
ID3	1	1	0	0	1	...	1
ID4	0	0	1	0	1	...	0
⋮			⋮				⋮

Output:

{F1}, {F3}, {F2, F5}, {F2, F5, F6}, ...

Significant Pattern Mining

- Detect patterns (feature combinations) that are statistically significantly enriched in a class of a dataset

Input:

	F1	F2	F3	F4	F5	...	Class
ID1	0	1	1	1	0	...	0
ID2	1	1	0	1	1	...	1
ID3	1	1	0	0	1	...	1
ID4	0	0	1	0	1	...	0
⋮			⋮				⋮

(x over F1–…, y over Class)

Output:

{F1}, {F3},
{F2, F5},
{F2, F5, F6}, ...

Significant Pattern Mining

- Detect patterns (feature combinations) that are statistically significantly enriched in a class of a dataset

Input:

	F1	F2	F3	F4	F5	...	Class
ID1	0	1	1	1	0	...	0
ID2	1	1	0	1	1	...	1
ID3	1	1	0	0	1	...	1
ID4	0	0	1	0	1	...	0
⋮			⋮				⋮

$\underbrace{\text{F1 ... }}_{}$ columns F1–… form $\boldsymbol{x}$; Class is y

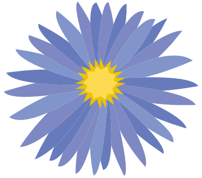
→ **Output:**

{F1}, {F3},
{F2, F5},
{F2, F5, F6}, ...

Example: Itemset Mining

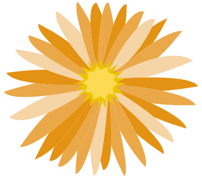
SNPs (items)

Case



ID 1: 0 0 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0
ID 2: 1 1 0 0 1 0 1 1 1 1 0 0 0 1 0 1 0 1 0 0
ID 3: 1 0 1 0 0 0 1 1 1 0 1 0 1 1 0 0 0 0 0 1
ID 4: 1 1 0 1 1 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1

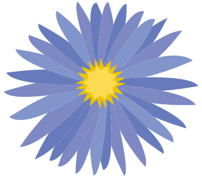
Control

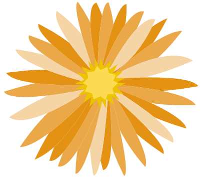


ID 5: 0 0 1 1 0 0 0 1 1 0 0 0 1 1 1 1 1 0 0 0
ID 6: 0 1 0 1 1 0 1 1 0 0 0 0 1 1 0 0 1 0 1 0
ID 7: 1 0 1 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 0
ID 8: 1 1 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0 1

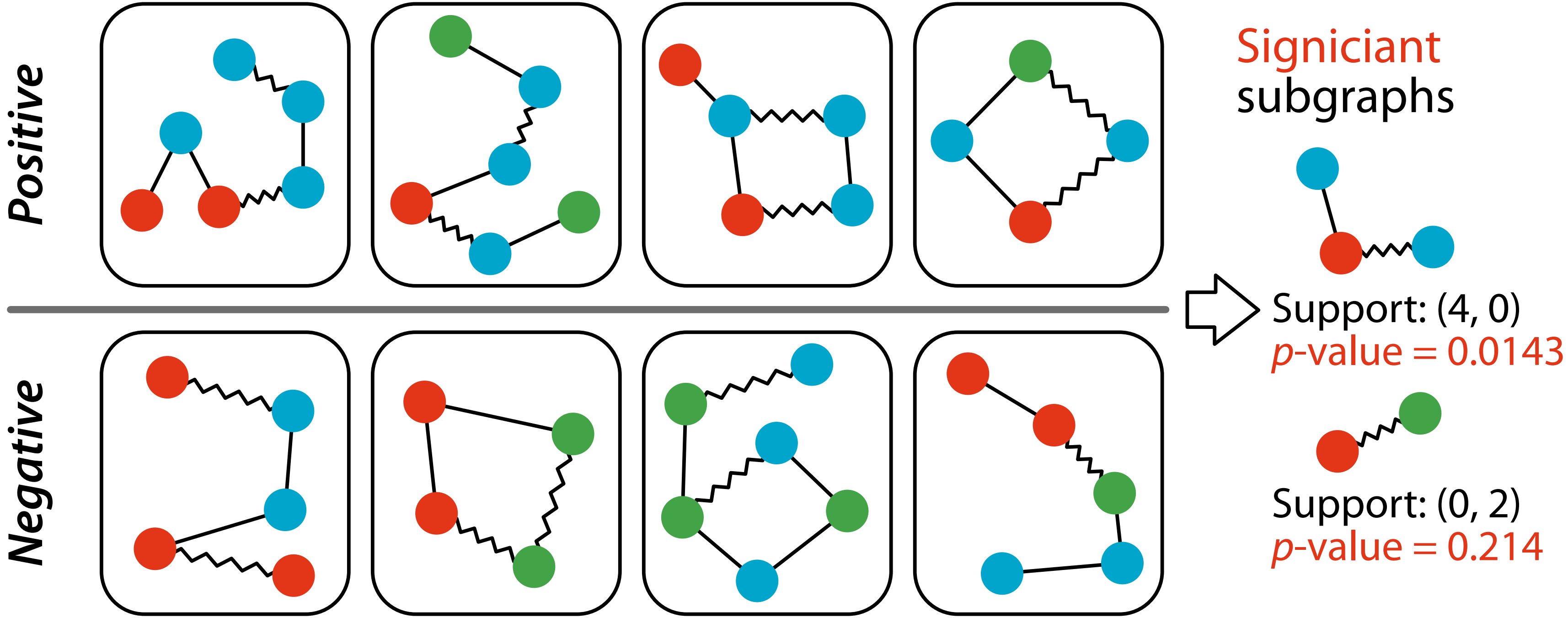
Example: Itemset Mining

		SNPs (items)	Signal
Case	ID 1:	00110011100111001110	1
	ID 2:	11001011110001010100	1
	ID 3:	10100011101011000001	1
	ID 4:	11011011111111010011	1
Control	ID 5:	00110001100011111000	0
	ID 6:	01011011000011001010	0
	ID 7:	10110010100000101000	0
	ID 8:	11001001010100010101	0

