DRMI: A Dataset Reduction Technology based on Mutual Information for Black-box Attacks

Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, and Jinwen He

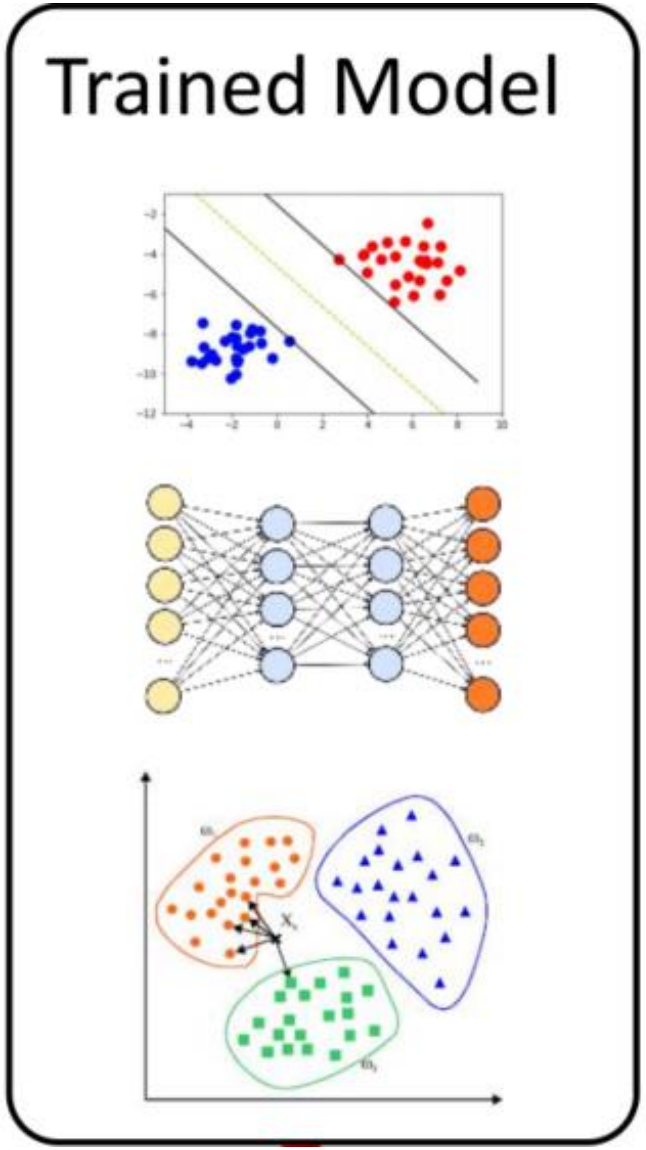
SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, China

