August 15, 2019
IJCAI 2019

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

PRESTO
SAKIGAKE

Finding Statistically Significant Interactions between Continuous Features

Mahito Sugiyama (National Institute of Informatics)
Karsten Borgwardt (ETH Zürich)

Our Proposal: *C-Tarone*

- Find all feature interactions that are significantly associated with class labels from multivariate data with controlling the FWER

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
⋮			⋮				⋮

(F1–... grouped as $\boldsymbol{x}$; Class as y)

Our Proposal: *C-Tarone*

- Find all feature interactions that are significantly associated with class labels from multivariate data with controlling the FWER

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
⋮			⋮				⋮

Output:

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

Existing Method: Significant Pattern Mining

- So far only binary (or discrete) data can be used
  $\rightarrow$ Results obtained by SPM via binarization can be uninformative!

Input:

	$\boldsymbol{x}$						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
⋮			⋮				⋮

$\longrightarrow$

Output:

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

We solve:

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

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 of the candidate space?
   - The number of possible interactions is 2^d for d features

Problem Formulation

- Define $X_{\mathcal{F}}$ as the binary random variable of joint occurrence for a feature combination $\mathcal{F} = \{F_i, F_{i+1}, \dots, F_{i+k}\}$
  - $X_{\mathcal{F}} = 1$ if $\mathcal{F}$ "occurs", $X_{\mathcal{F}} = 0$ otherwise
- Let Y be an output binary variable
- **Our 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
- We need to estimate the probability $\Pr(X_{\mathcal{F}})$ from data

Copula Support [Tatti, 2013] for $\Pr(X_{\mathcal{F}} = 1)$

	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

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

Contingency Tables

Expected (under null) 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 the condition:

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

  - $\boldsymbol{p}_{\mathrm{E}}$ 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

Multiple Testing Correction

- The FWER should be controlled
  - Probability that at least one feature combination is a false positive
  - If we naïvely test all combinations, $\alpha 2^d$ false positives could occur!!

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

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

$$D_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{p}_{\mathrm{E}}) < a \log \frac{1}{b} + (b-a) \log \frac{b-a}{(1-a)b} + (1-b) \log \frac{1}{(1-a)}$$

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

Tarone's Testability Trick

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

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

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

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

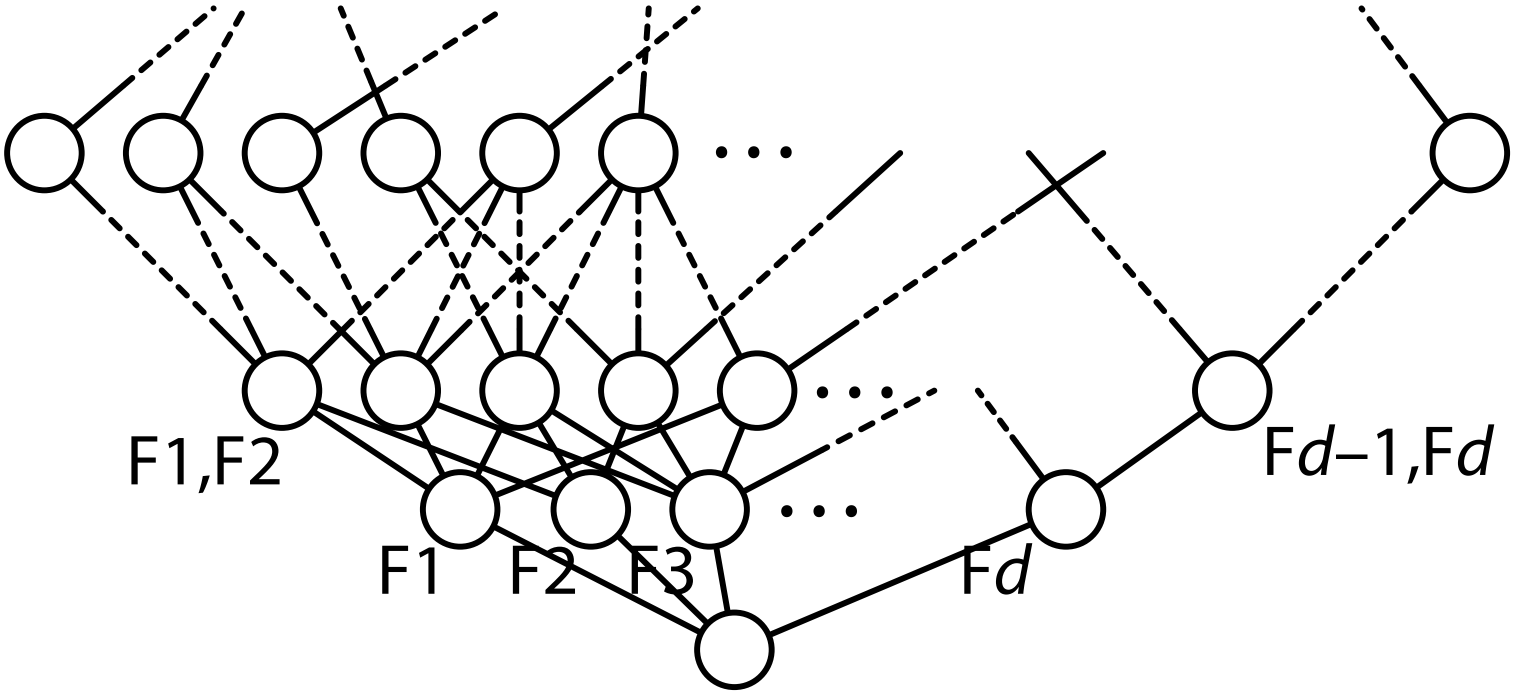
Untestable combinations $\rightarrow$ Prune without testing