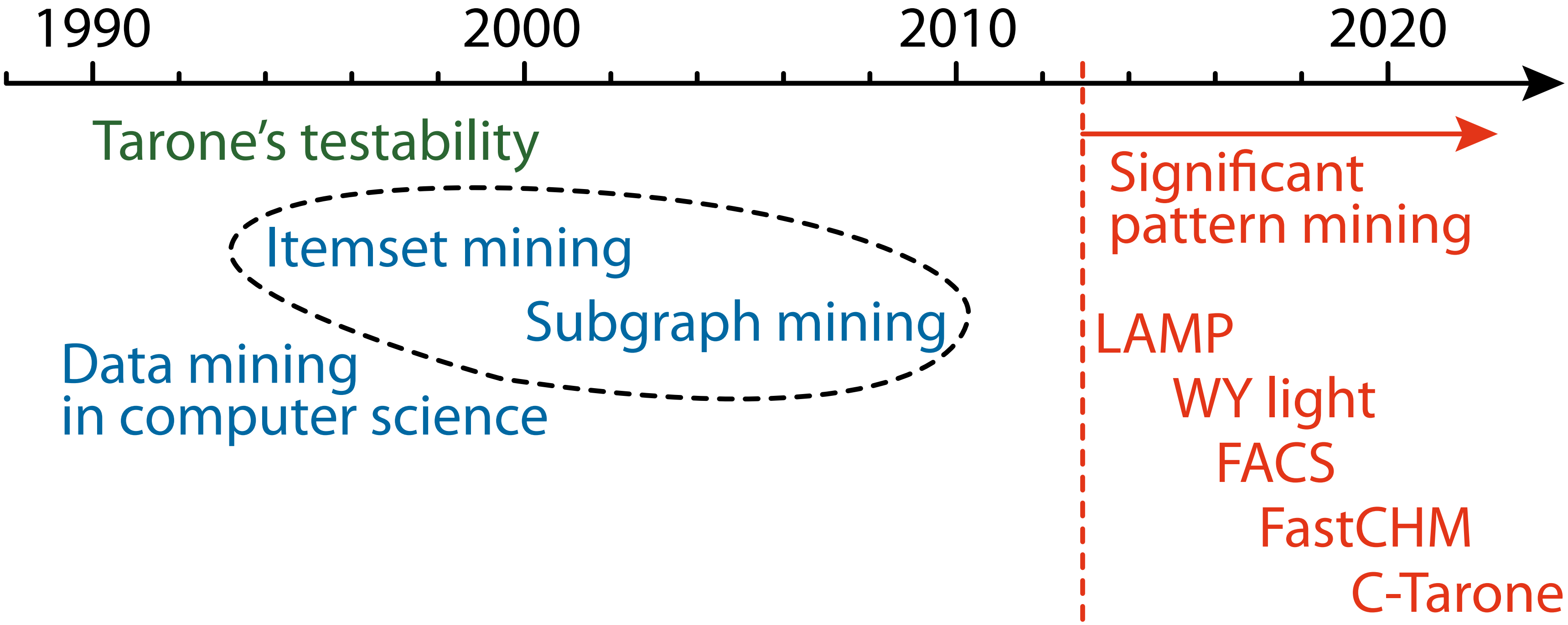




Example: Subgraph Mining



Timeline


1990
2000
2010
2020
Tarone's testability
Itemset mining
Subgraph mining
Data mining
in computer science
Significant
pattern mining
LAMP
WY light
FACS
FastCHM
C-Tarone

Recent Advances

- Webb, G.I., Petitjean, F.: **A Multiple Test Correction for Streams and Cascades of Statistical Hypothesis Tests**, KDD2016
- Pellegrina, L., Vandin, F.: **Efficient Mining of the Most Significant Patterns with Permutation Testing**, KDD2018
- Pellegrina, L., Riondato, M., Vandin, F.: **SPuManTE: Significant Pattern Mining with Unconditional Testing**, KDD2019
- Tran, T.Q., Fukuchi, K., Akimoto, Y., Sakuma, J.: **Statistically Significant Pattern Mining with Ordinal Utility**, KDD2020

Libraries

- CASMAP
  - Llinares-Lopez, et al.: **CASMAP: Detection of statistically significant combinations of SNPs in association mapping**, Bioinformatics (2019)
- MP-LAMP (for parallel computation)
  - Yoshizoe, K., Terada, A., Tsuda, K.: **MP-LAMP: parallel detection of statistically significant multi-loci markers on cloud platforms**, Bioinformatics (2018)

Key Challenges:

1. How to assess the significance for a multiplicative interaction of variables?

2. How to perform multiple testing correction?
   - How to control the FWER (family-wise error rate), the probability to detect one or more false positives?

3. How to manage combinatorial explosion (2^d for d variables) of the candidate space?

Problem Formulation

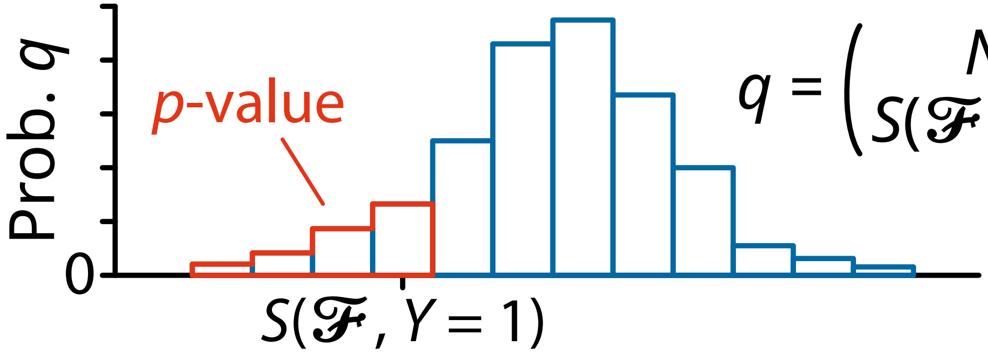
- Define $X_{\mathcal{F}}$ as binary random variable of joint occurrence for a feature combination $\mathcal{F} = \{F_i\}_{i \in I}, I \subseteq \{1, \dots, d\}$
  - $X_{\mathcal{F}} = 1$ if $\mathcal{F}$ "occurs", $X_{\mathcal{F}} = 0$ otherwise
- Let Y be an output binary variable

Problem Formulation

- Define $X_{\mathcal{F}}$ as binary random variable of joint occurrence for a feature combination $\mathcal{F} = \{F_i\}_{i \in I}, I \subseteq \{1, \ldots, d\}$
  - $X_{\mathcal{F}} = 1$ if $\mathcal{F}$ "occurs", $X_{\mathcal{F}} = 0$ otherwise
- Let Y be an output binary variable
- **Task:** Test the null hypothesis $X_{\mathcal{F}} \perp\!\!\!\perp Y$ for *all* $\mathcal{F} \in 2^V$
  - Testing statistical independence between $X_{\mathcal{F}}$ and Y

Fisher's Exact Test

	$X_{\mathcal{F}} = 1$	$X_{\mathcal{F}} = 0$	Total
$Y = 1$	$S(\mathcal{F}, Y = 1)$	$N_1 - S(\mathcal{F}, Y = 1)$	N_1
$Y = 0$	$S(\mathcal{F}, Y = 0)$	$N_0 - S(\mathcal{F}, Y = 0)$	N_0
Total	$S(\mathcal{F})$	$N - S(\mathcal{F})$	N



$$q = \binom{N_1}{S(\mathcal{F}, Y{=}1)}\binom{N_0}{S(\mathcal{F}, Y{=}0)} \Big/ \binom{N}{S(\mathcal{F})}$$

Multiple Testing Correction

- In each test, [probability of having a false positive] $\leq \alpha$
- If we repeat m tests, αm patterns can be false positives
  - Too many if m is large! For example in itemset mining:
    - For 100000 items, #patterns $= 2^{100000}$
    - Set significance level $\alpha = 0.01$
    - Number of false positives: $0.01 \cdot 2^{100000} = 10^{30101}$
- The FWER should be controlled
  - Probability at least one $\mathcal{F}$ is false positive

Controlling the FWER

- FWER $= \Pr(\mathrm{FP} > 0)$
  - FP: Number of false positives
- **Objective**: Maximize FWER(δ) subject to FWER(δ) $\leq \alpha$
  - FWER(δ): FWER at corrected significance level δ
    - Cannot be evaluated in closed form (simple but not easy!)
  - Bonferroni correction is popular: $\delta^*_{\text{Bon}} = \alpha/m$

Tarone's Testability Trick

- We use Tarone's testability trick, which requires the minimum achievable p-value $\psi(\mathcal{F})$ for $\mathcal{F}$

$$\psi(\mathcal{F}) = \binom{N_1}{S(\mathcal{F})} \Big/ \binom{N}{S(\mathcal{F})} \quad \text{in Fisher's exact test}$$

Tarone's Testability Trick

$$\mathcal{F}_1\,,\ \mathcal{F}_2\,,\ \mathcal{F}_3\ ,\ldots,\ \mathcal{F}_{m-1}\,,\ \mathcal{F}_m\,,\ \mathcal{F}_{m+1}\ ,\ldots,\ \mathcal{F}_{2^d} \qquad \big(\psi(\mathcal{F}_i) \leq \psi(\mathcal{F}_{i+1})\big)$$