School of Cyber Security, University of Chinese Academy of Sciences, China

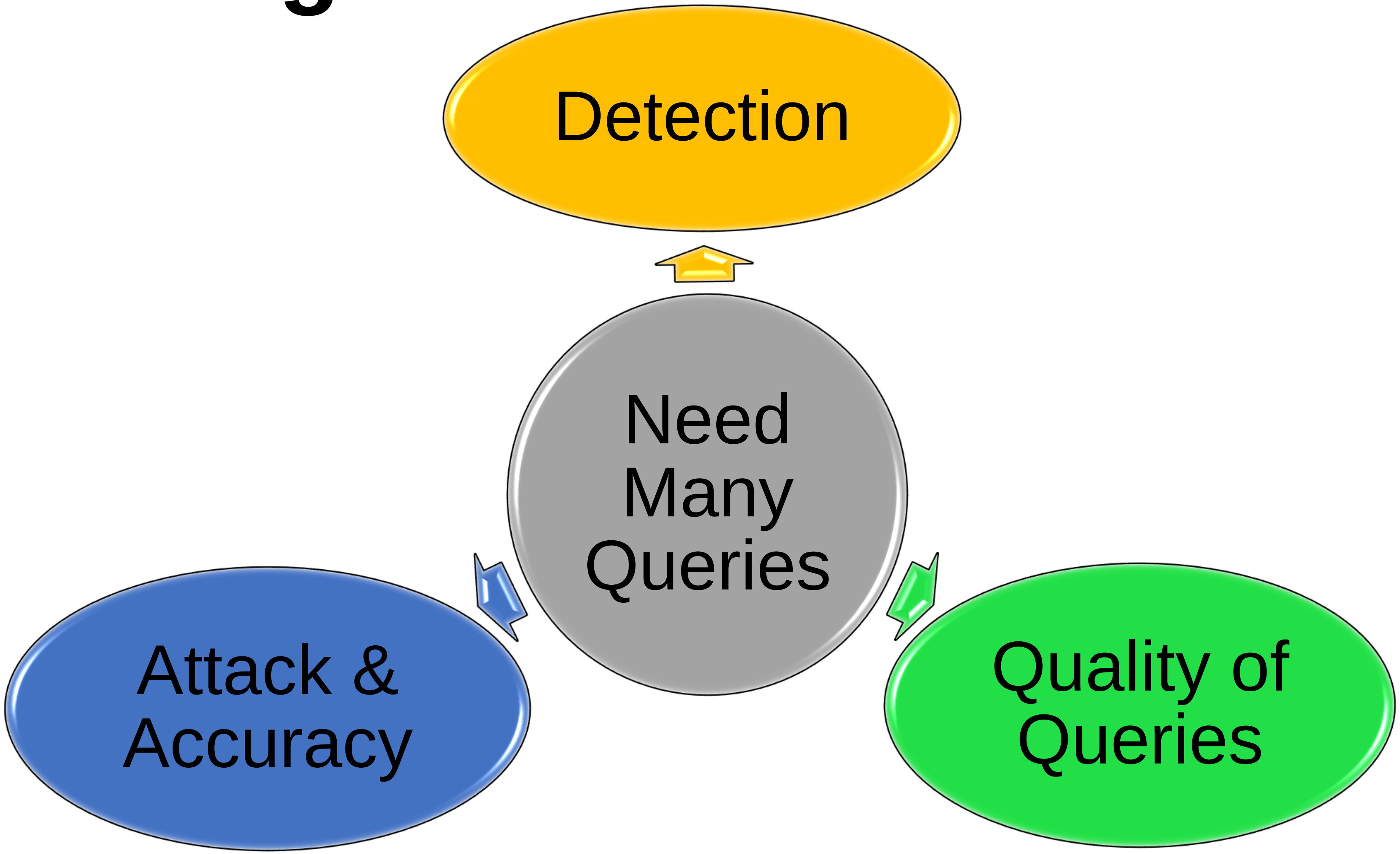
Background

Attacker

Steal



Challenges


Detection
Need Many Queries
Attack & Accuracy
Quality of Queries

Threat Model

We need

Dataset of Similar Distribution

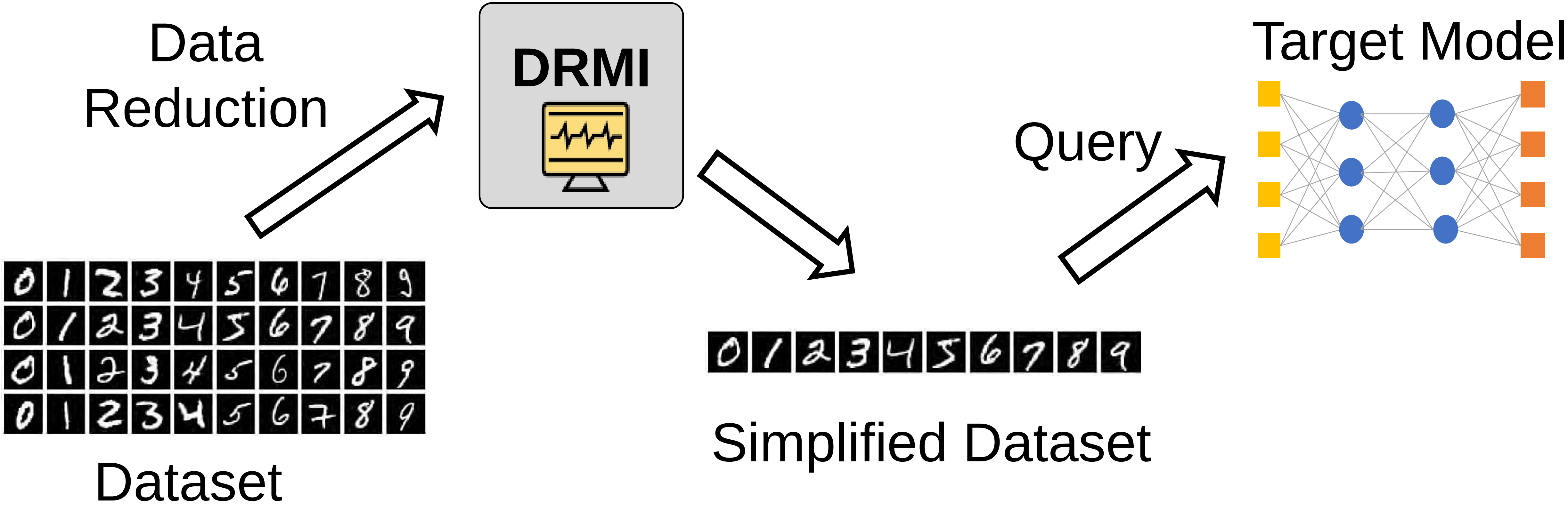
Classification Task

Model Architecture

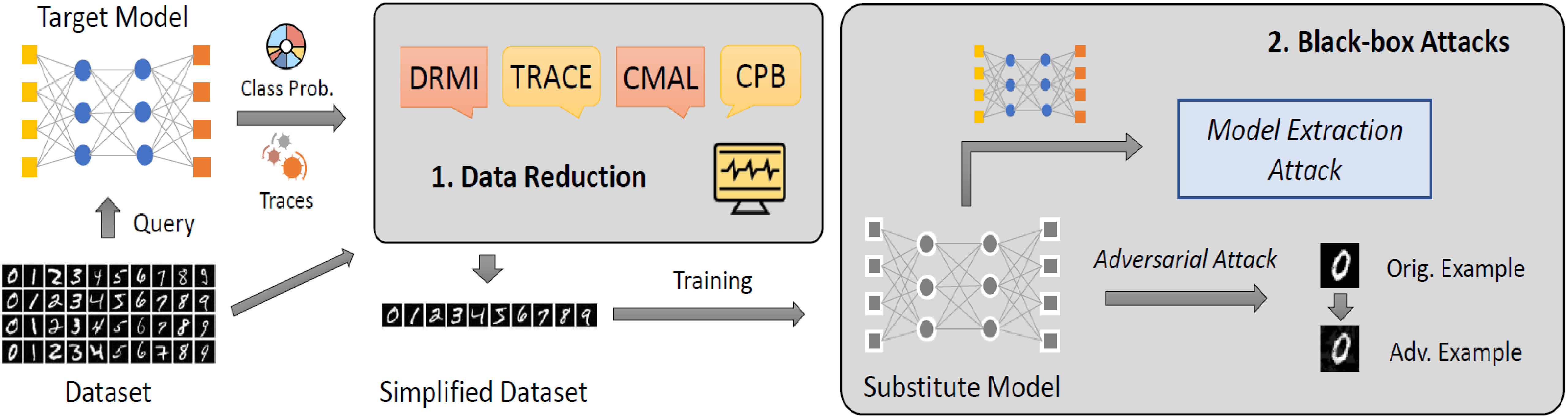
Target Dataset

Confidence Scores

Workflow


Data
Reduction
DRMI
Target Model
Query
Simplified Dataset
Dataset

Workflow



Target Model
Class Prob.
Traces
Query
Dataset
DRMI
TRACE
CMAL
CPB
1. Data Reduction
Simplified Dataset
Training
2. Black-box Attacks
Model Extraction Attack
Adversarial Attack
Orig. Example
Adv. Example
Substitute Model

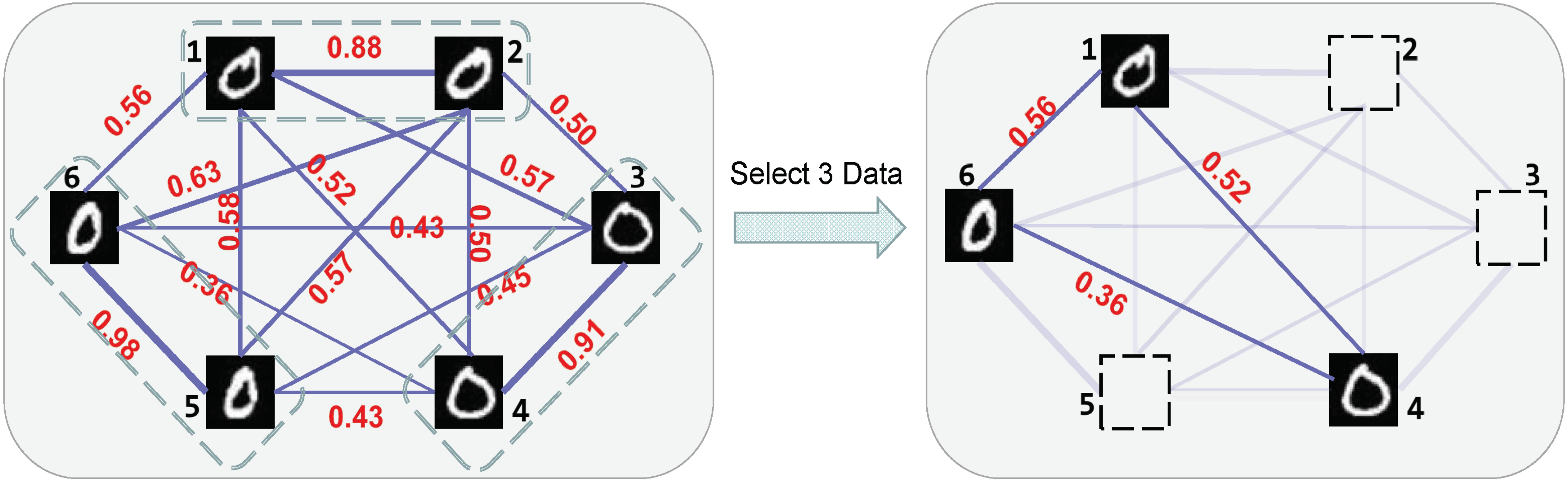
Approach: DRMI

- High mutual information means high redundancy
- Goal of DRMI
    - select a more representative reduced dataset through minimizing the mutual information value

Approach: DRMI

A case

The minimal sum of mutual information is 1.44



Approach: DRMI

- The mutual information value of image $\boldsymbol{u}$ and $\boldsymbol{v}$ is calculated as

$$MI(u)(v) = \sum_{i=0}^{R} \sum_{j=0}^{R} P_{uv}(i,j) \log \frac{P_{uv}(i,j)}{P_u(i)P_v(j)}$$

Approach: DRMI

- The mutual information value of image $\boldsymbol{u}$ and $\boldsymbol{v}$ is calculated as

$$MI(u)(v) = \sum_{i=0}^{R} \sum_{j=0}^{R} P_{uv}(i,j) \log \frac{P_{uv}(i,j)}{P_u(i)P_v(j)}$$

- Use a matrix $\boldsymbol{I}$ and a hyperparameter $\boldsymbol{\alpha}$ to represent the mutual information

$$I[u][v] = MI(u)(v)^{\alpha}$$

Approach: DRMI

- Formalized Goal of DRMI

$$\arg\min_{S} H = \frac{1}{2} \sum_{i \in S} \sum_{j \in S} I[i][j], \; i \neq j$$

Approach: DRMI

- Formalized Goal of DRMI

$$\arg\min_{S} H = \frac{1}{2} \sum_{i \in S} \sum_{j \in S} I[i][j], \; i \neq j$$