Testable combinations $\downarrow$

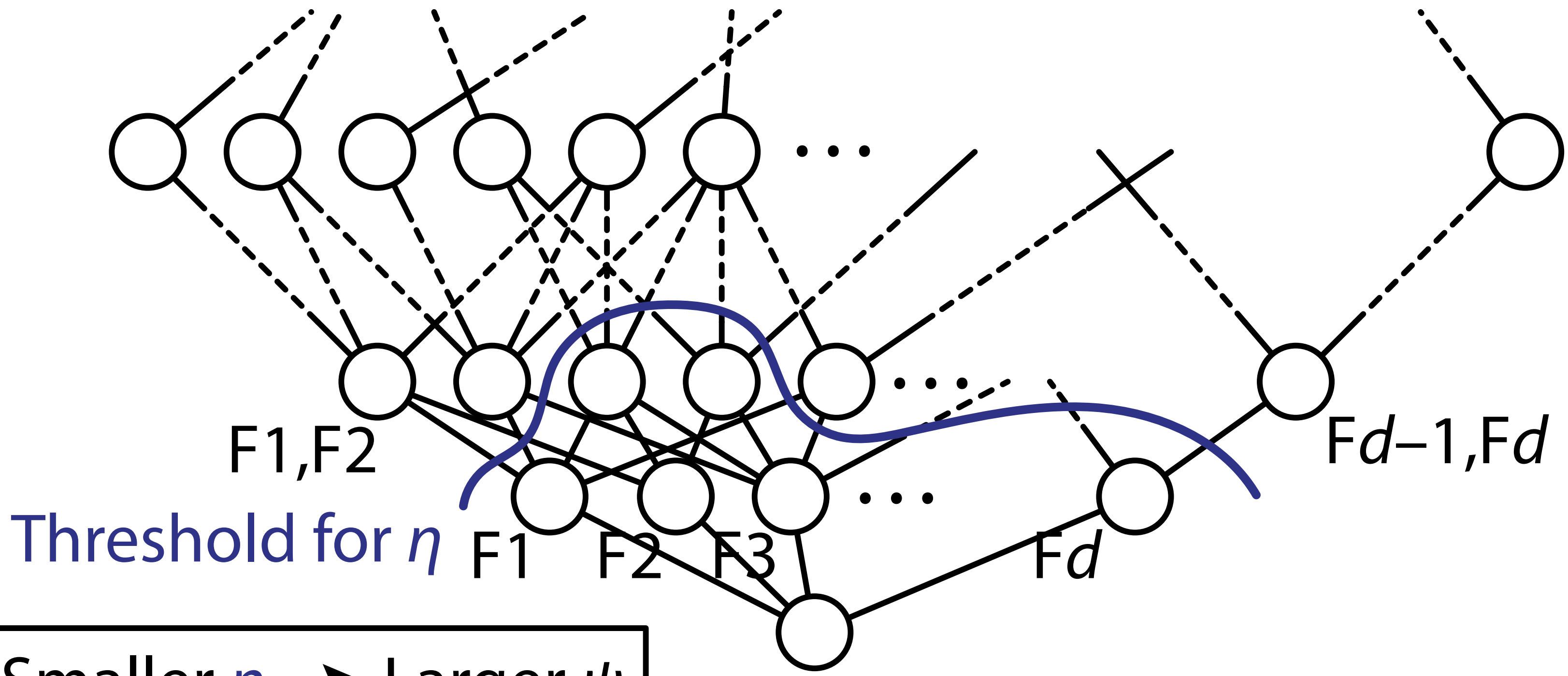
$\mathcal{F}_i$ is significant if: p-value$(\mathcal{F}_i) < \alpha / m$

