NEURAL INFORMATION
PROCESSING SYSTEMS

Boosting the Uniqueness of Neural Networks Fingerprints with Informative Triggers

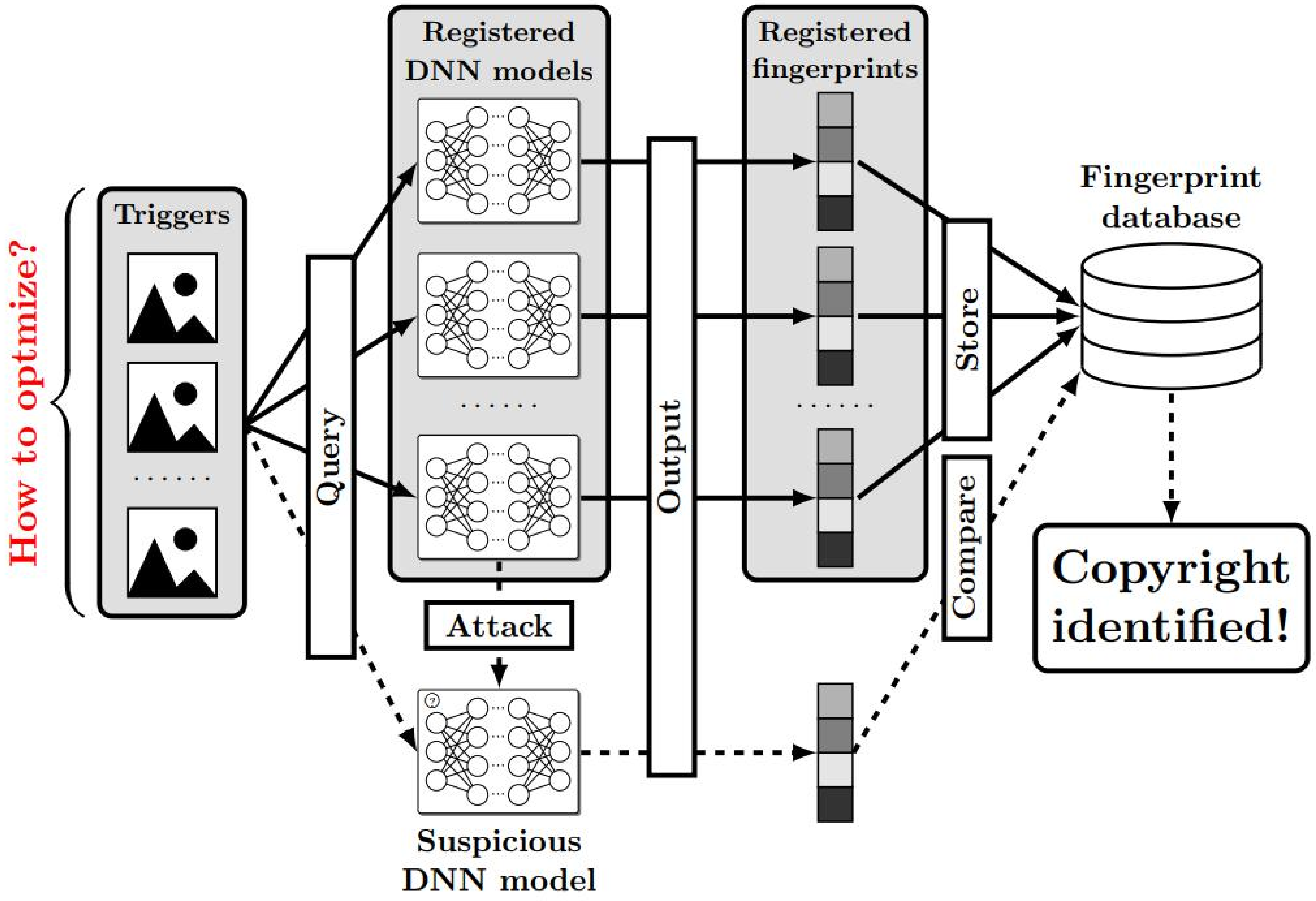
Zhuomeng Zhang, Fangqi Li, Hanyi Wang, Shi-lin Wang

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Introduction

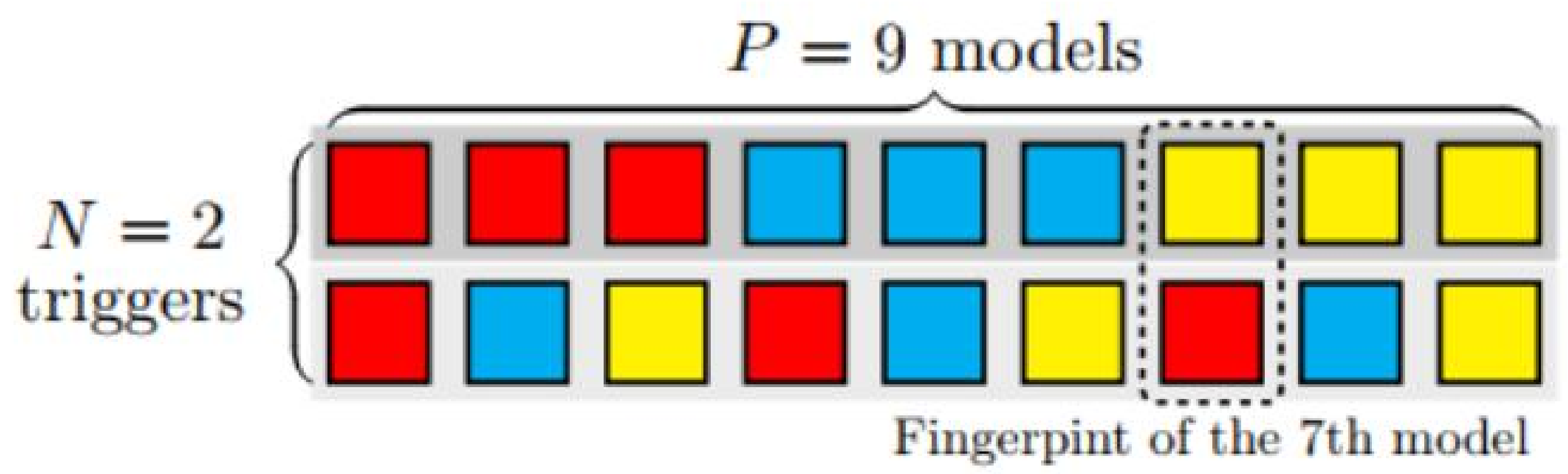
Contributions:

- We adopt an information theoretical perspective to measure the contribution to copyright tracing of each fingerprinting trigger by its conditional mutual information.
- We boost the copyright tracing performance by greedily optimizing the collection of triggers and validate this method through extensive experiments.
- We derive the first necessary condition for the number of fingerprinting triggers to ensure copyright tracing.