Tarone's Testability Trick

$$\boxed{m\,\psi(\mathcal{F}_m) < \alpha \quad \text{and} \quad (m+1)\psi(\mathcal{F}_{m+1}) \geq \alpha}$$

$$\mathcal{F}_1,\ \mathcal{F}_2,\ \mathcal{F}_3\ ,\ldots,\ \mathcal{F}_{m-1},\ \mathcal{F}_m,\ \big|\ \mathcal{F}_{m+1}\ ,\ldots,\ \mathcal{F}_{2d} \qquad \big(\psi(\mathcal{F}_i) \leq \psi(\mathcal{F}_{i+1})\big)$$

Tarone's Testability Trick

$$\boxed{m\,\psi(\mathcal{F}_m) < \alpha \quad \text{and} \quad (m+1)\psi(\mathcal{F}_{m+1}) \geq \alpha}$$

$$\underbrace{\mathcal{F}_1,\ \mathcal{F}_2,\ \mathcal{F}_3\ ,\ldots,\ \mathcal{F}_{m-1},\ \mathcal{F}_m,}_{\text{Testable combinations}}\ \Big|\ \underbrace{\mathcal{F}_{m+1}\ ,\ldots,\ \mathcal{F}_{2^d}}_{\text{Untestable combinations}} \qquad \big(\psi(\mathcal{F}_i) \leq \psi(\mathcal{F}_{i+1})\big)$$

Testable combinations

Untestable combinations → Prune without testing

↓

$\mathcal{F}_i$ is significant if: p-value$(\mathcal{F}_i) < \alpha\,/m$ — Correction factor

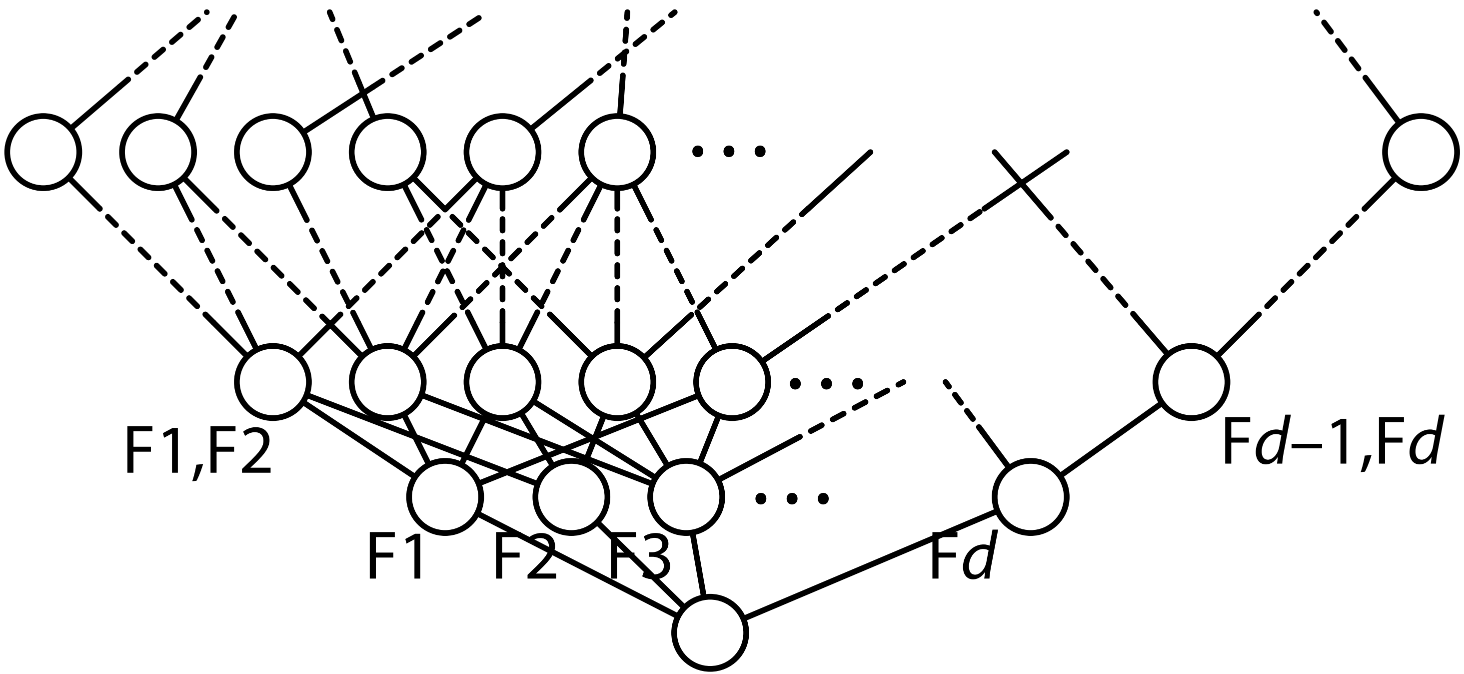
Tarone's Testability Trick with Apriori

- We use Tarone's testability trick, which requires the minimum achievable p-value $\psi(\mathcal{F})$ for $\mathcal{F}$

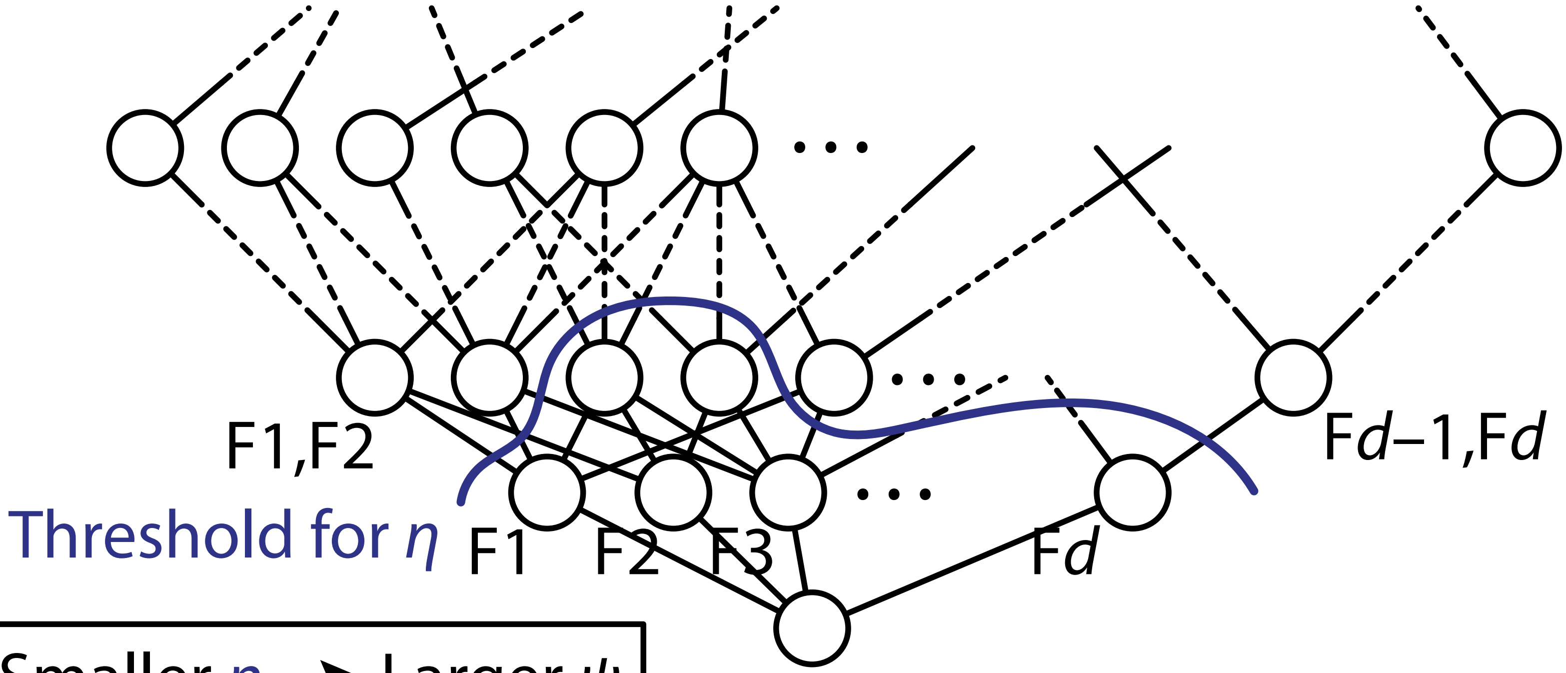
$$\psi(\mathcal{F}) = \binom{N_1}{S(\mathcal{F})} \Big/ \binom{N}{S(\mathcal{F})} \quad \text{in Fisher's exact test}$$