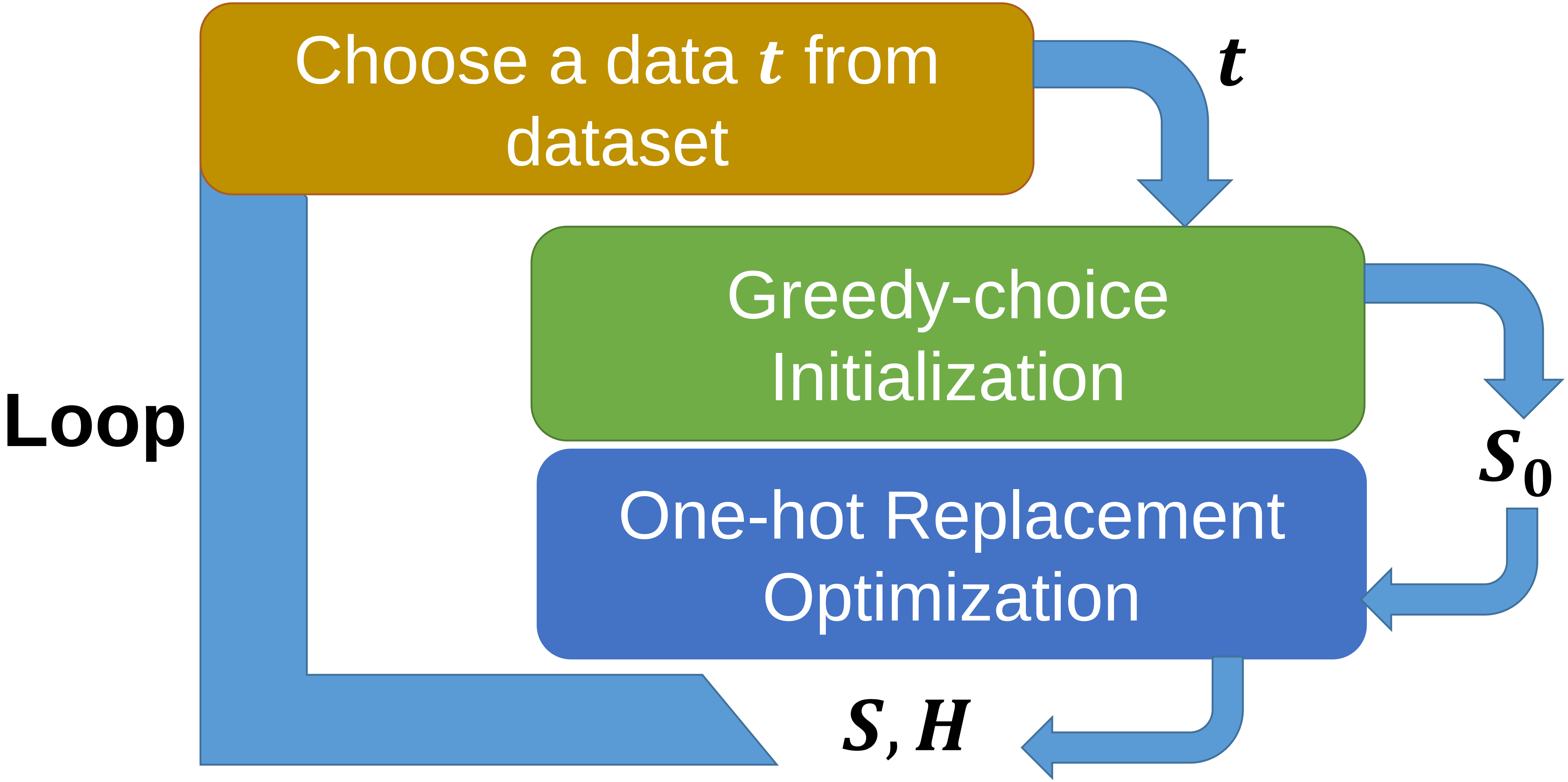
- Mapping it to Graph Theory

$$\arg\min_{G[S]} H = \sum_{e=(u,v)} w(e), \;\; u, v \in S, u \neq v, \; and \; e \in E$$

Approach: DRMI

- Proof of NP-Complete
  - Proof of NP
    - Verifiable in polynomial time
  - Proof of NP-Hard
    - The maximum independent set problem can be reduced to ours.

Approach: DRMI



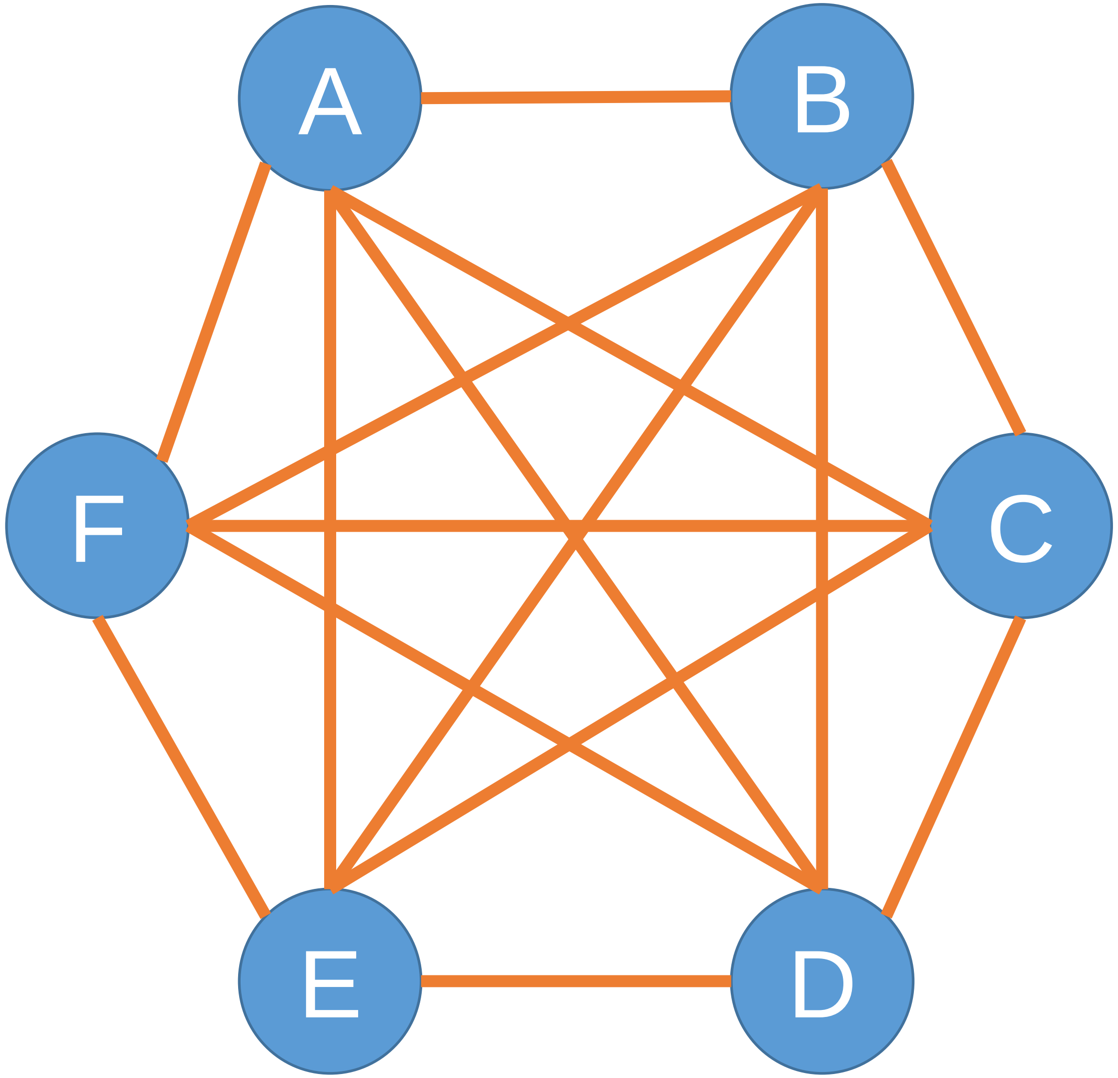
Approach: DRMI

Greedy-choice Initialization

6→3

$S_0 = \emptyset$

MI	A	B	C	D	E	F
A	/	0.88	0.57	0.52	0.58	0.56
B	0.88	/	0.50	0.50	0.57	0.63
C	0.57	0.50	/	0.91	0.45	0.43
D	0.52	0.50	0.91	/	0.43	0.36
E	0.58	0.57	0.45	0.43	/	0.98
F	0.56	0.63	0.43	0.36	0.98	/