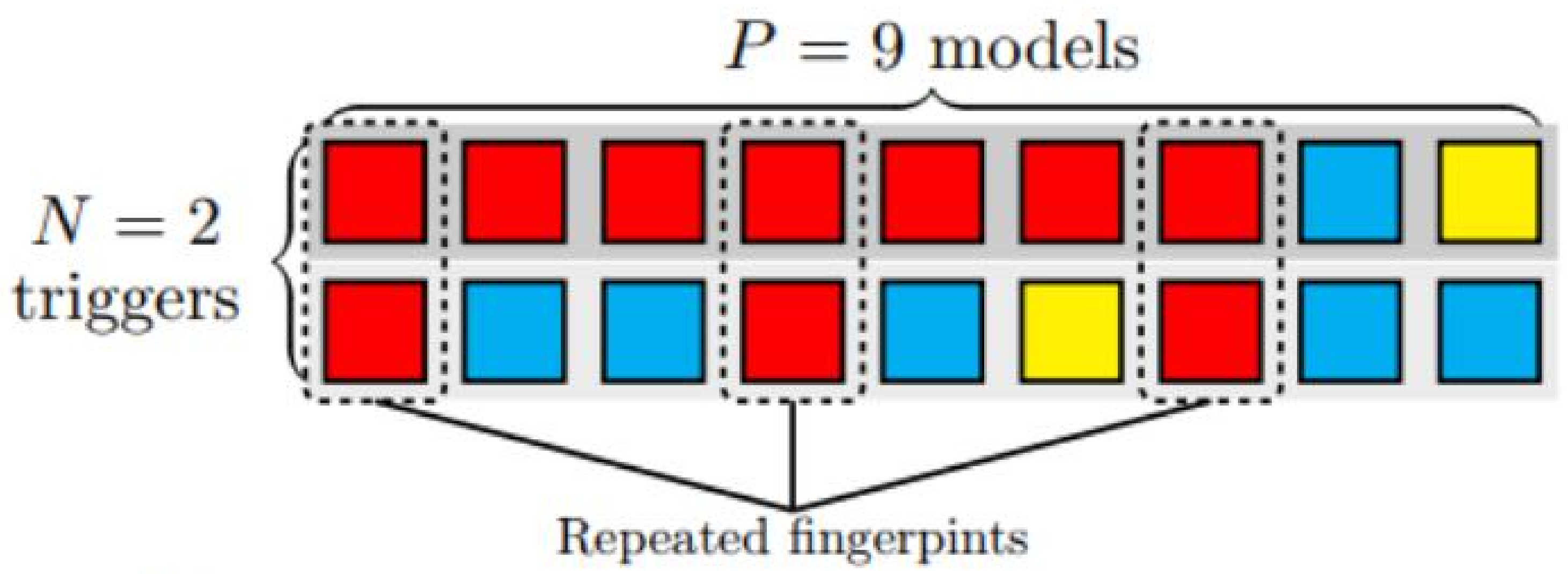


The framework of a fingerprint-based DNN copyright tracing system

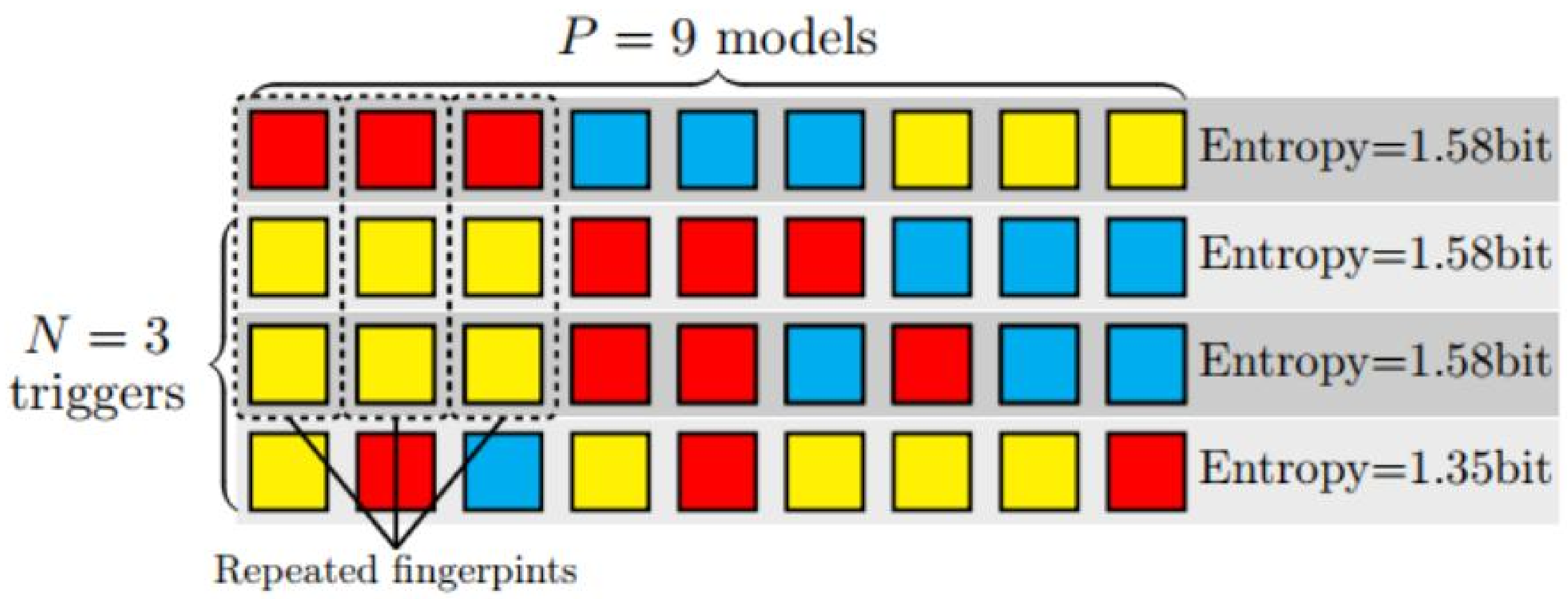
Introduction



(a) The idealistic setting.



