Stop, Don’t Click Here Anymore

Boosting Website Fingerprinting By Considering Sets of Webpages

Asya Mitseva and Andriy Panchenko

Chair of IT Security
Brandenburg University of Technology (BTU Cottbus, Germany)

August 15, 2024

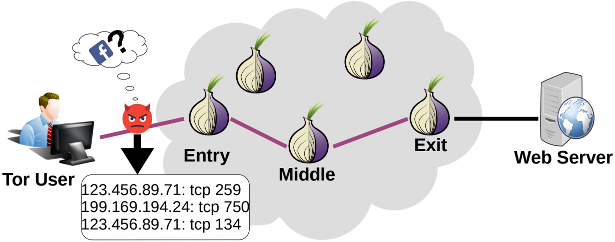
Brandenburg
University of Technology
Cottbus - Senftenberg

Motivation

- **The Tor network:** *most popular low-latency anonymization network*

Motivation

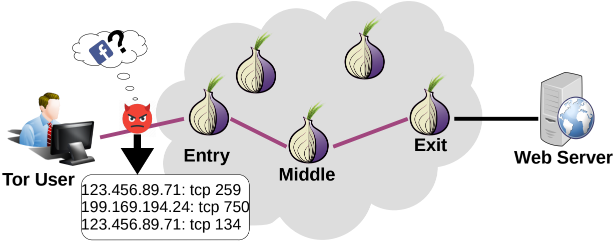
- **The Tor network:** *most popular low-latency anonymization network*
- *Problem:* **Tor is vulnerable to website fingerprinting (WFP)**
  - Infer website accessed without breaking the encryption
  - Using only packet sizes, direction, and timestamps
  - High efficiency in laboratory settings



```
123.456.89.71: tcp 259
199.169.194.24: tcp 750
123.456.89.71: tcp 134
```

Motivation

- **The Tor network:** *most popular low-latency anonymization network*
- *Problem:* **Tor is vulnerable to website fingerprinting (WFP)**
  - Infer website accessed without breaking the encryption
  - Using only packet sizes, direction, and timestamps
  - High efficiency in laboratory settings



- **Limitations**
  - *Scalability* in real-world settings still *under research*
  - Typical scenario: detection of *isolated webpage loads*

However,

users visit multiple pages of a website sequentially, e.g., by following links!

Our Contributions

- **Novel evaluation setting**
    - *Consecutive user's visits* of multiple pages of a website
    - *Detect the website*, not the webpage

Our Contributions

- **Novel evaluation setting**
  - *Consecutive user's visits* of multiple pages of a website
  - *Detect the website*, not the webpage

- **Systematic analysis of state-of-the-art webpage classifiers**
  - *Goal:* Investigate their suitability for *website* fingerprinting
  - 20 to 30% decrease in accuracy for website fingerprinting

Our Contributions

- **Novel evaluation setting**
  - *Consecutive user's visits* of multiple pages of a website
  - *Detect the website*, not the webpage
- **Systematic analysis of state-of-the-art webpage classifiers**
  - *Goal:* Investigate their suitability for *website* fingerprinting
  - 20 to 30% decrease in accuracy for website fingerprinting
- **Novel fingerprinting strategies for our new evaluation setting**
  - *Use of voting* to boost existing webpage classifiers
    - Six different voting-based fingerprinting strategies
  - *Set-aware classifier* based on multi-instance learning (MIL)

Our Contributions

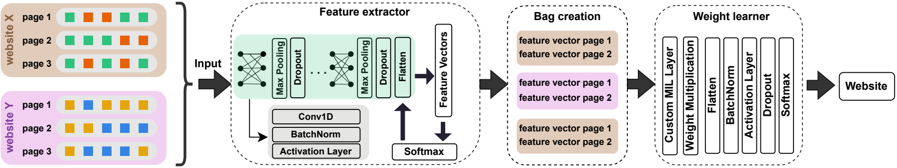
- **Novel evaluation setting**
  - *Consecutive user's visits* of multiple pages of a website
  - *Detect the website*, not the webpage
- **Systematic analysis of state-of-the-art webpage classifiers**
  - *Goal:* Investigate their suitability for *website* fingerprinting
  - 20 to 30% decrease in accuracy for website fingerprinting
- **Novel fingerprinting strategies for our new evaluation setting**
  - *Use of voting* to boost existing webpage classifiers
    - Six different voting-based fingerprinting strategies
  - *Set-aware classifier* based on multi-instance learning (MIL)
- **Limited protection provided by existing WFP defenses**
  - Up to five times less effective than expected or even completely useless

Our Novel Fingerprinting Techniques

- **Voting-based strategies**
  - Train state-of-the-art webpage classifier on multiple pages of websites
  - Compute probability of a single page to belong to a website
  - Calculate join probability of pages belonging to a website

Our Novel Fingerprinting Techniques

- **Voting-based strategies**
  - Train state-of-the-art webpage classifier on multiple pages of websites
  - Compute probability of a single page to belong to a website
  - Calculate join probability of pages belonging to a website
- **Set-aware MIL-based classifier**
  - A set of pages belonging to single website forms a *bag*
  - Learn a classification model to predict labels of bags
  - Adaptive learning of weights for single pages in a bag



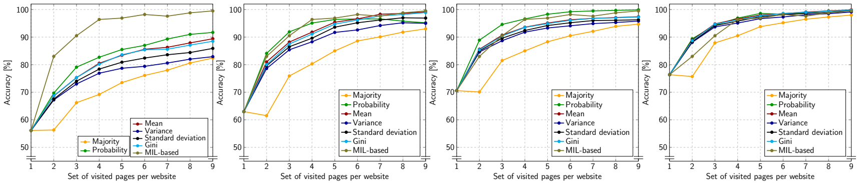
Analysis of Our Fingerprinting Techniques (1/3)