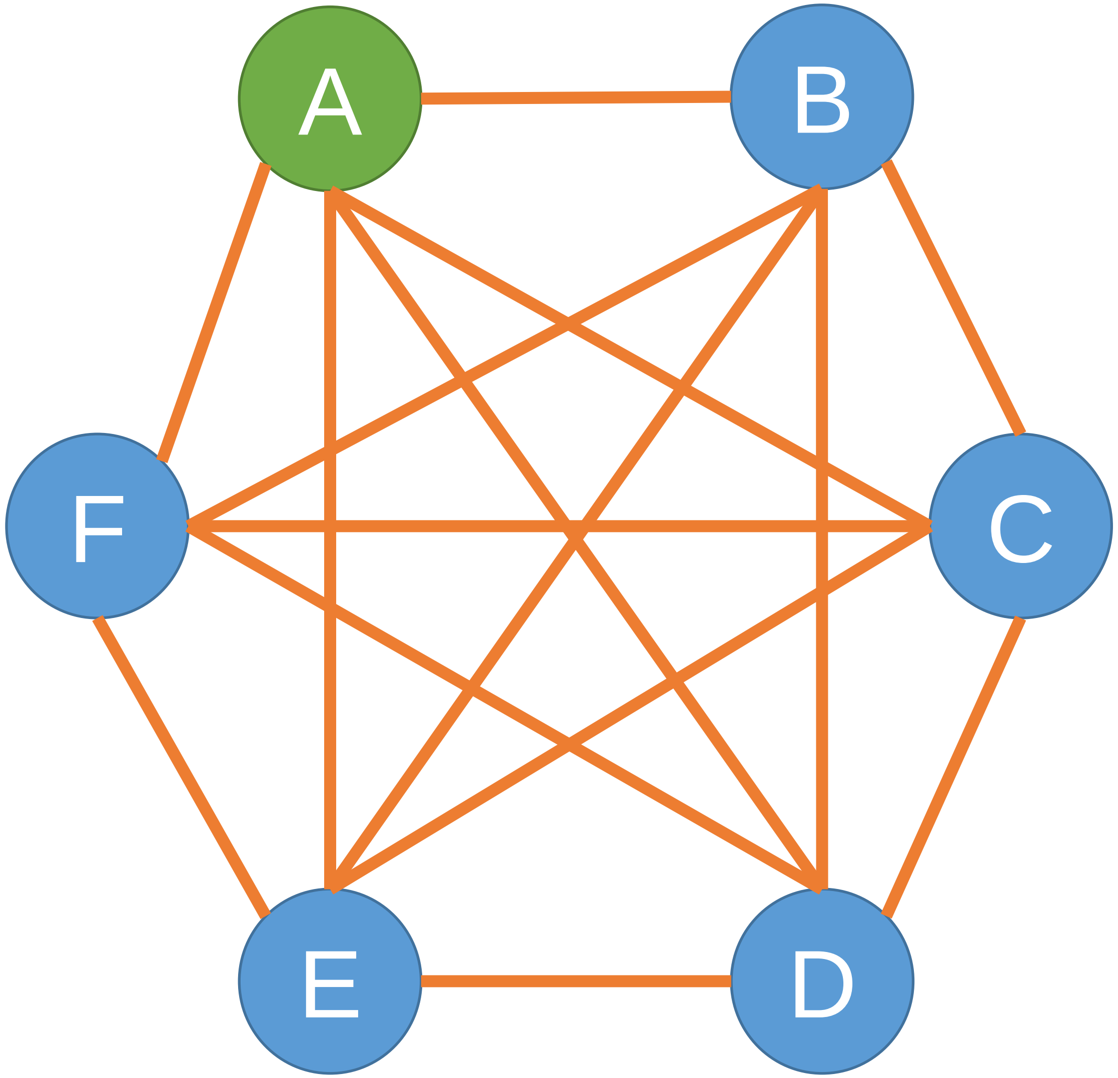
Approach: DRMI

Greedy-choice Initialization

$$t = A$$

$$S_0 = \{A\}$$

MI	A	B	C	D	E	F
A	/	0.88	0.57	0.52	0.58	0.56
B	0.88	/	0.50	0.50	0.57	0.63
C	0.57	0.50	/	0.91	0.45	0.43
D	0.52	0.50	0.91	/	0.43	0.36
E	0.58	0.57	0.45	0.43	/	0.98
F	0.56	0.63	0.43	0.36	0.98	/

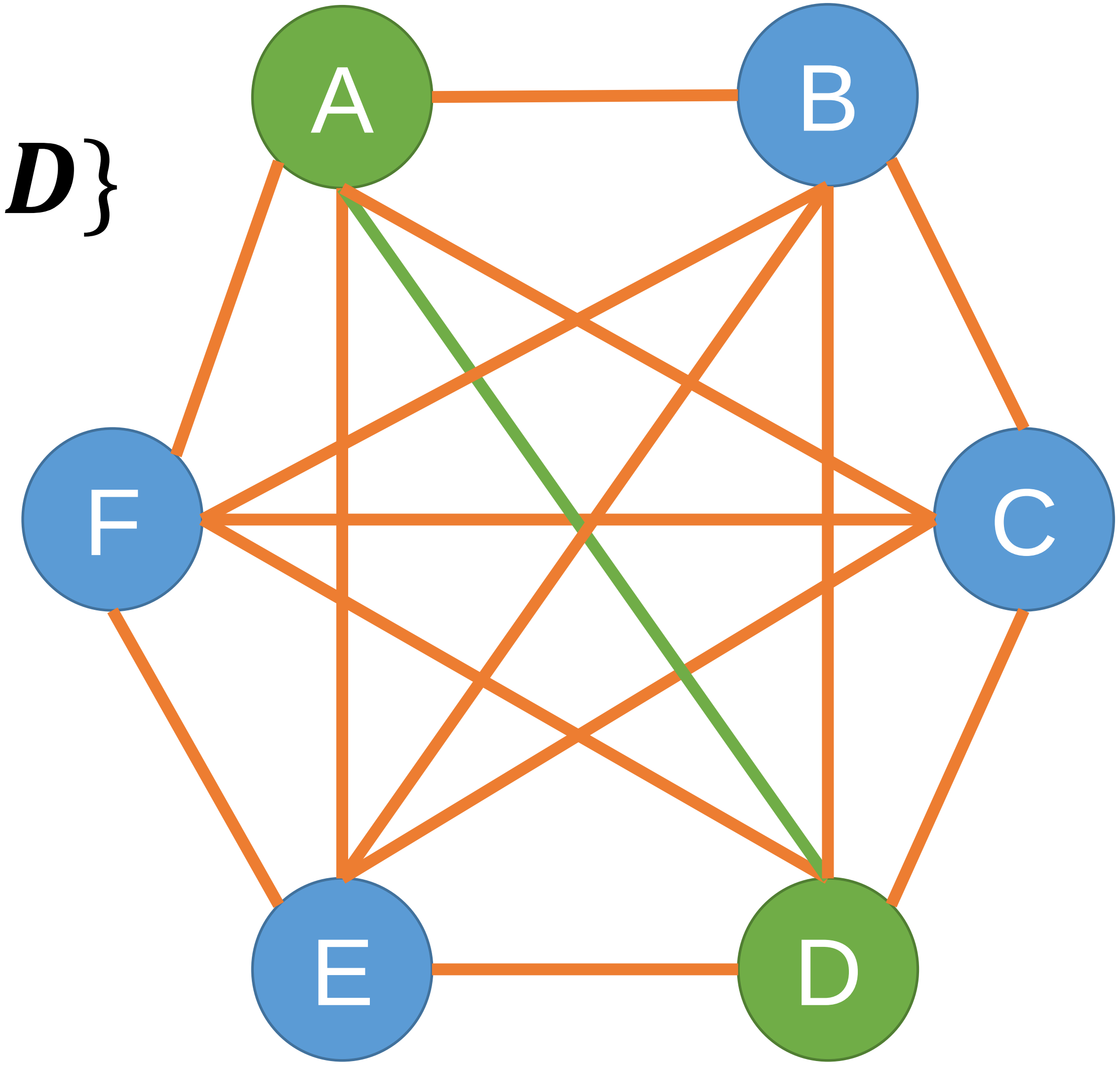


Approach: DRMI

Greedy-choice Initialization

$$S_0 = \{A, D\}$$

MI	A	B	C	D	E	F
A	/	0.88	0.57	**0.52**	0.58	0.56
B	0.88	/	0.50	0.50	0.57	0.63
C	0.57	0.50	/	0.91	0.45	0.43
D	0.52	0.50	0.91	/	0.43	0.36
E	0.58	0.57	0.45	0.43	/	0.98
F	0.56	0.63	0.43	0.36	0.98	/