(b) Triggers with biased outputs.



(c) Triggers with the largest entropy fail uniqueness.

Challenges:

- The number of triggers is not arbitrarily large. It is necessary to evaluate the performance of DNN fingerprints when the number of triggers is limited.
- The value of each trigger has been overestimated. DNN models' outputs on a trigger might not be uniformly distributed, as shown in (b). Moreover, the triggers are not independent of each other. Using triggers with independently the largest entropy might turn out to be misleading as shown in (c).
- The influence of adversarial modifications is unclear.

Information in DNN Copyright Tracing

Table 1: Frequently used notations in this paper.

Symbol	Meaning
$\mathcal{X}$	Input space of DNN models.
C	Number of classes.
$\mathcal{Y}$	Output space of DNN models.
P	Number of registered DNN models.
$\mathbf{F}$	Registered DNN models, $\mathbf{F} = \{f_p\}_{p=1}^{P}$.
N	Number of fingerprinting triggers.
$\hat{N}$	Number of greedily selected triggers.
$\mathcal{T}$	Distribution of triggers.
$\mathbf{T}$	Fingerprinting triggers, $\mathbf{T} = \{\mathbf{t}_n\}_{n=1}^{N}$.
ϵ	Verifier's tolerance on the adversary's attack.
$\mathbf{A}_\epsilon$	Verifier's threat model.
$\mathcal{F}$	A randomly selected model from $\mathbf{F}$.
$\mathcal{A}$	A randomly selected attack from $\mathbf{A}_\epsilon$.
ϕ_n	The randomly selected model's prediction on $\mathbf{t}_n$.
u	Uniqueness rate
$I_\epsilon\left(\mathbf{t}_n \vert \mathbf{t}_{1:(n-1)}\right)$	Conditional mutual information of $\mathbf{t}_n$.

$$u = 1 - |\mathbf{F}_\epsilon(\mathbf{T})|/|\mathbf{F}|.$$

$$u = \frac{2^{\log_2 uP}}{P} \leq \frac{2^{-u\log_2 u-(1-u)\log_2(1-u)+\log_2 uP}}{P} \leq \frac{2^{I(\Phi;\mathcal{F})}}{P}.$$

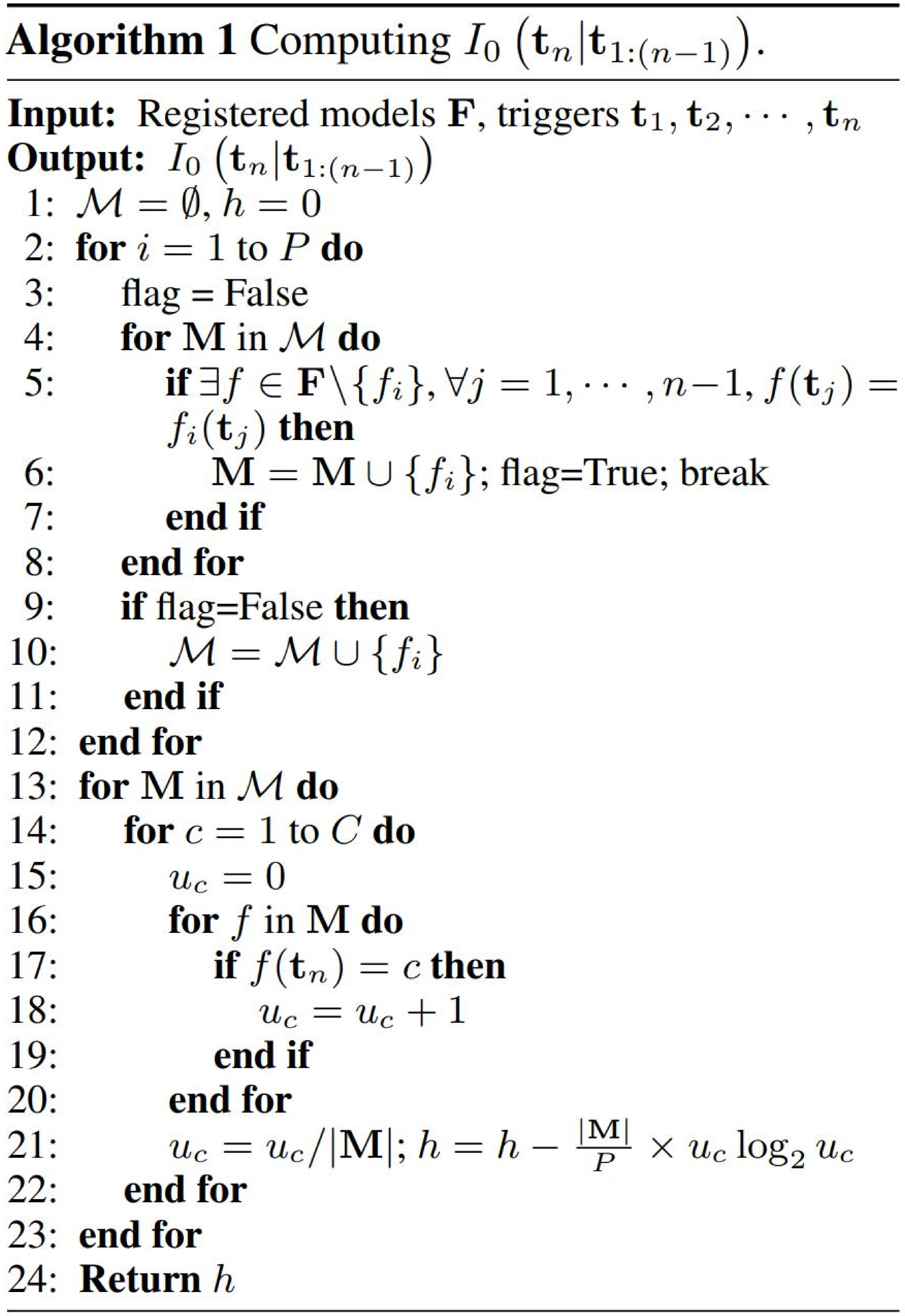
$$I_\epsilon\left(\mathbf{t}_n \vert \mathbf{t}_{1:(n-1)}\right) = H(\phi_n|\phi_1,\cdots,\phi_{n-1}) - H(\phi_n|\phi_1,\cdots,\phi_{n-1},\mathcal{F}).$$