Correction factor

Enumeration Algorithm Based on Apriori

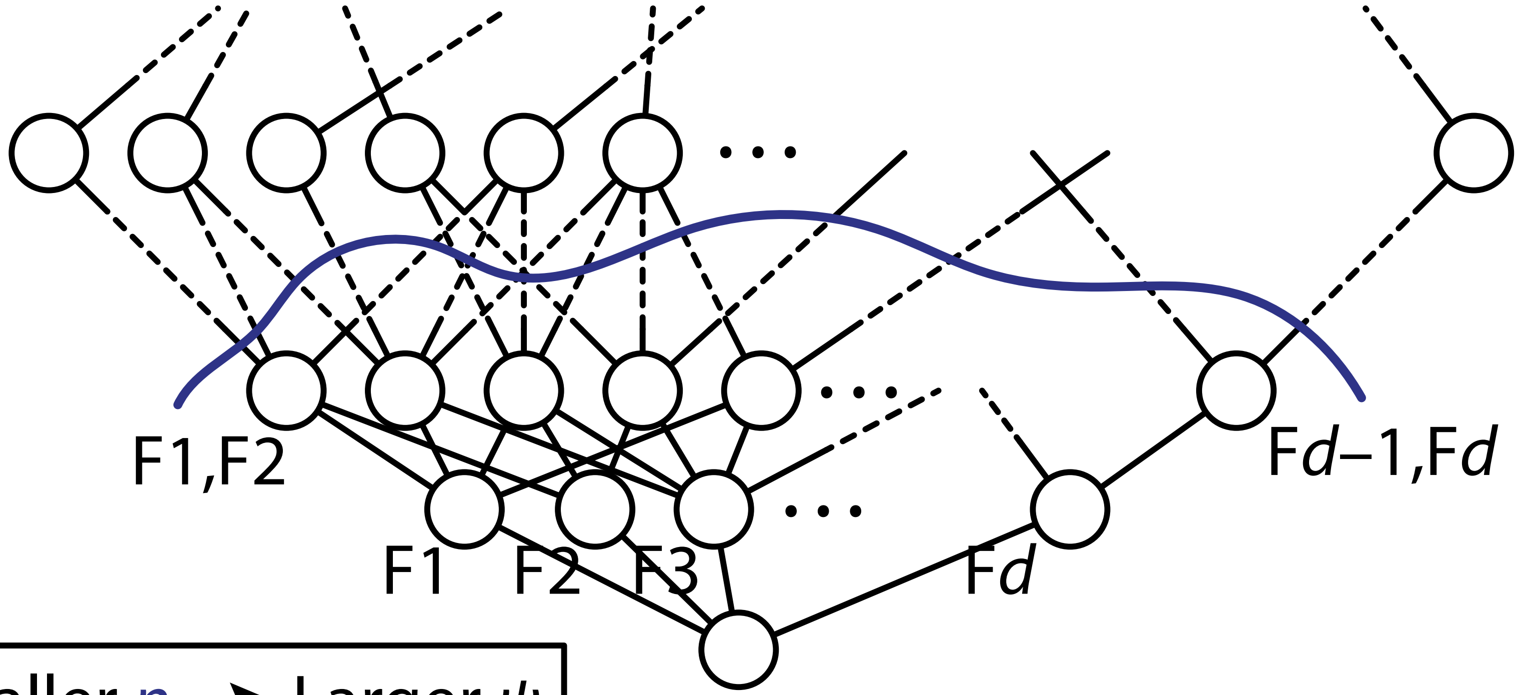


Enumeration Algorithm Based on Apriori



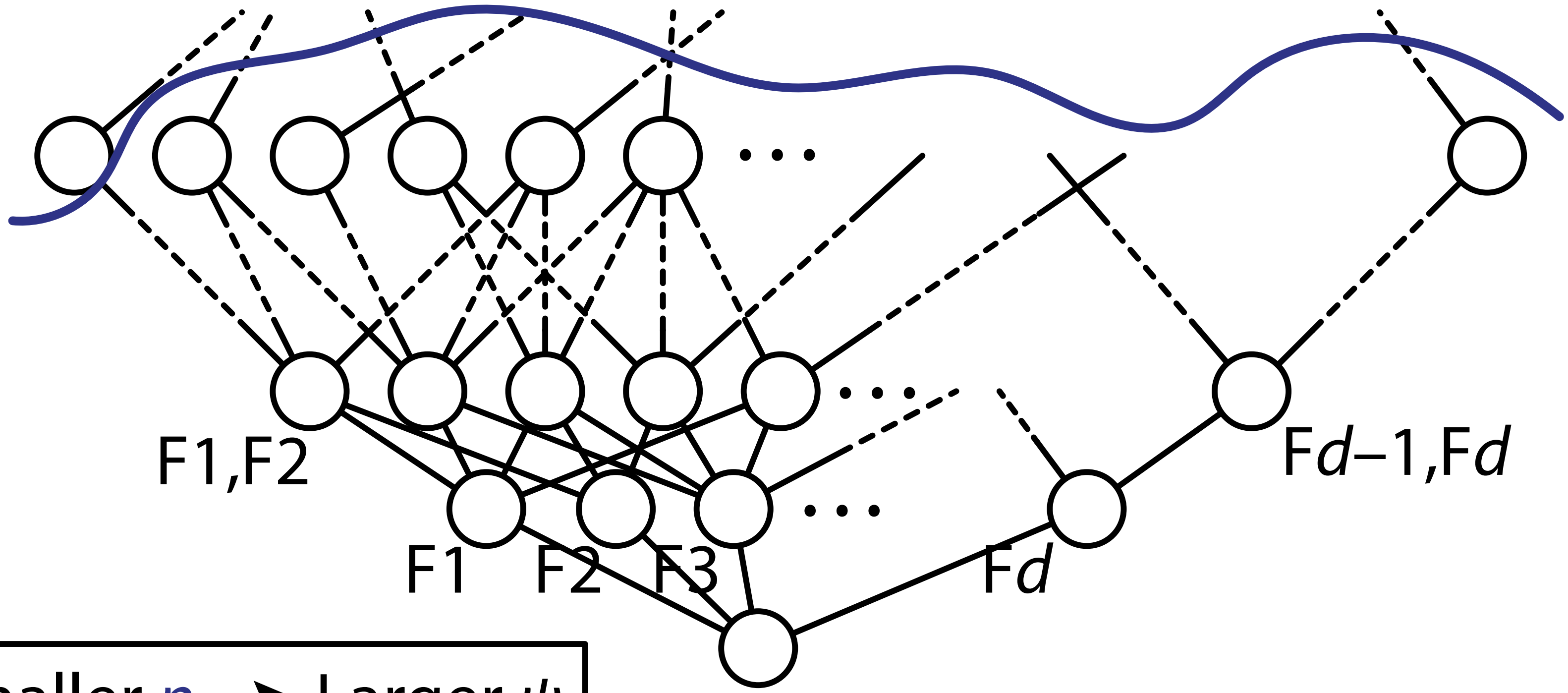