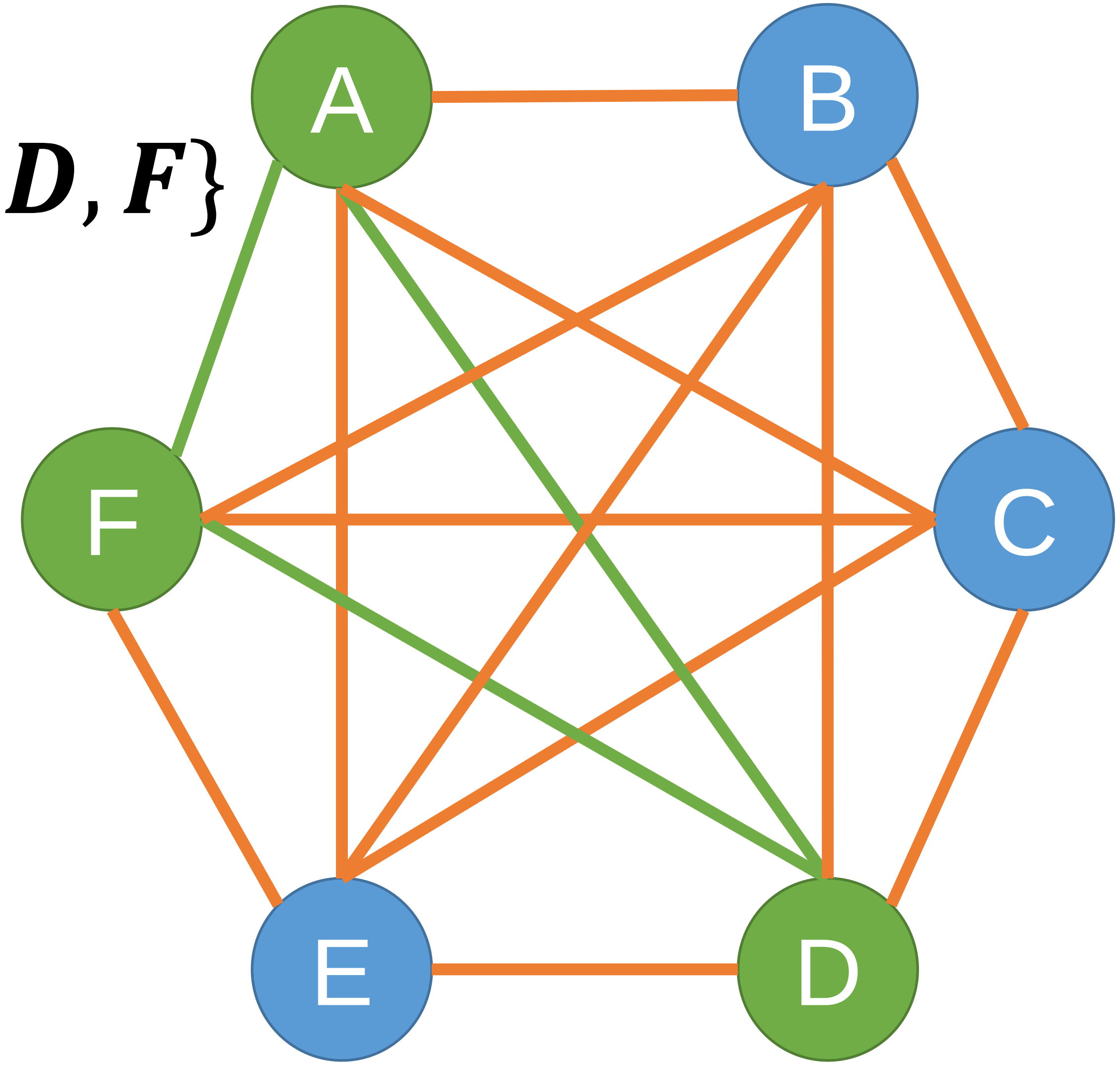


Approach: DRMI

Greedy-choice Initialization

$$S_0 = \{A, D, F\}$$

MI	A	B	C	D	E	F
A	/	0.88	0.57	**0.52**	0.58	0.56
B	0.88	/	0.50	0.50	0.57	0.63
C	0.57	0.50	/	0.91	0.45	0.43
D	0.52	0.50	0.91	/	0.43	0.36
E	0.58	0.57	0.45	0.43	/	0.98
F	**0.56**	0.63	0.43	**0.36**	0.98	/



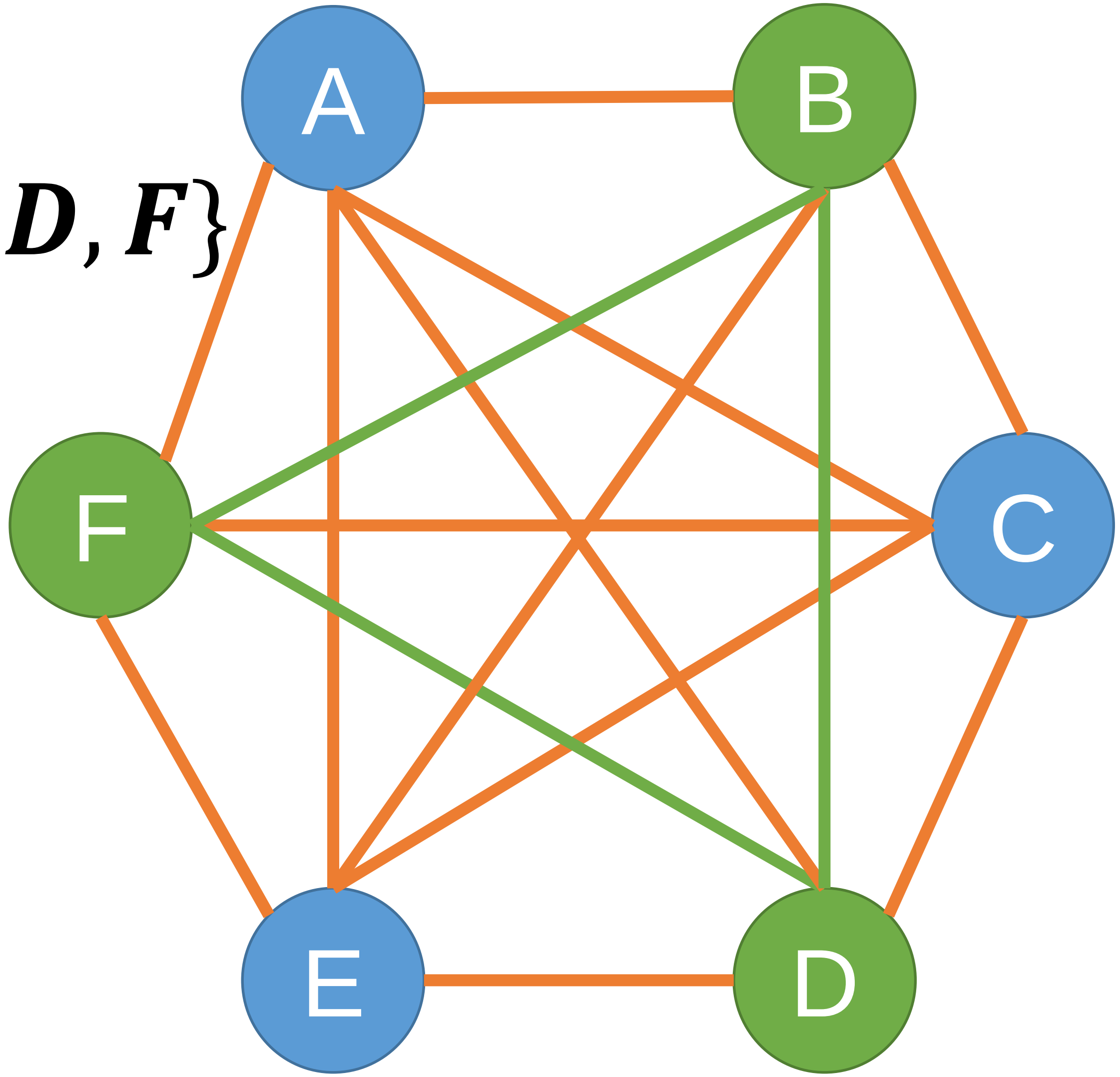
Approach: DRMI

Greedy-choice Initialization

$$t = B$$

$$S_0 = \{B, D, F\}$$

MI	A	B	C	D	E	F
A	/	0.88	0.57	0.52	0.58	0.56
B	0.88	/	0.50	**0.50**	0.57	**0.63**
C	0.57	0.50	/	0.91	0.45	0.43
D	0.52	0.50	0.91	/	0.43	**0.36**
E	0.58	0.57	0.45	0.43	/	0.98
F	0.56	0.63	0.43	0.36	0.98	/