$$I(\Phi;\mathcal{F}) = H(\Phi) - H(\Phi|\mathcal{F}) = \sum_{n=1}^{N} I_\epsilon\left(\mathbf{t}_n \vert \mathbf{t}_{1:(n-1)}\right).$$

$$I_0\left(\mathbf{t}_n \vert \mathbf{t}_{1:(n-1)}\right) = H(\phi_n|\phi_1,\cdots,\phi_{n-1}).$$

Information in DNN Copyright Tracing

$$O(C^n) \rightarrow O(max\{P^2n, PC\})$$



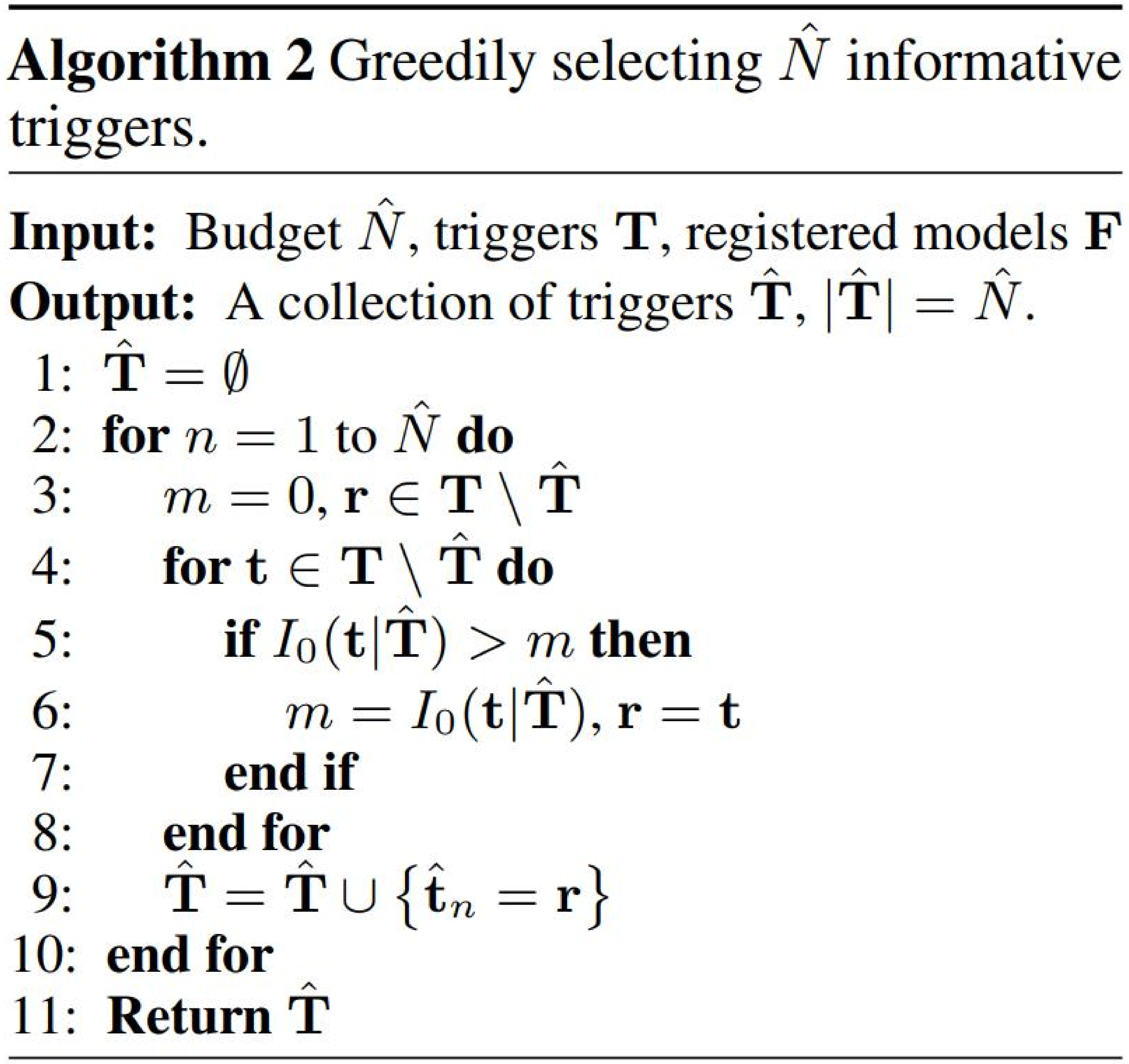
Algorithm 1 Computing $I_0\left(\mathbf{t}_n|\mathbf{t}_{1:(n-1)}\right)$.