Smaller $\eta \longrightarrow$ Larger ψ

Enumeration Algorithm Based on Apriori



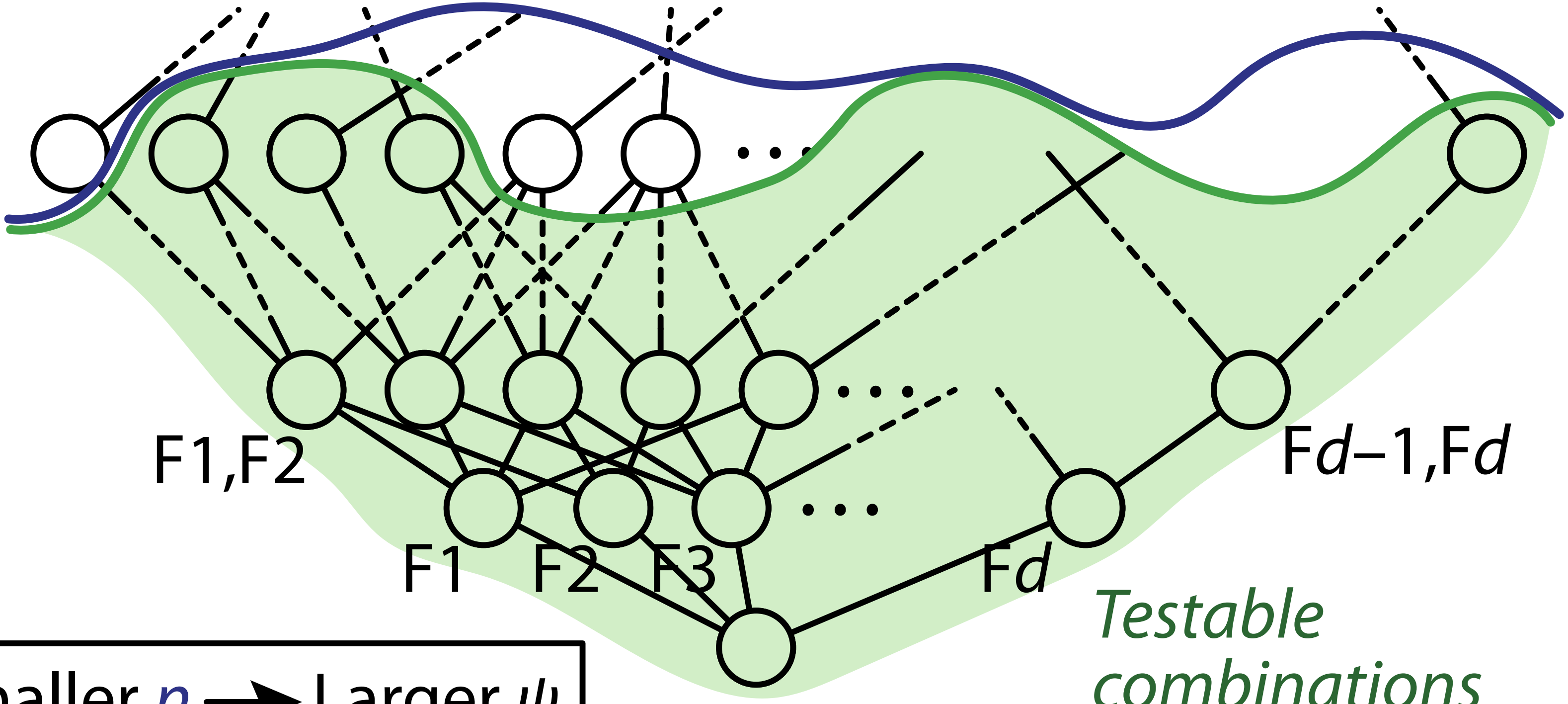
Smaller η → Larger ψ

Enumeration Algorithm Based on Apriori



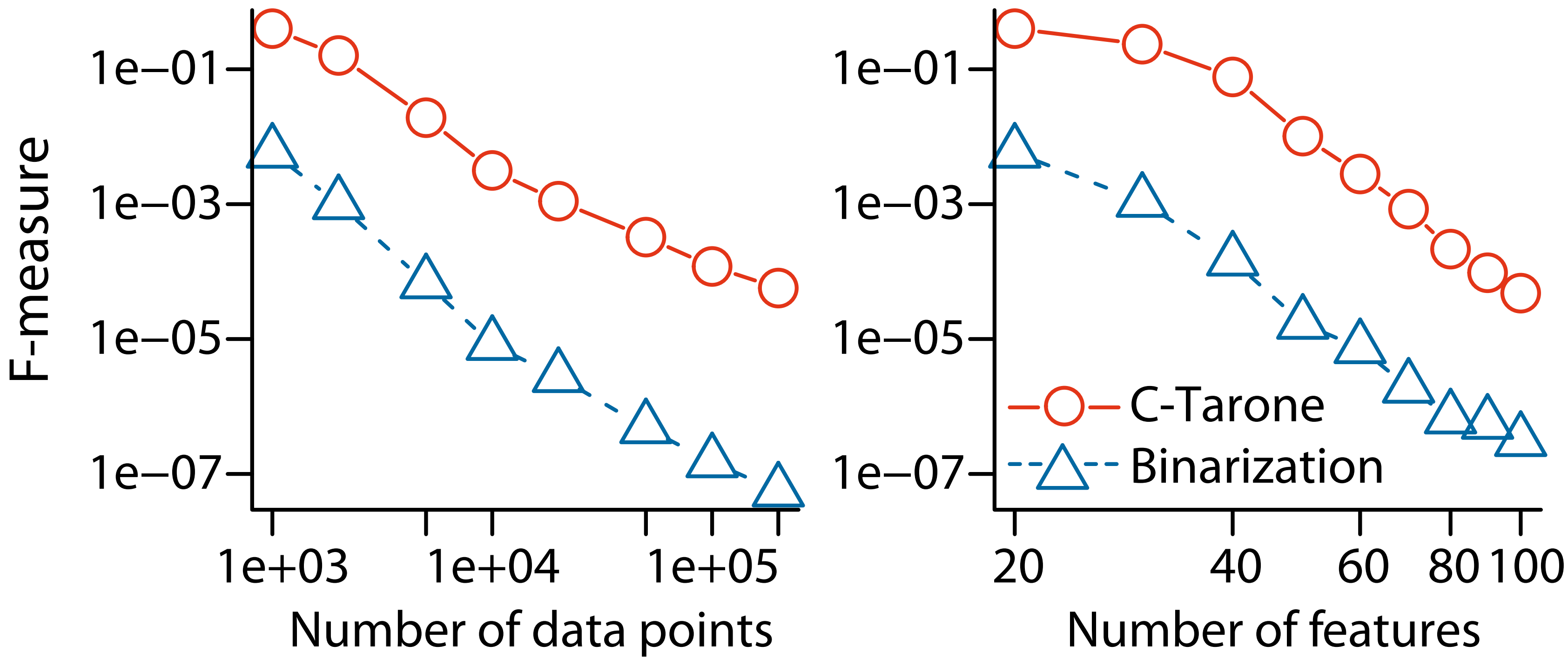