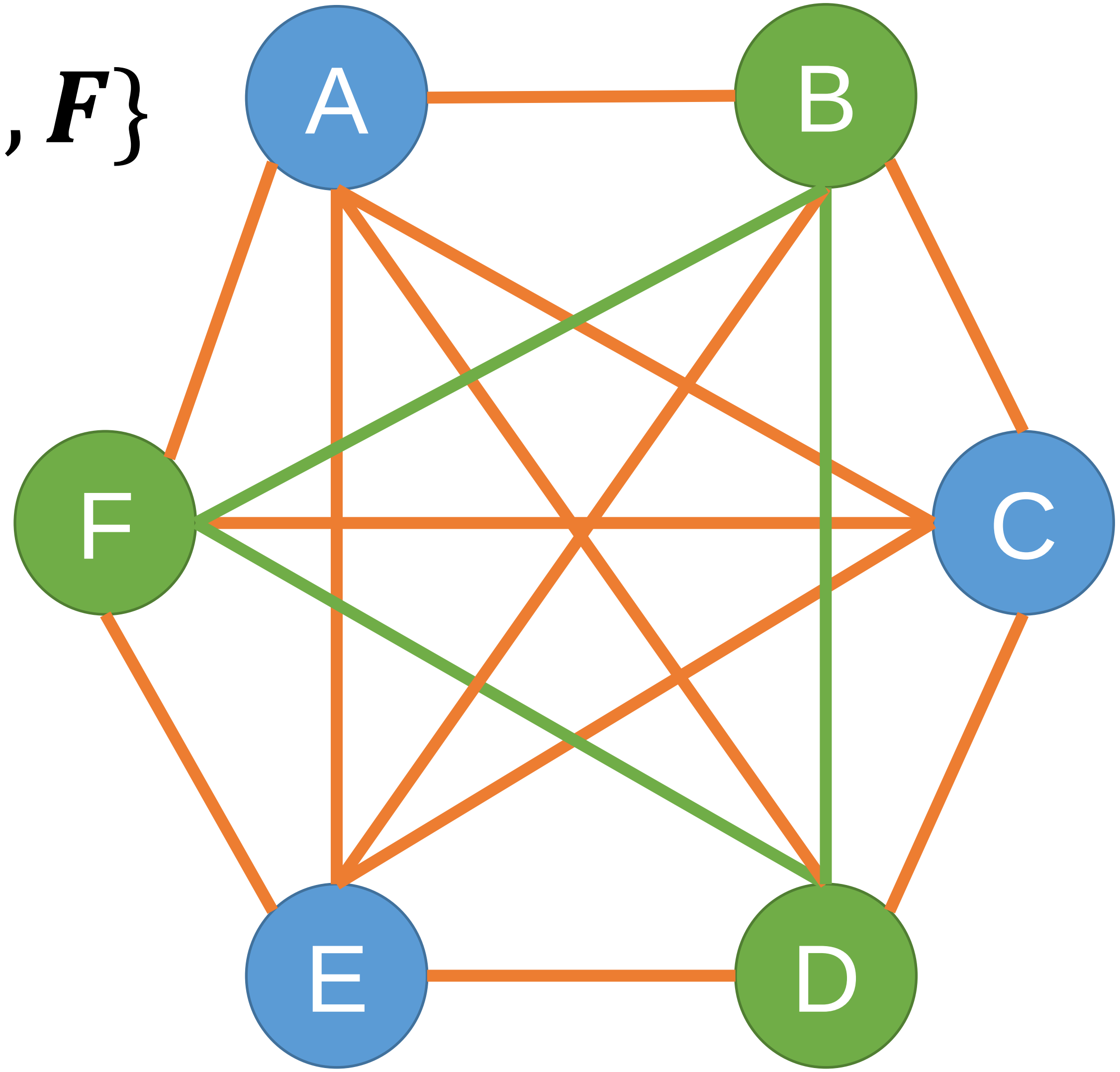
Approach: DRMI

One-hot Replacement Optimization

$$S_0 = \{B, D, F\}$$

$$S = S_0$$

MI	A	B	C	D	E	F
A	/	0.88	0.57	0.52	0.58	0.56
B	0.88	/	0.50	**0.50**	0.57	**0.63**
C	0.57	0.50	/	0.91	0.45	0.43
D	0.52	0.50	0.91	/	0.43	**0.36**
E	0.58	0.57	0.45	0.43	/	0.98
F	0.56	0.63	0.43	0.36	0.98	/

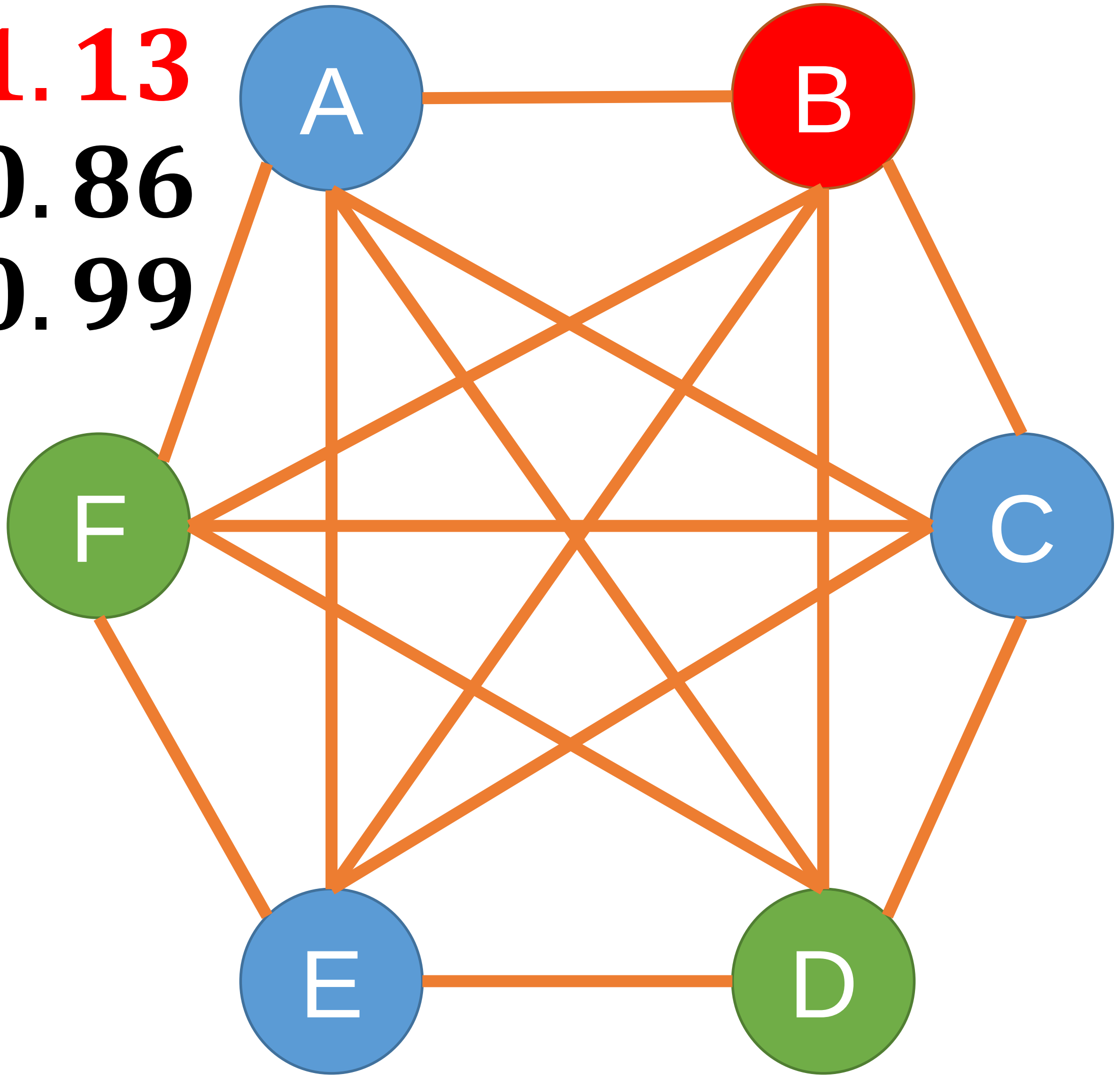


Approach: DRMI

One-hot Replacement Optimization

$$\boldsymbol{B} \rightarrow \boldsymbol{D}, \boldsymbol{F}: \mathbf{1.13}$$
$$\boldsymbol{D} \rightarrow \boldsymbol{B}, \boldsymbol{F}: \mathbf{0.86}$$
$$\boldsymbol{F} \rightarrow \boldsymbol{B}, \boldsymbol{D}: \mathbf{0.99}$$

MI	A	B	C	D	E	F
A	/	0.88	0.57	0.52	0.58	0.56
B	0.88	/	0.50	**0.50**	0.57	**0.63**
C	0.57	0.50	/	0.91	0.45	0.43
D	0.52	0.50	0.91	/	0.43	**0.36**
E	0.58	0.57	0.45	0.43	/	0.98
F	0.56	0.63	0.43	0.36	0.98	/