Input: Registered models $\mathbf{F}$, triggers $\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_n$

Output: $I_0\left(\mathbf{t}_n|\mathbf{t}_{1:(n-1)}\right)$

```
$\mathcal{M} = \emptyset$, $h = 0$
for $i = 1$ to $P$ do
    flag = False
    for M in $\mathcal{M}$ do
        if $\exists f \in \mathbf{F} \setminus \{f_i\}, \forall j = 1, \cdots, n-1, f(\mathbf{t}_j) = f_i(\mathbf{t}_j)$ then
            $\mathbf{M} = \mathbf{M} \cup \{f_i\}$; flag=True; break
        end if
    end for
    if flag=False then
        $\mathcal{M} = \mathcal{M} \cup \{f_i\}$
    end if
end for
for M in $\mathcal{M}$ do
    for $c = 1$ to $C$ do
        $u_c = 0$
        for $f$ in M do
            if $f(\mathbf{t}_n) = c$ then
                $u_c = u_c + 1$
            end if
        end for
        $u_c = u_c/|\mathbf{M}|$; $h = h - \frac{|\mathbf{M}|}{P} \times u_c \log_2 u_c$
    end for
end for
Return $h$
```

Greedy Optimization of Triggers



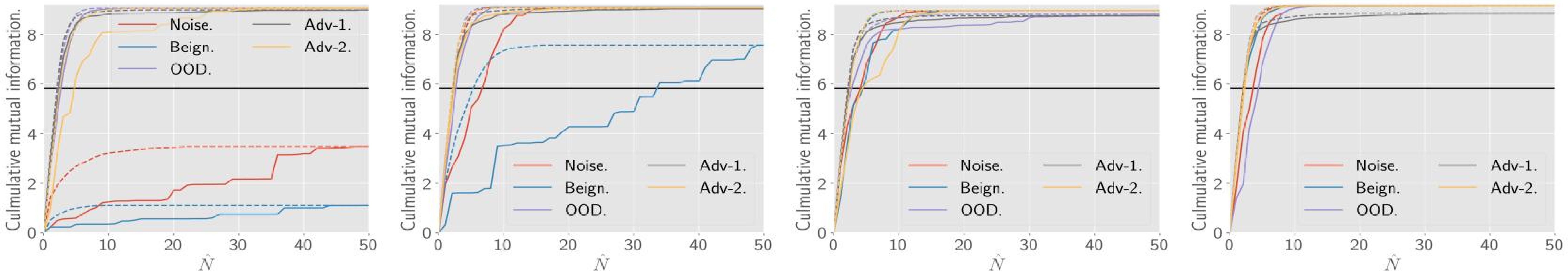
Algorithm 2 Greedily selecting $\hat{N}$ informative triggers.

```
Input: Budget $\hat{N}$, triggers $\mathbf{T}$, registered models $\mathbf{F}$
Output: A collection of triggers $\hat{\mathbf{T}}$, $|\hat{\mathbf{T}}| = \hat{N}$.
$\hat{\mathbf{T}} = \emptyset$
for $n = 1$ to $\hat{N}$ do
    $m = 0$, $\mathbf{r} \in \mathbf{T} \setminus \hat{\mathbf{T}}$
    for $\mathbf{t} \in \mathbf{T} \setminus \hat{\mathbf{T}}$ do
        if $I_0(\mathbf{t}|\hat{\mathbf{T}}) > m$ then
            $m = I_0(\mathbf{t}|\hat{\mathbf{T}})$, $\mathbf{r} = \mathbf{t}$
        end if
    end for
    $\hat{\mathbf{T}} = \hat{\mathbf{T}} \cup \{\hat{\mathbf{t}}_n = \mathbf{r}\}$
end for
Return $\hat{\mathbf{T}}$
```

$$\sum_{n=1}^{\hat{N}} \underline{I}_{\epsilon}\left(\hat{\mathbf{t}}_n | \hat{\mathbf{t}}_{1:(n-1)}\right) \geq \left(1 - \frac{1}{e}\right) \sum_{n=1}^{\hat{N}} \underline{I}_{\epsilon}\left(\tilde{\mathbf{t}}_n | \tilde{\mathbf{t}}_{1:(n-1)}\right).$$

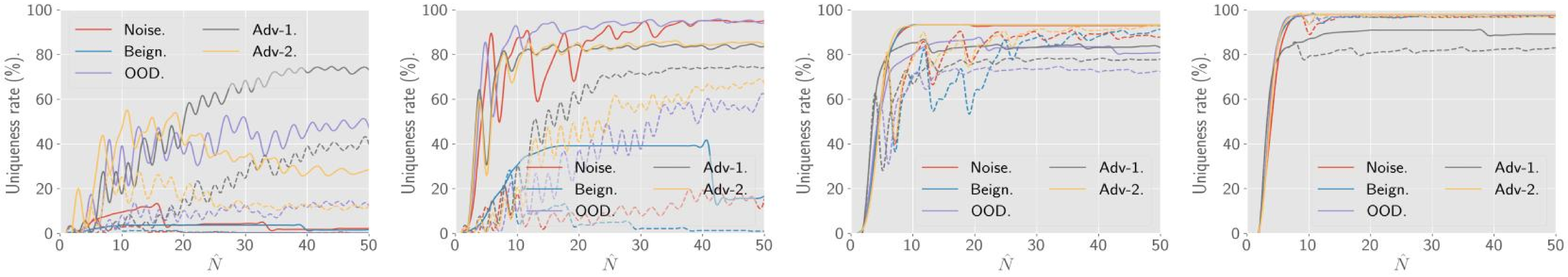
$$N \geq \min\left\{\hat{N} : \sum_{n=1}^{\hat{N}} \underline{I}_{\epsilon}\left(\hat{\mathbf{t}}_n | \hat{\mathbf{t}}_{1:(n-1)}\right) \geq \left(1 - \frac{1}{e}\right) \log_2 P\right\}.$$

Experimental Results



(a) MNIST. (b) FashionMNIST. (c) CIFAR-10. (d) ImageNet.

Figure 4: The cumulative mutual information (in bit) provided by triggers in the original order (the solid curves) and triggers selected by the greedy algorithm (the dashed curves) when $\epsilon = 0$. The black line marks $\left(1 - \frac{1}{e}\right) \log_2 600$ bits.