Smaller η ⟶ Larger ψ

Enumeration Algorithm Based on Apriori

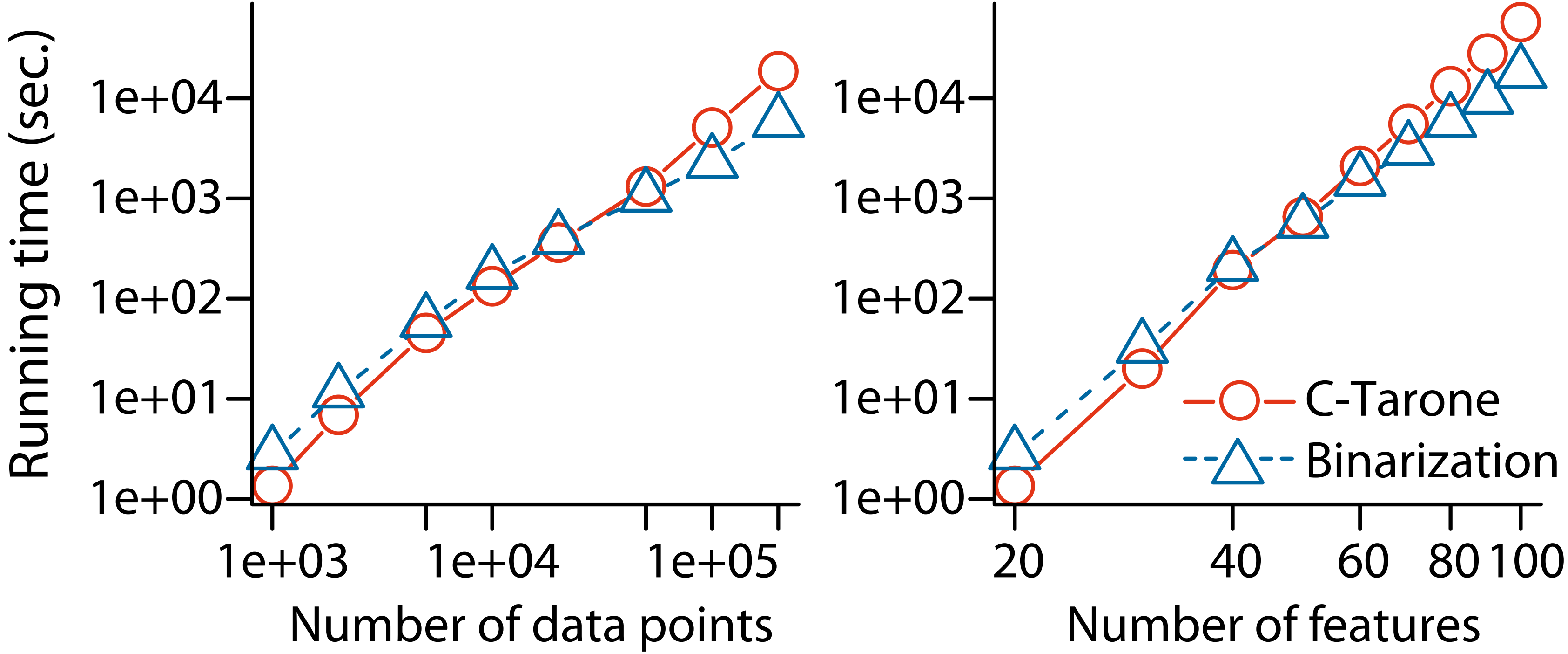


Smaller $\eta \longrightarrow$ Larger ψ

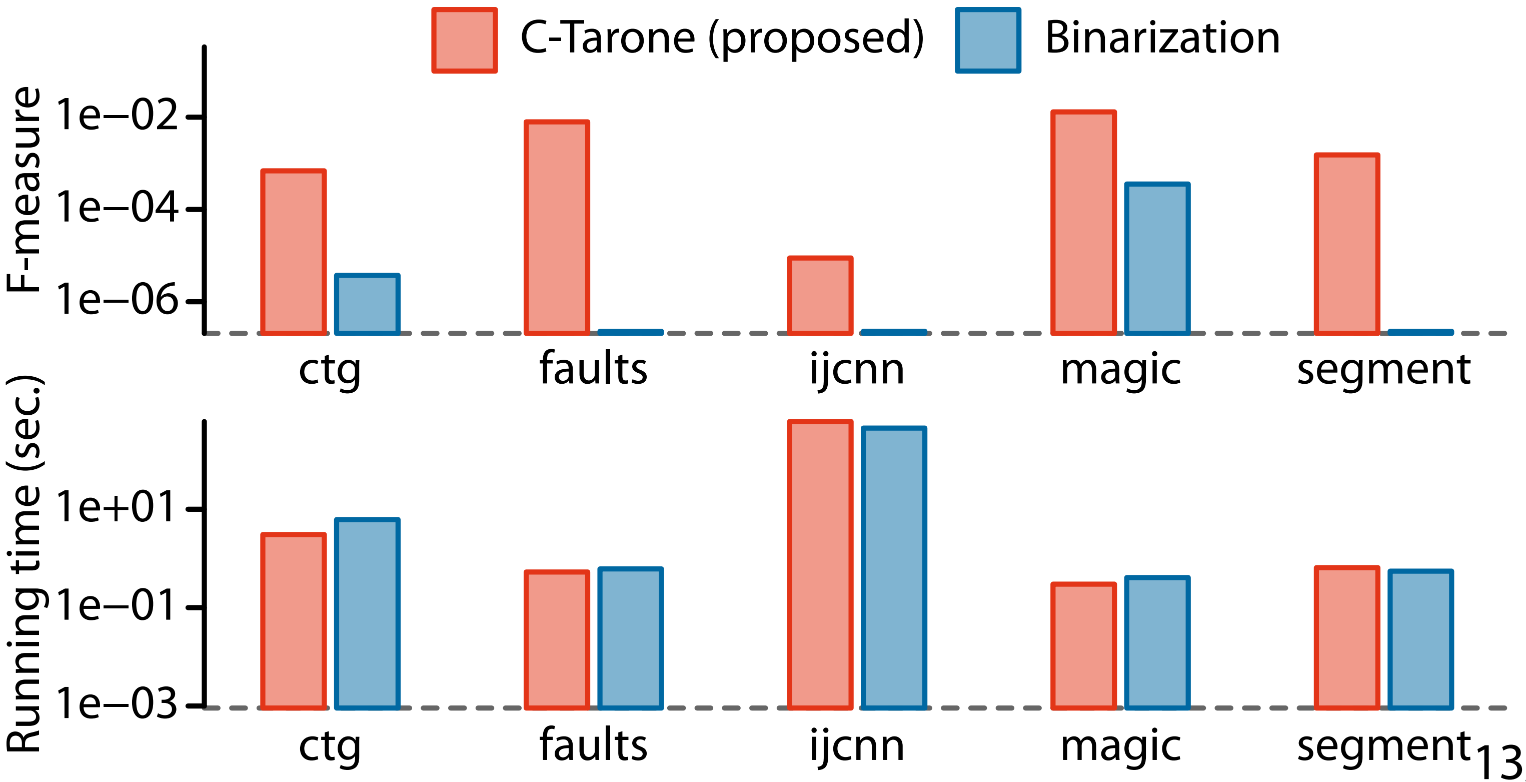