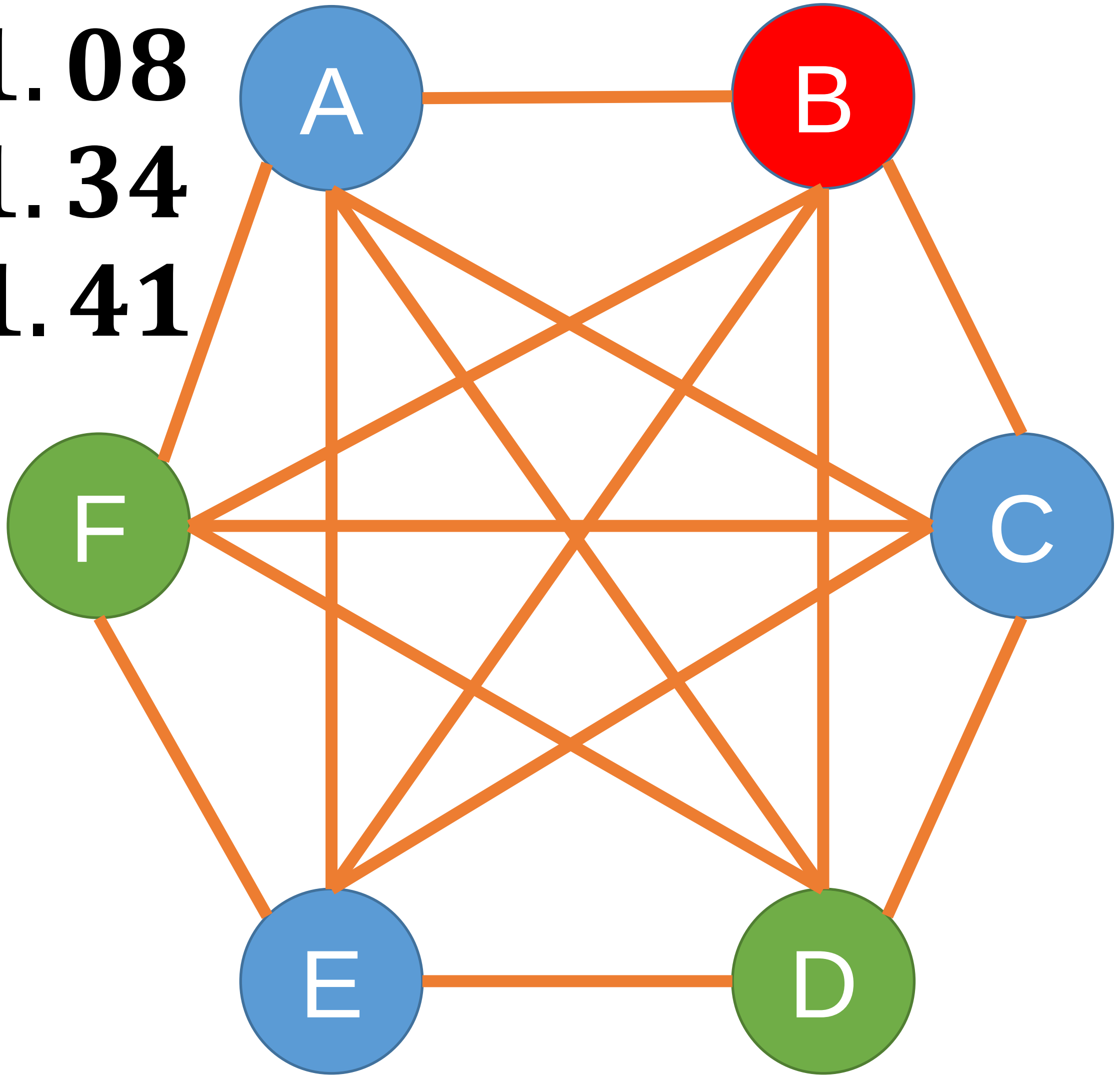


Approach: DRMI

One-hot Replacement Optimization

$A \rightarrow D, F: 1.08$
$C \rightarrow D, F: 1.34$
$E \rightarrow D, F: 1.41$

MI	A	B	C	D	E	F
A	/	0.88	0.57	0.52	0.58	0.56
B	0.88	/	0.50	0.50	0.57	0.63
C	0.57	0.50	/	0.91	0.45	0.43
D	0.52	0.50	0.91	/	0.43	0.36
E	0.58	0.57	0.45	0.43	/	0.98
F	0.56	0.63	0.43	0.36	0.98	/



Approach: DRMI

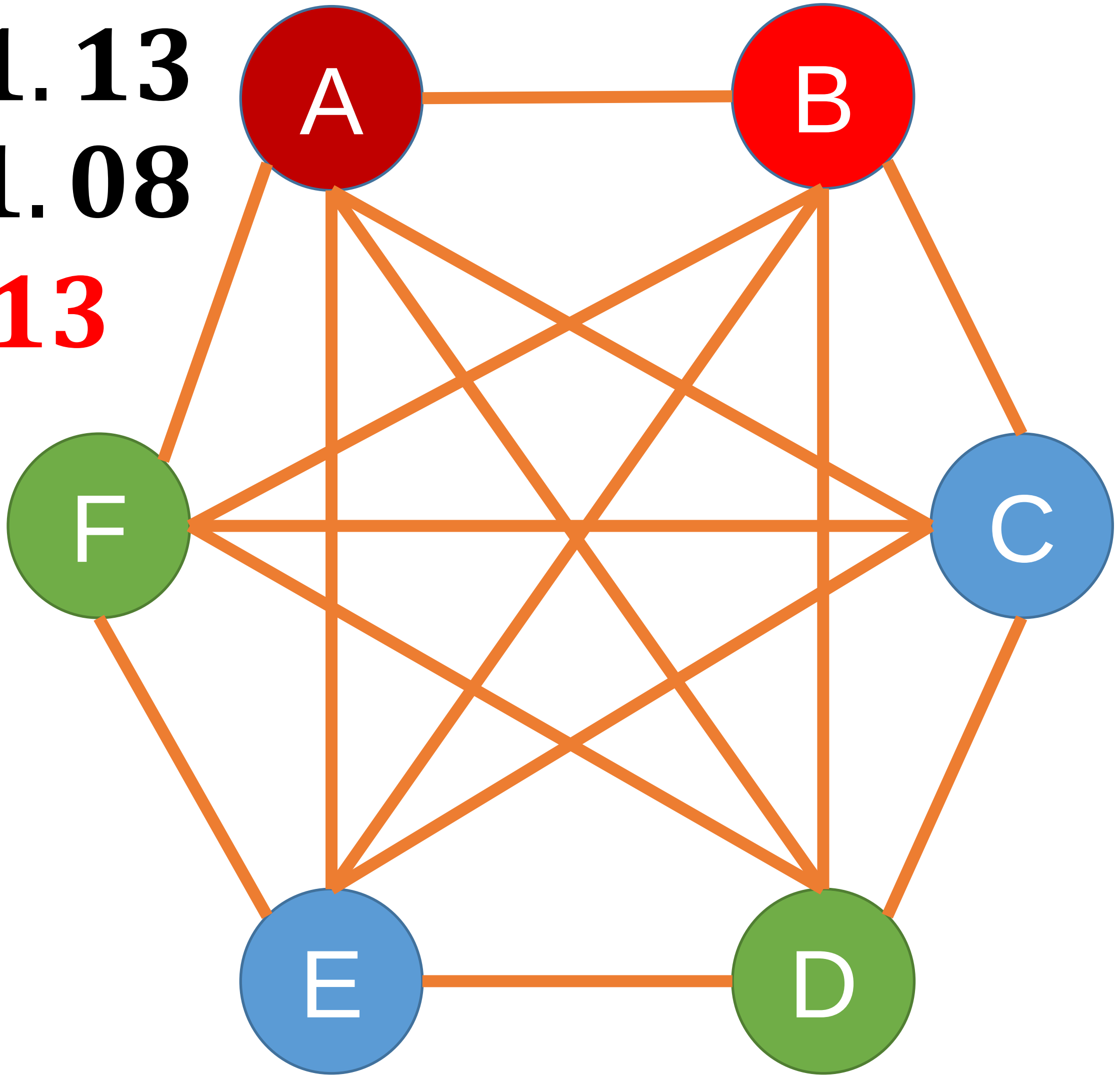
One-hot Replacement Optimization

$\boldsymbol{B \rightarrow D, F: 1.13}$

$\boldsymbol{A \rightarrow D, F: 1.08}$

$\boldsymbol{1.08 < 1.13}$

MI	A	B	C	D	E	F
A	/	0.88	0.57	**0.52**	0.58	**0.56**
B	0.88	/	0.50	**0.50**	0.57	**0.63**
C	0.57	0.50	/	0.91	0.45	0.43
D	0.52	0.50	0.91	/	0.43	0.36
E	0.58	0.57	0.45	0.43	/	0.98
F	0.56	0.63	0.43	0.36	0.98	/