- This method is particularly effective if the relationship "Smaller $\eta(\mathcal{F}) \to$ Larger $\psi(\mathcal{F})$" holds
  - For each pattern $\mathcal{F}$,
    $S(\mathcal{F})$: Support (how many times $\mathcal{F}$ occurs in a dataset)
    $\eta(\mathcal{F}) = S(\mathcal{F})/N$: Frequency

Enumeration Based on Apriori

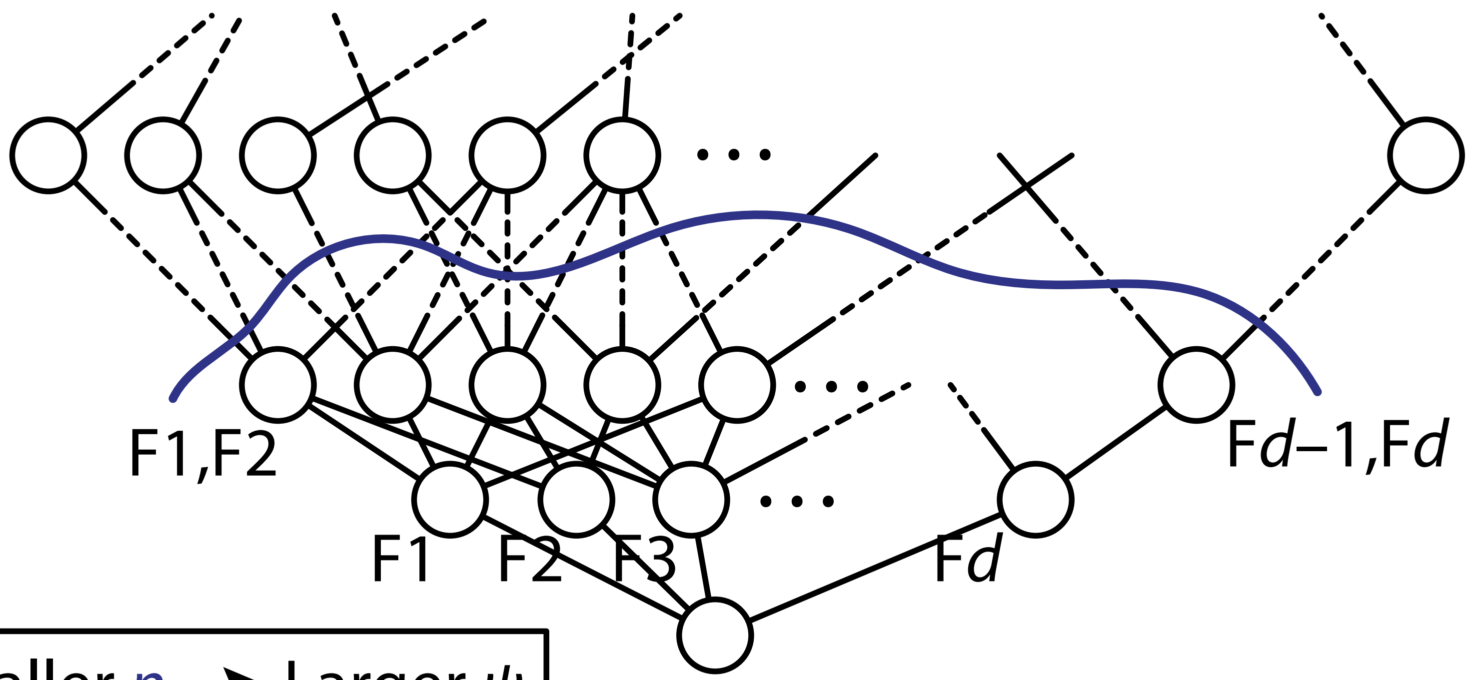


Enumeration Based on Apriori



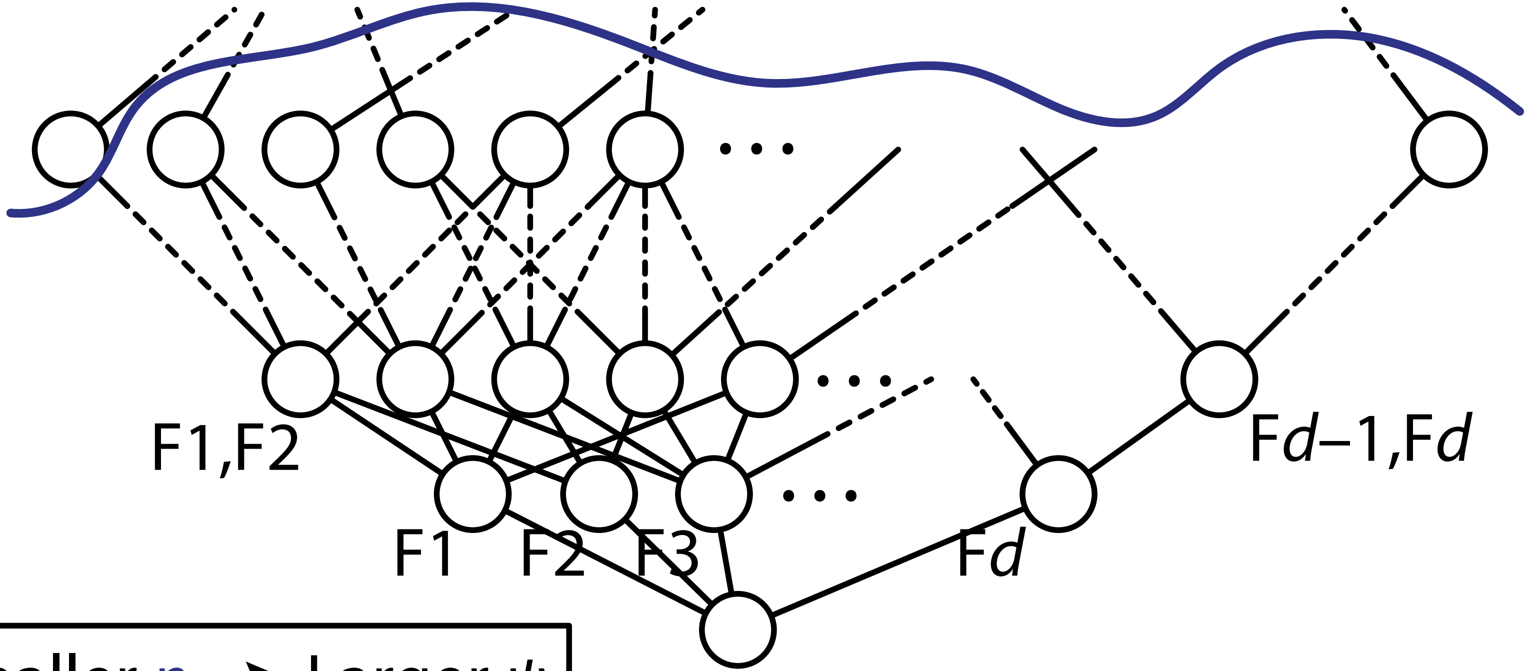
Smaller $\eta \longrightarrow$ Larger ψ

Enumeration Based on Apriori



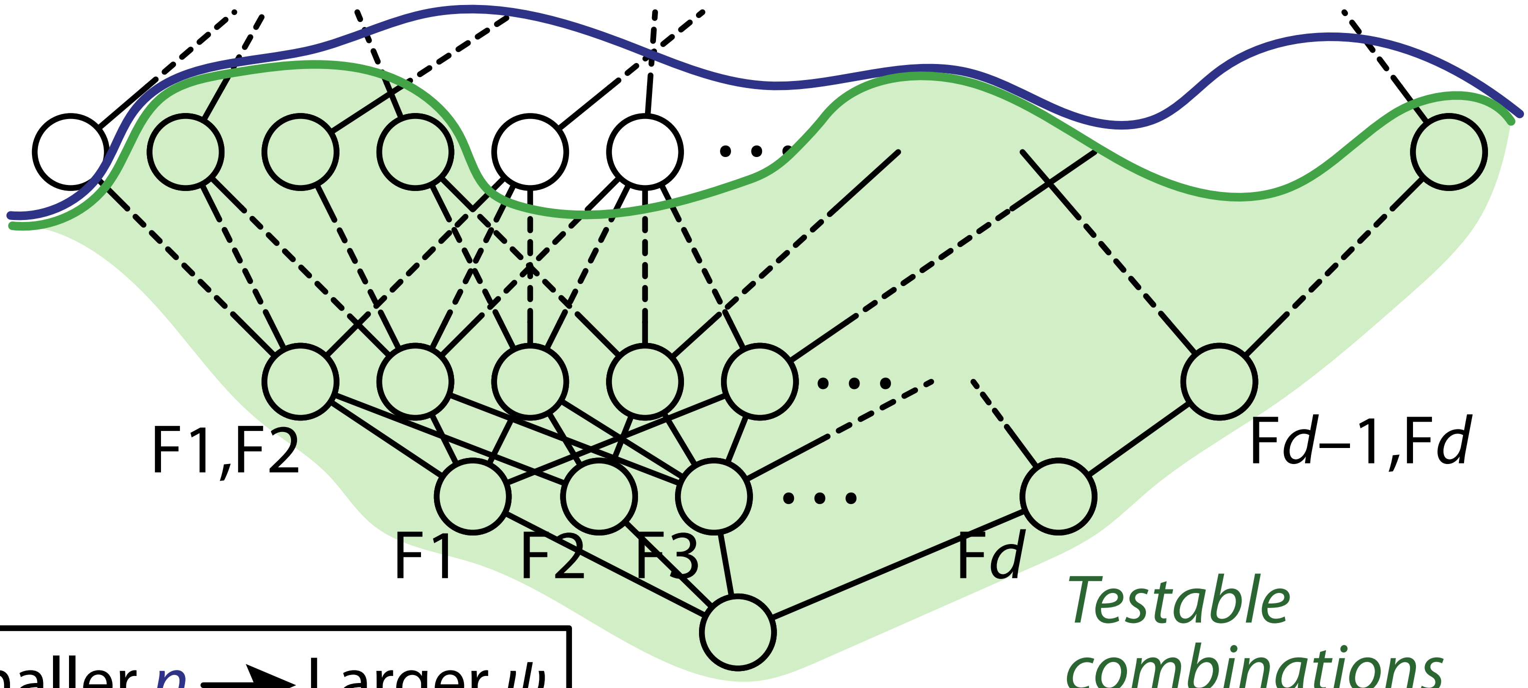
Smaller $\eta \longrightarrow$ Larger ψ