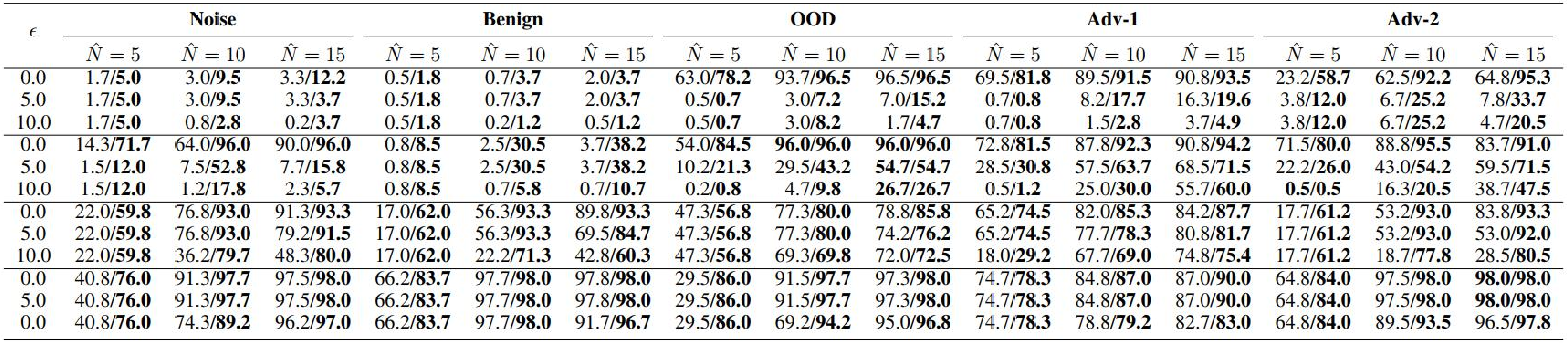


(a) MNIST. (b) FashionMNIST. (c) CIFAR-10. (d) ImageNet.

Figure 5: Uniqueness rate (%) provided by greedily selected informative triggers when $\epsilon = 2.5\%$ (the solid curves) and $\epsilon = 10.0\%$ (the dashed curves).

Experimental Results

Table 2: Uniqueness rate of registered models (%). For A/B in each entry, A is the uniqueness rate provided by the original order of triggers, and B is the uniqueness rate provided by greedily selected triggers. The dataset is MNIST, FashionMNIST, CIFAR-10, and ImageNet from top to bottom. The better scheme in each setting is highlighted in bold.



ϵ	**Noise**			**Benign**			**OOD**			**Adv-1**			**Adv-2**		
	$\hat{N}=5$	$\hat{N}=10$	$\hat{N}=15$	$\hat{N}=5$	$\hat{N}=10$	$\hat{N}=15$	$\hat{N}=5$	$\hat{N}=10$	$\hat{N}=15$	$\hat{N}=5$	$\hat{N}=10$	$\hat{N}=15$	$\hat{N}=5$	$\hat{N}=10$	$\hat{N}=15$
0.0	1.7/**5.0**	3.0/**9.5**	3.3/**12.2**	0.5/**1.8**	0.7/**3.7**	2.0/**3.7**	63.0/**78.2**	93.7/**96.5**	96.5/**96.5**	69.5/**81.8**	89.5/**91.5**	90.8/**93.5**	23.2/**58.7**	62.5/**92.2**	64.8/**95.3**
5.0	1.7/**5.0**	3.0/**9.5**	3.3/**3.7**	0.5/**1.8**	0.7/**3.7**	2.0/**3.7**	0.5/**0.7**	3.0/**7.2**	7.0/**15.2**	0.7/**0.8**	8.2/**17.7**	16.3/**19.6**	3.8/**12.0**	6.7/**25.2**	7.8/**33.7**
10.0	1.7/**5.0**	0.8/**2.8**	0.2/**3.7**	0.5/**1.8**	0.2/**1.2**	0.5/**1.2**	0.5/**0.7**	3.0/**8.2**	1.7/**4.7**	0.7/**0.8**	1.5/**2.8**	3.7/**4.9**	3.8/**12.0**	6.7/**25.2**	4.7/**20.5**
0.0	14.3/**71.7**	64.0/**96.0**	90.0/**96.0**	0.8/**8.5**	2.5/**30.5**	3.7/**38.2**	54.0/**84.5**	**96.0/96.0**	**96.0/96.0**	72.8/**81.5**	87.8/**92.3**	90.8/**94.2**	71.5/**80.0**	88.8/**95.5**	83.7/**91.0**
5.0	1.5/**12.0**	7.5/**52.8**	7.7/**15.8**	0.8/**8.5**	2.5/**30.5**	3.7/**38.2**	10.2/**21.3**	29.5/**43.2**	**54.7/54.7**	28.5/**30.8**	57.5/**63.7**	68.5/**71.5**	22.2/**26.0**	43.0/**54.2**	59.5/**71.5**
10.0	1.5/**12.0**	1.2/**17.8**	2.3/**5.7**	0.8/**8.5**	0.7/**5.8**	0.7/**10.7**	0.2/**0.8**	4.7/**9.8**	**26.7/26.7**	0.5/**1.2**	25.0/**30.0**	55.7/**60.0**	**0.5/0.5**	16.3/**20.5**	38.7/**47.5**
0.0	22.0/**59.8**	76.8/**93.0**	91.3/**93.3**	17.0/**62.0**	56.3/**93.3**	89.8/**93.3**	47.3/**56.8**	77.3/**80.0**	78.8/**85.8**	65.2/**74.5**	82.0/**85.3**	84.2/**87.7**	17.7/**61.2**	53.2/**93.0**	83.8/**93.3**
5.0	22.0/**59.8**	76.8/**93.0**	79.2/**91.5**	17.0/**62.0**	56.3/**93.3**	69.5/**84.7**	47.3/**56.8**	77.3/**80.0**	74.2/**76.2**	65.2/**74.5**	77.7/**78.3**	80.8/**81.7**	17.7/**61.2**	53.2/**93.0**	53.0/**92.0**
10.0	22.0/**59.8**	36.2/**79.7**	48.3/**80.0**	17.0/**62.0**	22.2/**71.3**	42.8/**60.3**	47.3/**56.8**	69.3/**69.8**	72.0/**72.5**	18.0/**29.2**	67.7/**69.0**	74.8/**75.4**	17.7/**61.2**	18.7/**77.8**	28.5/**80.5**
0.0	40.8/**76.0**	91.3/**97.7**	97.5/**98.0**	66.2/**83.7**	97.7/**98.0**	97.8/**98.0**	29.5/**86.0**	91.5/**97.7**	97.3/**98.0**	74.7/**78.3**	84.8/**87.0**	87.0/**90.0**	64.8/**84.0**	97.5/**98.0**	**98.0/98.0**
5.0	40.8/**76.0**	91.3/**97.7**	97.5/**98.0**	66.2/**83.7**	97.7/**98.0**	97.8/**98.0**	29.5/**86.0**	91.5/**97.7**	97.3/**98.0**	74.7/**78.3**	84.8/**87.0**	87.0/**90.0**	64.8/**84.0**	97.5/**98.0**	**98.0/98.0**
0.0	40.8/**76.0**	74.3/**89.2**	96.2/**97.0**	66.2/**83.7**	97.7/**98.0**	91.7/**96.7**	29.5/**86.0**	69.2/**94.2**	95.0/**96.8**	74.7/**78.3**	78.8/**79.2**	82.7/**83.0**	64.8/**84.0**	89.5/**93.5**	96.5/**97.8**