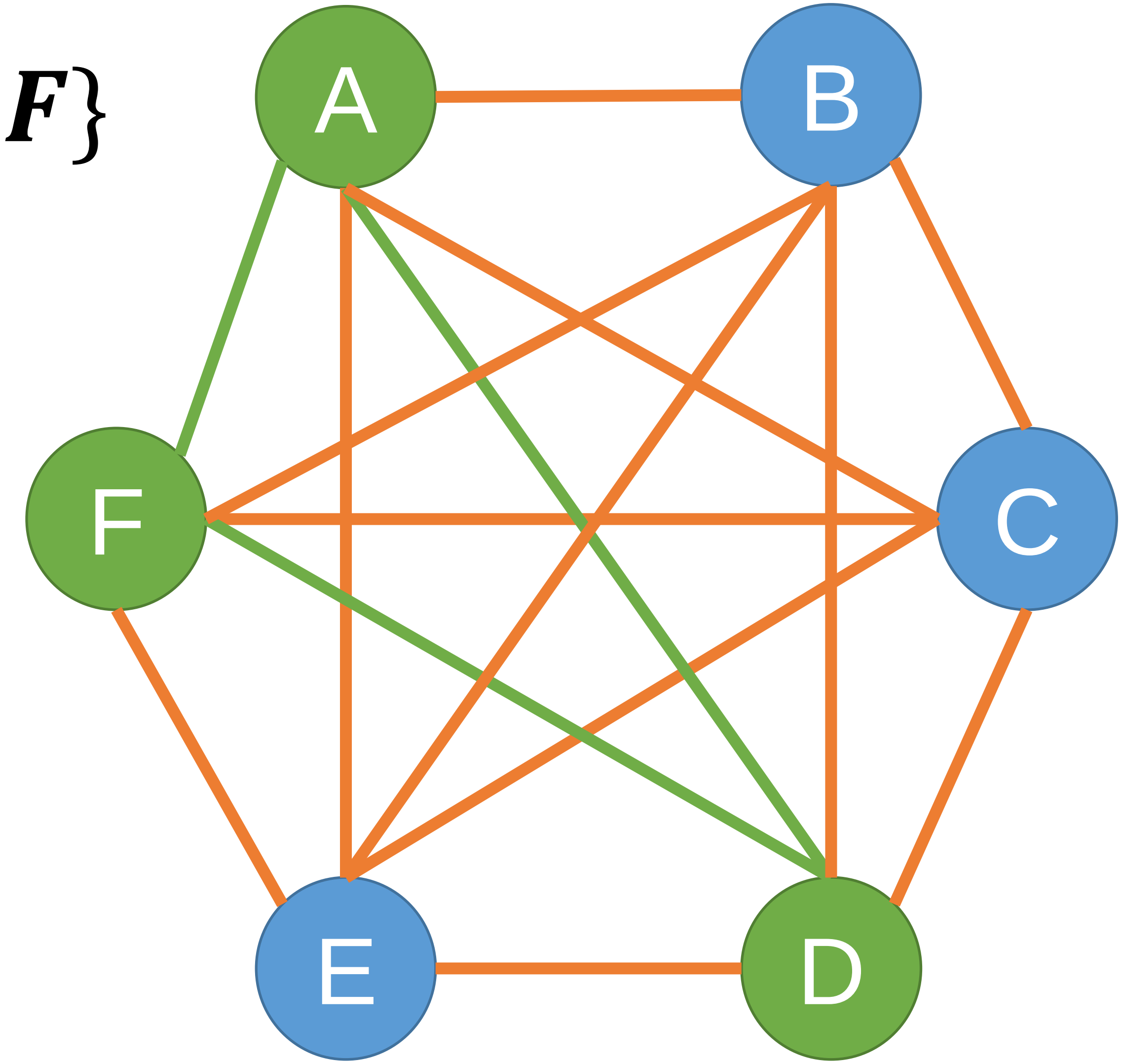
Approach: DRMI

One-hot Replacement Optimization

$$\boldsymbol{S} = \{\boldsymbol{A}, \boldsymbol{D}, \boldsymbol{F}\}$$

$$\boldsymbol{H} = \mathbf{1.44}$$

MI	A	B	C	D	E	F
A	/	0.88	0.57	**0.52**	0.58	**0.56**
B	0.88	/	0.50	0.50	0.57	0.63
C	0.57	0.50	/	0.91	0.45	0.43
D	0.52	0.50	0.91	/	0.43	**0.36**
E	0.58	0.57	0.45	0.43	/	0.98
F	0.56	0.63	0.43	0.36	0.98	/



Evaluation

Different model architecture

Method	architecture	Q=600	Q=300	Q=150
DRMI	LeNet-5	96.38%	94.29%	92.13%
	C3F2	97.25%	94.41%	91.12%
Baseline		91.91%	88.48%	84.97%

Accuracy on substitute model

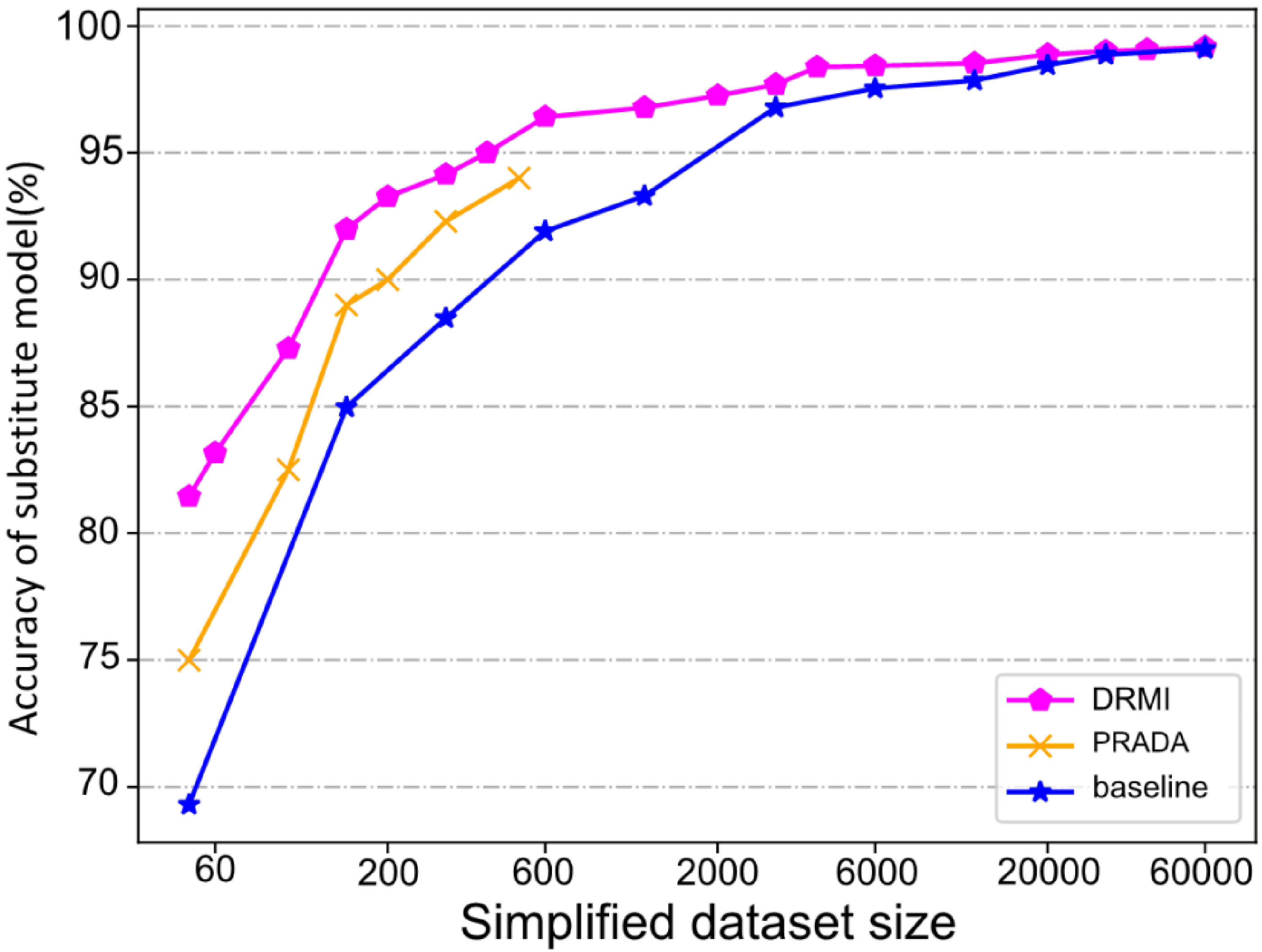
Evaluation

Different distribution of the target dataset

Test Acc.		Query 600	Query 300	Query 150
MNIST	5,000	94.83%	92.40%	90.51%
USPS	5,000	93.36%	91.88%	89.57%

Evaluation

Accuracy of substitute model



Evaluation

Black-box attack

Queries	Target model	Transferability	Accuracy
150	LeNet-5	68.32%	92.13%
	PRADA	29%	89%
300	LeNet-5	69.80%	94.34%
	PRADA	39%	91%

Conclusion

- Reduce the number of queries and high accuracy
- Black-box attacks based on substitute model
- Measurement of the quality of queries

Thanks for listening!

Q&A

heyingzhe@iie.ac.cn