Enumeration Based on Apriori



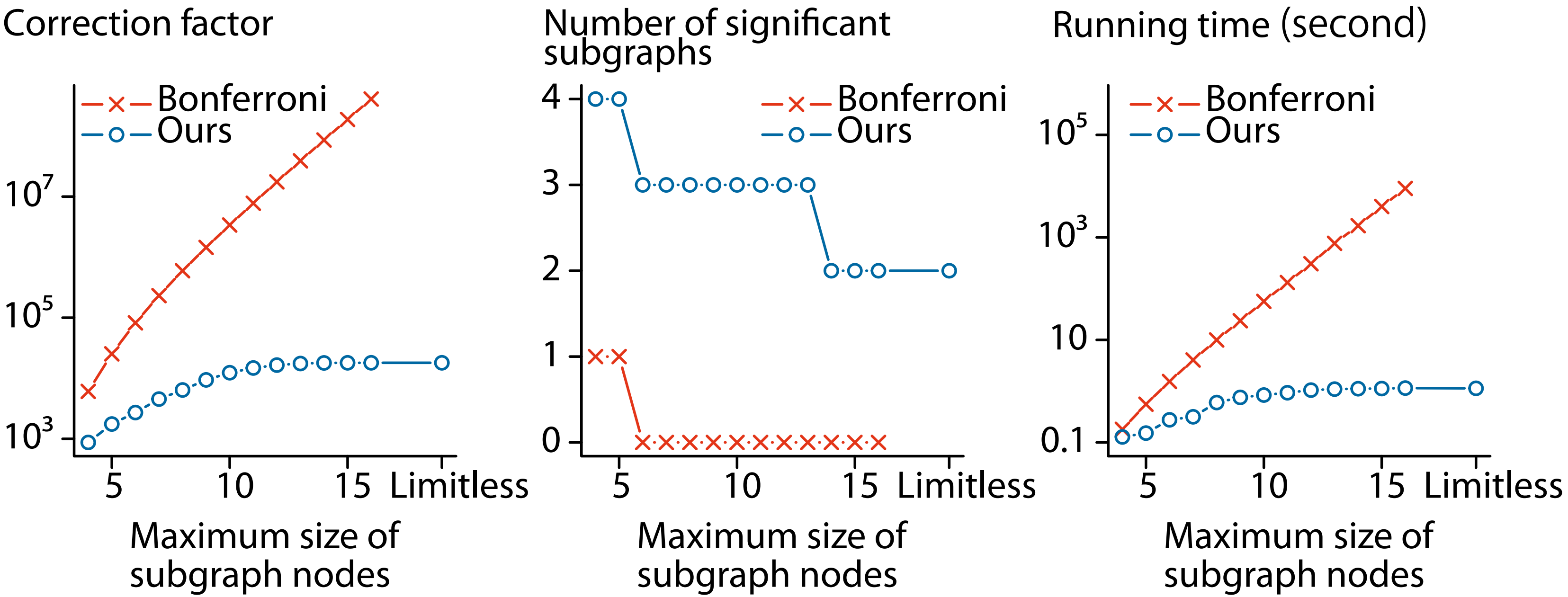
Smaller η ⟶ Larger ψ

Enumeration Based on Apriori



Smaller $\eta \longrightarrow$ Larger ψ

Power of Testability



The PTC (Predictive Toxicology Challenge) dataset with 601 chemical compounds

Summary

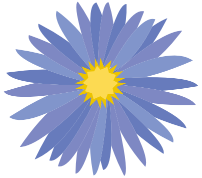
1. Formulate significance test for each pattern
   - Fisher's exact test is standard,
     while there are more possibilities

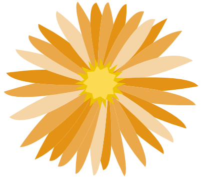
2. Enumerate testable patterns via
   Tarone's testability + Apriori (DFS)