Experimental Results on Synthetic Data



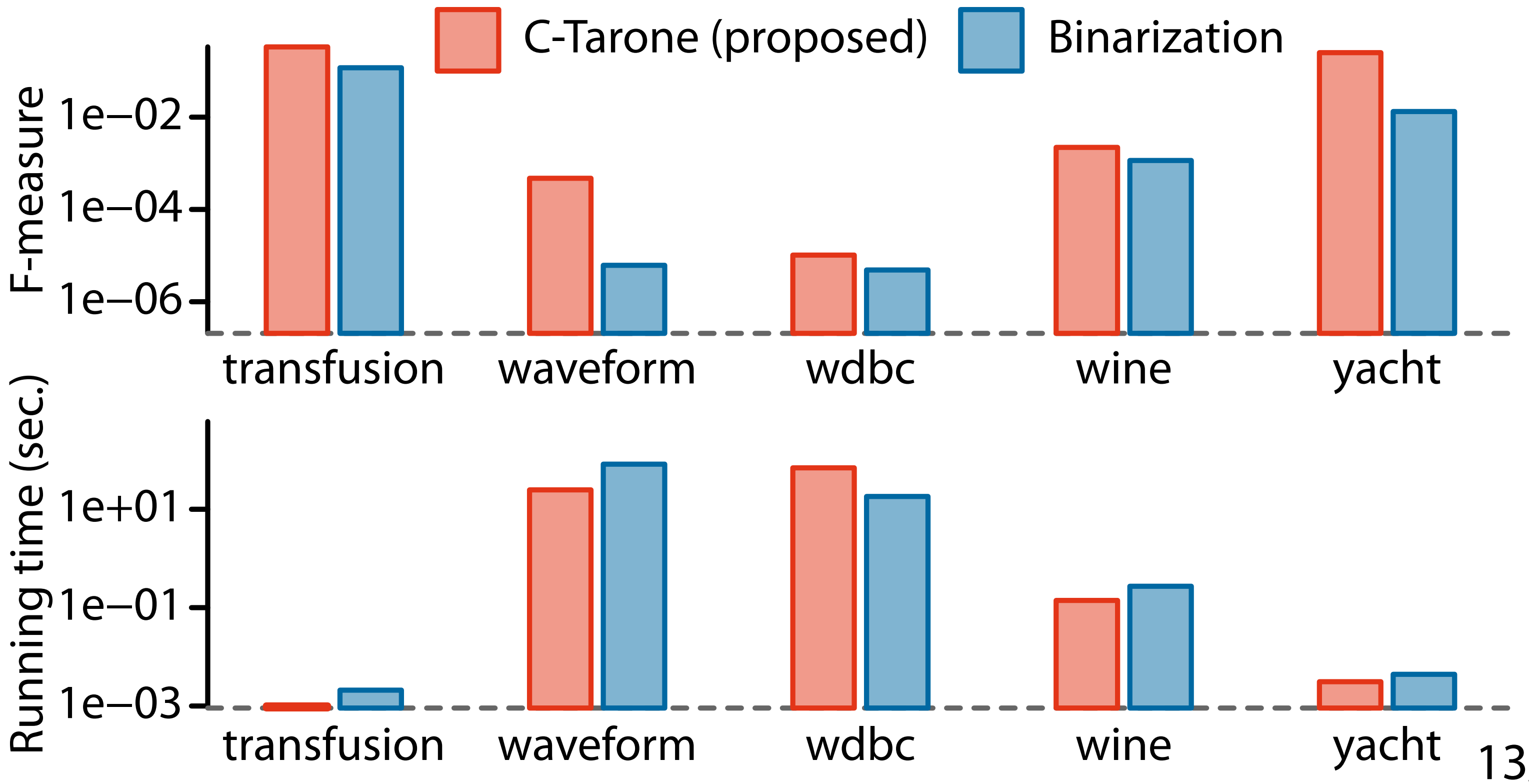
Experimental Results on Synthetic Data



Experimental Results on Real Data



Experimental Results on Real Data



Conclusion

- We have proposed C-Tarone, a solution to the open problem of finding *all* multiplicative interactions between continuous features significantly associated with an output variable
  - Significance is rigorously controlled for multiple testing

- Our work opens the door to many applications of searching significant feature combinations, in which the data is not adequately described by binary features