Experimental Results

Scalability to Online Copyright Tracing:

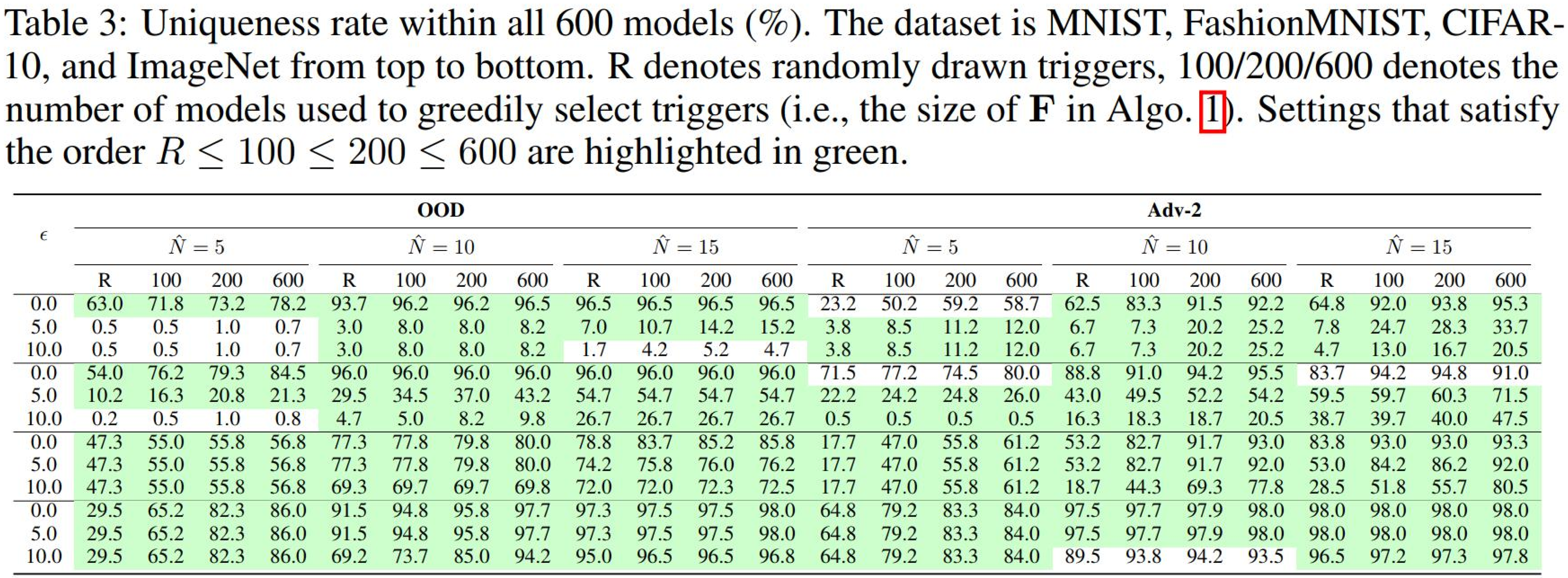


Table 3: Uniqueness rate within all 600 models (%). The dataset is MNIST, FashionMNIST, CIFAR-10, and ImageNet from top to bottom. R denotes randomly drawn triggers, 100/200/600 denotes the number of models used to greedily select triggers (i.e., the size of **F** in Algo. 1). Settings that satisfy the order $R \leq 100 \leq 200 \leq 600$ are highlighted in green.