3. Test each testable pattern

LAMP and WY light

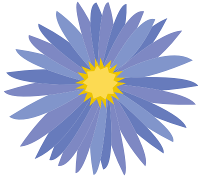
		SNPs (items)	Signal
Case	ID 1:	0 0 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0	1
	ID 2:	1 1 0 0 1 0 1 1 1 1 0 0 0 1 0 1 0 1 0 0	1
	ID 3:	1 0 1 0 0 0 1 1 1 0 1 0 1 1 0 0 0 0 0 1	1
	ID 4:	1 1 0 1 1 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1	1
Control	ID 5:	0 0 1 1 0 0 0 1 1 0 0 0 1 1 1 1 1 0 0 0	0
	ID 6:	0 1 0 1 1 0 1 1 0 0 0 0 1 1 0 0 1 0 1 0	0
	ID 7:	1 0 1 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 0	0
	ID 8:	1 1 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0 1	0

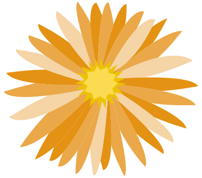




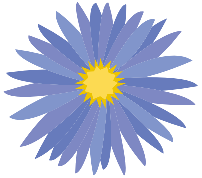
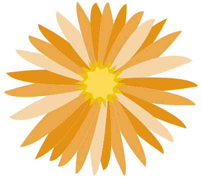
FACS [Papaxanthos et al. 2016] for Covariates

		SNPs (items)	Cov.	Signal
Case	ID 1:	0 0 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0	Europe	1
	ID 2:	1 1 0 0 1 0 1 1 1 1 0 0 0 1 0 1 0 1 0 0	Europe	1
	ID 3:	1 0 1 0 0 0 1 1 1 0 1 0 1 1 0 0 0 0 0 1	Asia	1
	ID 4:	1 1 0 1 1 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1	Asia	1
Control	ID 5:	0 0 1 1 0 0 0 1 1 0 0 0 1 1 1 1 1 0 0 0	Europe	0
	ID 6:	0 1 0 1 1 0 1 1 0 0 0 0 1 1 0 0 1 0 1 0	Europe	0
	ID 7:	1 0 1 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 0	Asia	0
	ID 8:	1 1 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0 1	Asia	0



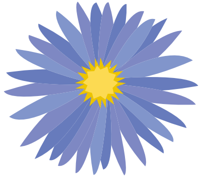


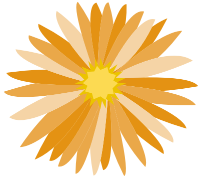
FAIS [Llinares-López et al. 2015] for Intervals

		SNPs (items)	Signal
Case 	ID 1:	001100111 001 11001110	1
	ID 2:	110010111 100 01010100	1
	ID 3:	101000111 010 11000001	1
	ID 4:	110110111 111 11010011	1
Control 	ID 5:	001100011 000 11111000	0
	ID 6:	010110110 000 11001010	0
	ID 7:	101100101 000 00101000	0
	ID 8:	110010010 101 00010101	1

FastCHM for Intervals + Cov.

		SNPs (items)	Cov.	Signal
Case	ID 1:	00110011100111001110	Europe	1
	ID 2:	11001011110001010100	Europe	1
	ID 3:	10100011101011000001	Asia	1
	ID 4:	11011011111111010011	Asia	1
Control	ID 5:	00110001100011111000	Europe	0
	ID 6:	01011011000011001010	Europe	0
	ID 7:	10110010100000101000	Asia	0
	ID 8:	11001001010100010101	Asia	1





C-Tarone [Sugiyama & Borgwardt, 2019]

- Find all feature interactions form continuous data

Input:

	F1	F2	F3	F4	F5	…	Class
ID1	–0.96	–3.03	3.38	2.57	–6.06	…	0
ID2	–1.80	4.45	–4.35	0.82	8.90	…	1
ID3	–3.29	1.39	–4.44	–0.77	2.78	…	1
ID4	–0.53	–1.96	–3.43	–4.42	–3.92	…	0
⋮			⋮				⋮

C-Tarone [Sugiyama & Borgwardt, 2019]

- Find all feature interactions form continuous data

Input:

	x						y
	F1	F2	F3	F4	F5	…	Class
ID1	–0.96	–3.03	3.38	2.57	–6.06	…	0
ID2	–1.80	4.45	–4.35	0.82	8.90	…	1
ID3	–3.29	1.39	–4.44	–0.77	2.78	…	1
ID4	–0.53	–1.96	–3.43	–4.42	–3.92	…	0
⋮			⋮				⋮

$\longrightarrow$

Output:

{F1}, {F3},
{F2, F5},
{F2, F5, F6}, …

Use Copula Support [Tatti, 2013]

	F1	F2	F3
$\boldsymbol{x}_1$	–0.96	–3.03	3.38
$\boldsymbol{x}_2$	–1.80	4.45	–4.35
$\boldsymbol{x}_3$	–3.29	1.39	–4.44
$\boldsymbol{x}_4$	–0.53	–1.96	–3.43

$\xrightarrow{\text{Rank}}$

R(F1)	R(F2)	R(F3)
2	0	3
1	3	1
0	2	0
3	1	2

$\xrightarrow{\text{Norm.}}$

π(F1)	π(F2)	π(F3)
0.67	0.00	1.00
0.34	1.00	0.34
0.00	0.67	0.00
1.00	0.34	0.67

$\xrightarrow{\text{Prod.}}$

0.00
0.11
0.00
0.22

$\xrightarrow{\text{Sum / 4}}$ $0.083 = \Pr(X_{\{F1,F2,F3\}} = 1) = \eta(\{F1,F2,F3\})$

Contingency Tables

- For each pattern (variable combination), we construct two types of contingency tables
  - One is from the expected situation under null
  - The other is from the observed situation from data
- Significance is assessed by comparison of the two tables
  - Each table is represented as a four-dimensional vector

Expected for $\boldsymbol{p}_{\mathrm{E}}$	$X_{\mathcal{F}} = 1$	$X_{\mathcal{F}} = 0$	Total
$Y = 1$	$\eta(\mathcal{F})\, r_1$	$r_1 - \eta(\mathcal{F})\, r_1$	r_1
$Y = 0$	$\eta(\mathcal{F})\, r_0$	$r_0 - \eta(\mathcal{F})\, r_0$	r_0
Total	$\eta(\mathcal{F})$	$1 - \eta(\mathcal{F})$	1

Observed for $\boldsymbol{p}_{\mathrm{O}}$	$X_{\mathcal{F}} = 1$	$X_{\mathcal{F}} = 0$	Total
$Y = 1$	$\eta(\mathcal{F}, Y = 1)$	$r_1 - \eta(\mathcal{F}, Y = 1)$	r_1
$Y = 0$	$\eta(\mathcal{F}, Y = 0)$	$r_0 - \eta(\mathcal{F}, Y = 0)$	r_0
Total	$\eta(\mathcal{F})$	$1 - \eta(\mathcal{F})$	1

Significance Test

- The independence $X_{\mathcal{F}} \perp\!\!\!\perp Y$ is translated into:

  $$H_0 : D_{\mathrm{KL}}(\boldsymbol{p}_{\mathrm{O}}, \boldsymbol{p}_{\mathrm{E}}) = 0, \quad H_1 : D_{\mathrm{KL}}(\boldsymbol{p}_{\mathrm{O}}, \boldsymbol{p}_{\mathrm{E}}) \neq 0$$

  - $\boldsymbol{p}_{\mathrm{E}}$ and $\boldsymbol{p}_{\mathrm{O}}$ are vectorized contingency tables:

    $$\boldsymbol{p}_{\mathrm{E}} = \big(\eta(\mathcal{F})r_1, \eta(\mathcal{F})r_0, r_1 - \eta(\mathcal{F})r_1, r_0 - \eta(\mathcal{F})r_0\big)$$

    $$\boldsymbol{p}_{\mathrm{O}} = \big(\eta(\mathcal{F}, Y=1), \eta(\mathcal{F}, Y=0), r_1 - \eta(\mathcal{F}, Y=1), r_0 - \eta(\mathcal{F}, Y=0)\big)$$

- We apply G-test: the statistic $\lambda = 2ND_{\mathrm{KL}}(\boldsymbol{p}_{\mathrm{O}}, \boldsymbol{p}_{\mathrm{E}})$ follows the χ^2-distribution with the d.f. 1

KL Divergence Bound

- **Theorem** (tight upper bound of KL divergence):

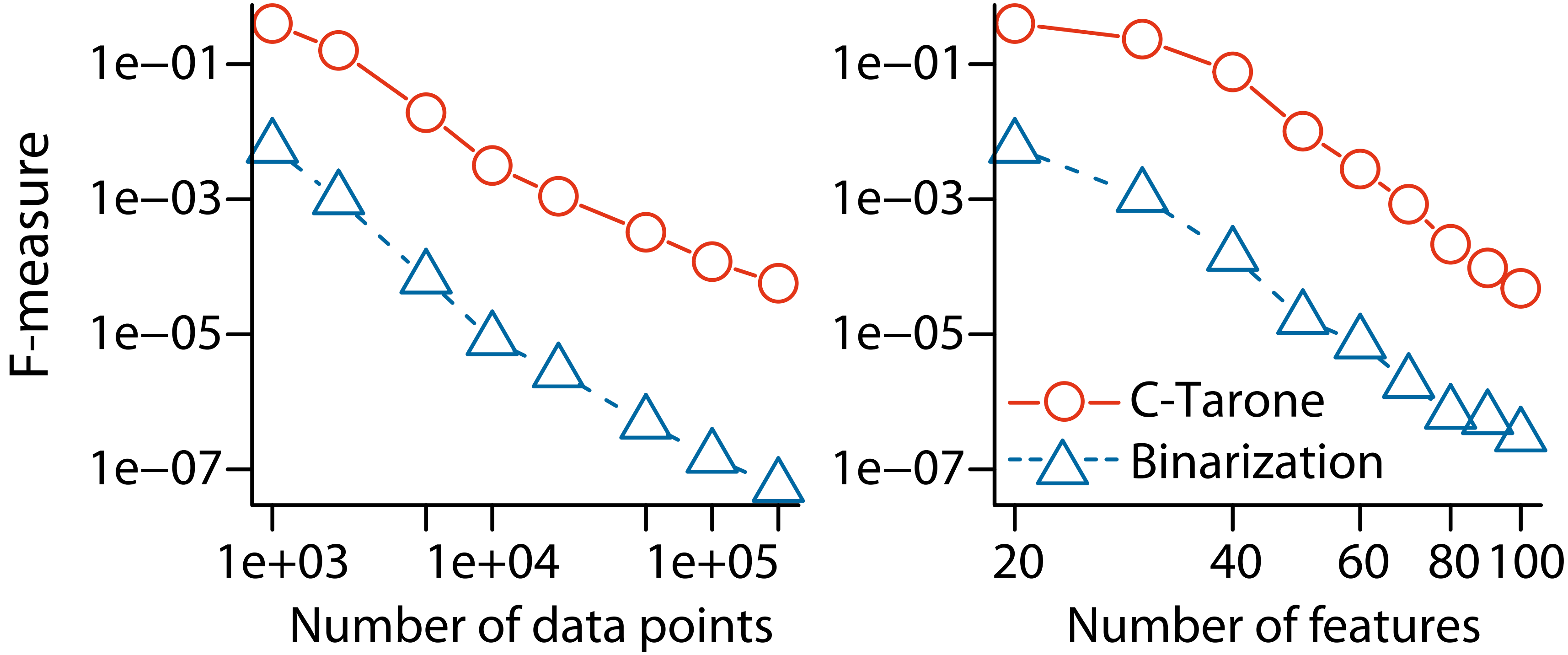
  $$D_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{p}_{\mathrm{E}})$$

  $$< a \log \frac{1}{b} + (b - a) \log \frac{b - a}{(1 - a)b} + (1 - b) \log \frac{1}{(1 - a)}$$

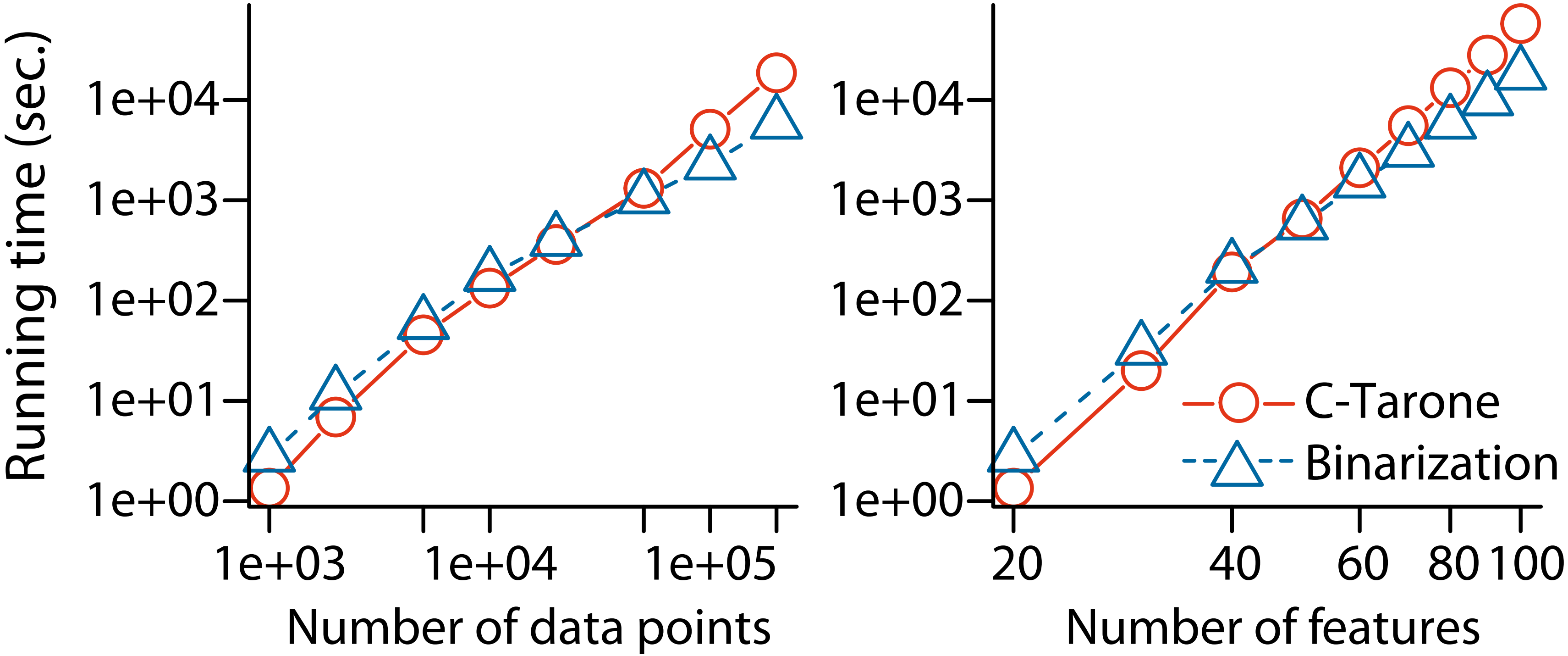
  - $\boldsymbol{p}_{\mathrm{E}} = (ab, a(1 - b), (1 - a)b, (1 - a)(1 - b))$,
    $\boldsymbol{p} \in \{ \boldsymbol{p} \in \mathcal{P} \mid p_1 + p_2 = a,\ p_1 + p_3 = b \}$