ϵ	OOD												Adv-2											
	$\hat{N}=5$				$\hat{N}=10$				$\hat{N}=15$				$\hat{N}=5$				$\hat{N}=10$				$\hat{N}=15$			
	R	100	200	600	R	100	200	600	R	100	200	600	R	100	200	600	R	100	200	600	R	100	200	600
0.0	63.0	71.8	73.2	78.2	93.7	96.2	96.2	96.5	96.5	96.5	96.5	96.5	23.2	50.2	59.2	58.7	62.5	83.3	91.5	92.2	64.8	92.0	93.8	95.3
5.0	0.5	0.5	1.0	0.7	3.0	8.0	8.0	8.2	7.0	10.7	14.2	15.2	3.8	8.5	11.2	12.0	6.7	7.3	20.2	25.2	7.8	24.7	28.3	33.7
10.0	0.5	0.5	1.0	0.7	3.0	8.0	8.0	8.2	1.7	4.2	5.2	4.7	3.8	8.5	11.2	12.0	6.7	7.3	20.2	25.2	4.7	13.0	16.7	20.5
0.0	54.0	76.2	79.3	84.5	96.0	96.0	96.0	96.0	96.0	96.0	96.0	96.0	71.5	77.2	74.5	80.0	88.8	91.0	94.2	95.5	83.7	94.2	94.8	91.0
5.0	10.2	16.3	20.8	21.3	29.5	34.5	37.0	43.2	54.7	54.7	54.7	54.7	22.2	24.2	24.8	26.0	43.0	49.5	52.2	54.2	59.5	59.7	60.3	71.5
10.0	0.2	0.5	1.0	0.8	4.7	5.0	8.2	9.8	26.7	26.7	26.7	26.7	0.5	0.5	0.5	0.5	16.3	18.3	18.7	20.5	38.7	39.7	40.0	47.5
0.0	47.3	55.0	55.8	56.8	77.3	77.8	79.8	80.0	78.8	83.7	85.2	85.8	17.7	47.0	55.8	61.2	53.2	82.7	91.7	93.0	83.8	93.0	93.0	93.3
5.0	47.3	55.0	55.8	56.8	77.3	77.8	79.8	80.0	74.2	75.8	76.0	76.2	17.7	47.0	55.8	61.2	53.2	82.7	91.7	92.0	53.0	84.2	86.2	92.0
10.0	47.3	55.0	55.8	56.8	69.3	69.7	69.7	69.8	72.0	72.0	72.3	72.5	17.7	47.0	55.8	61.2	18.7	44.3	69.3	77.8	28.5	51.8	55.7	80.5
0.0	29.5	65.2	82.3	86.0	91.5	94.8	95.8	97.7	97.3	97.5	97.5	98.0	64.8	79.2	83.3	84.0	97.5	97.7	97.9	98.0	98.0	98.0	98.0	98.0
5.0	29.5	65.2	82.3	86.0	91.5	94.8	95.8	97.7	97.3	97.5	97.5	98.0	64.8	79.2	83.3	84.0	97.5	97.7	97.9	98.0	98.0	98.0	98.0	98.0
10.0	29.5	65.2	82.3	86.0	69.2	73.7	85.0	94.2	95.0	96.5	96.5	96.8	64.8	79.2	83.3	84.0	89.5	93.8	94.2	93.5	96.5	97.2	97.3	97.8

Experimental Results

Overhead:

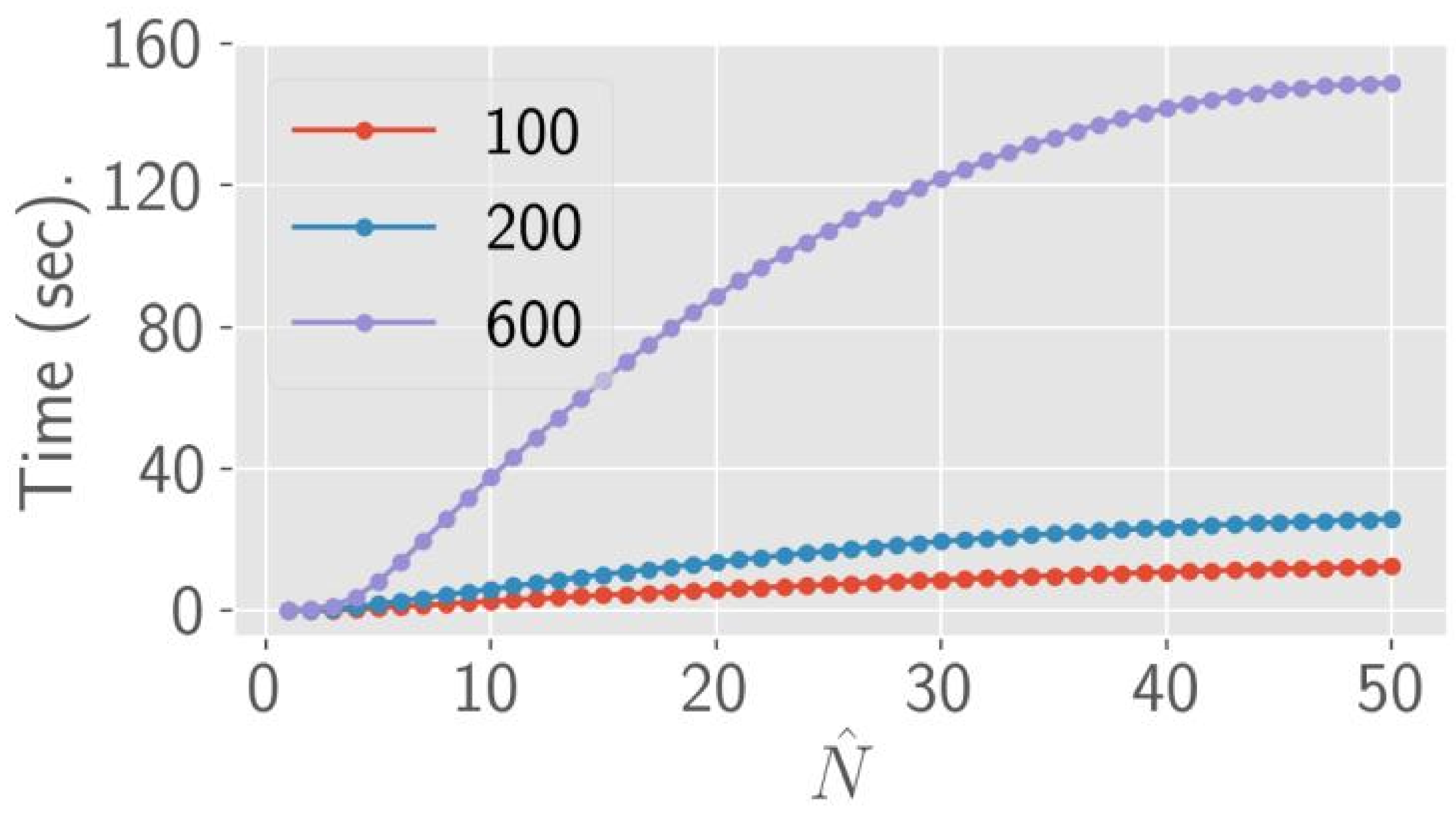


Figure 7: The time consumption of greedy trigger selection.

Thanks

Paper

Code