- The p-value for this upper bound is the minimum achievable p-value

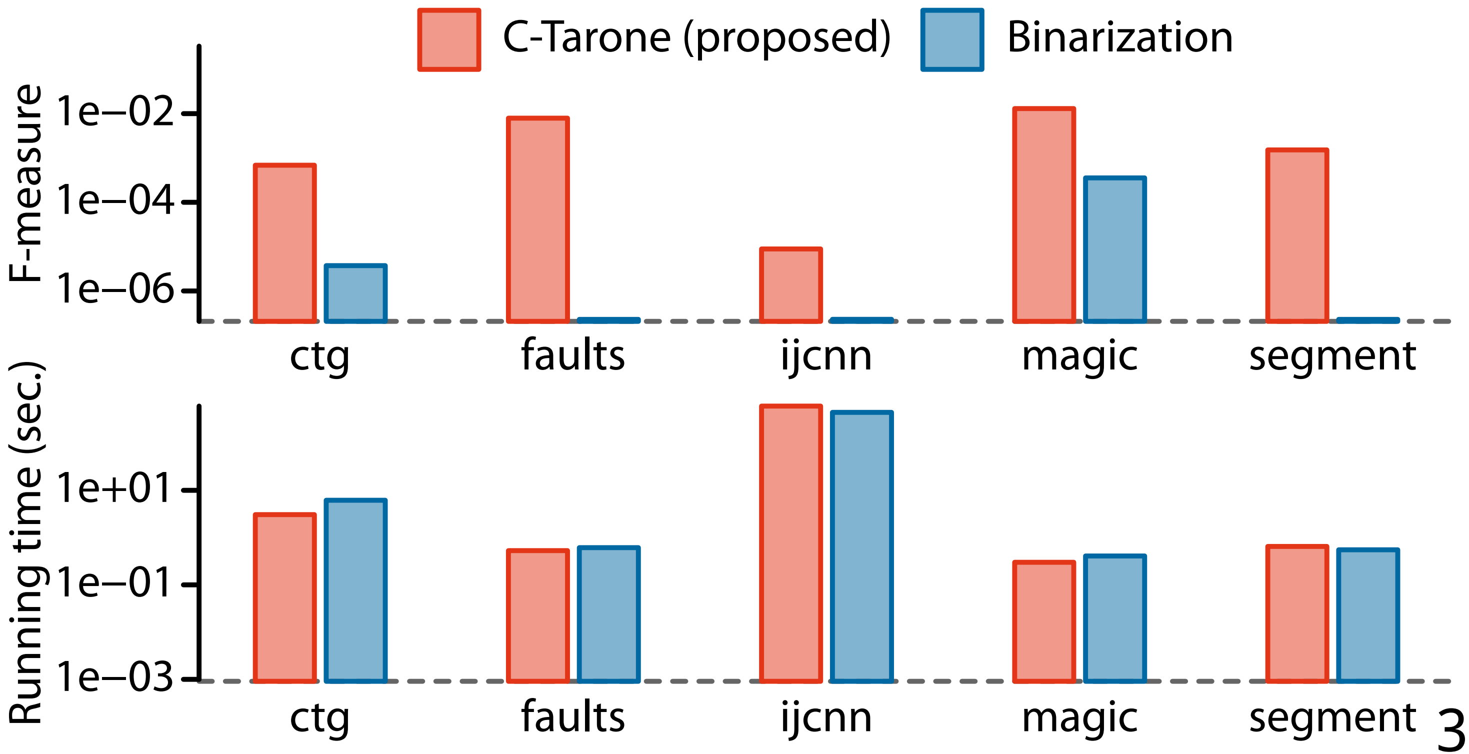
Exp. Results on Synthetic Data



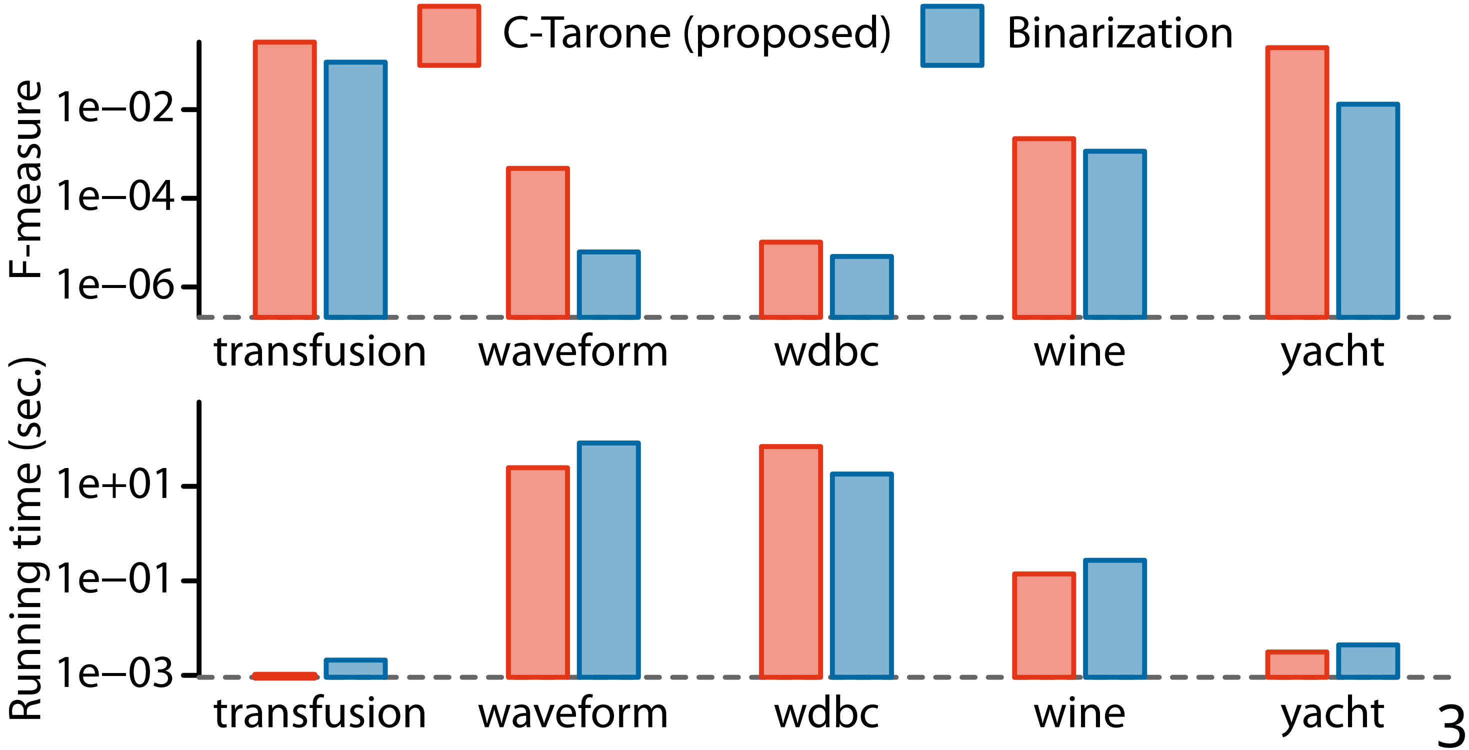
Exp. Results on Synthetic Data



Experimental Results on Real Data



Experimental Results on Real Data



Conclusion

- Significant pattern mining is introduced
  - Find significant interactions while controlling the FWER
  - pattern mining (data mining) + multiple testing correction (statistics)
- Key to solve the problem is Tarone's testability trick
  - This method can be used if the minimum achievable p-value ψ exists
  - If we have the relationship "Smaller $\eta \rightarrow$ Larger ψ", Apriori can be used to efficiently enumerate testable patterns