- **Dataset and evaluation setup**
  - *100 monitored websites* from different categories, layout, and content
  - *90 different pages* per website
  - Four state-of-the-art webpage classifiers
  - 10-fold cross-validation for all experiments

Analysis of Our Fingerprinting Techniques (1/3)

- **Dataset and evaluation setup**
  - *100 monitored websites* from different categories, layout, and content
  - *90 different pages* per website
  - Four state-of-the-art webpage classifiers
  - 10-fold cross-validation for all experiments

- **Evaluation in laboratory settings**
  - Users browse consecutively multiple pages of a single website
  - Increase of the detection rate *by almost 30% to 40%*



(a) CUMUL (b) k-FP (c) DF (d) Var-CNN

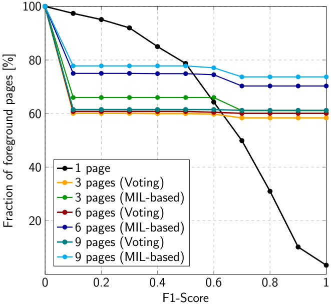
Analysis of Our Fingerprinting Techniques (2/3)

- **Impact of different training tactics**
  - *Our voting-based strategies*
    - 70 training pages are enough to obtain high accuracy
  - *Our set-aware MIL-based classifier*
    - Use of two training bags
    - Over 90% accuracy when four pages are consecutively visited
  - We refer to our paper for the extensive analysis of our methods

Analysis of Our Fingerprinting Techniques (2/3)

- **Impact of different training tactics**
  - *Our voting-based strategies*
    - 70 training pages are enough to obtain high accuracy
  - *Our set-aware MIL-based classifier*
    - Use of two training bags
    - Over 90% accuracy when four pages are consecutively visited
  - We refer to our paper for the extensive analysis of our methods

- **Real-world evaluation**
  - Use of 5000 unmonitored websites
  - Our methods achieve *F1-scores of 1.0* for *more than half* of the websites
  - When visiting *at least three consecutive pages* of a website



Analysis of Our Fingerprinting Techniques (3/3)

- **Increased robustness against existing WFP defenses**
  - Increase of the detection rate *up to 5 times*
  - No protection by defenses with low implementation costs

Defense	Classifier	Set of pages								
		1	2	3	4	5	6	7	8	9
Tamaraw	Voting	4.61	7.20	9.93	12.47	12.67	14.07	16.27	17.80	18.93
	MIL-based	–	5.37	7.25	8.81	10.79	12.19	13.77	14.73	16.36
CS-Buflo	Voting	**10.89**	18.13	23.33	33.27	37.40	43.93	46.93	52.47	**56.00**
	MIL-based	–	12.89	19.25	24.77	29.62	34.19	37.18	40.21	**43.17**
TrafficSliver-Net	Voting	19.92	29.93	34.48	38.73	40.45	42.79	43.80	44.85	46.55
	MIL-based	–	10.40	14.48	18.62	22.06	25.67	28.69	32.18	35.21
WTF-PAD	Voting	90.72	99.20	99.73	100.00	100.00	100.00	100.00	100.00	100.00
	MIL-based	–	98.28	99.61	99.89	99.99	99.99	99.99	100.00	100.00
RegulaTor	Voting	**17.17**	27.67	38.27	44.20	50.20	56.20	61.53	63.60	**64.87**
	MIL-based	–	16.11	22.83	27.77	31.89	36.19	40.29	43.44	46.48
FRONT	Voting	**67.00**	88.60	**96.87**	98.73	99.40	99.67	99.87	99.93	**100.00**
	MIL-based	–	86.41	**94.82**	97.70	98.85	99.38	99.55	99.77	**99.86**

Conclusion

- **Website fingerprinting is privacy threat for Tor users**
  - Especially those browsing multiple pages of website sequentially

Conclusion

- **Website fingerprinting is privacy threat for Tor users**
  - Especially those browsing multiple pages of website sequentially
- **Novel strategies using implicit knowledge on browsing behavior**
  - Six voting-based strategies to boost existing webpage classifiers
  - Novel set-aware classifier based on multi-instance learning
  - *Order of visiting* pages is *not necessary* for our methods

Conclusion

- **Website fingerprinting is privacy threat for Tor users**
    - Especially those browsing multiple pages of website sequentially
- **Novel strategies using implicit knowledge on browsing behavior**
    - Six voting-based strategies to boost existing webpage classifiers
    - Novel set-aware classifier based on multi-instance learning
    - *Order of visiting* pages is *not necessary* for our methods
- **More consecutive pages visited, higher detection rate**
    - Significant improvement of the detection rate in real-world settings
- **Significant reduction in protection by existing defenses**

Conclusion

- **Website fingerprinting is privacy threat for Tor users**
  - Especially those browsing multiple pages of website sequentially
- **Novel strategies using implicit knowledge on browsing behavior**
  - Six voting-based strategies to boost existing webpage classifiers
  - Novel set-aware classifier based on multi-instance learning
  - *Order of visiting* pages is *not necessary* for our methods
- **More consecutive pages visited, higher detection rate**
  - Significant improvement of the detection rate in real-world settings
- **Significant reduction in protection by existing defenses**

Stop, Don't Click Here Anymore: Boosting Website Fingerprinting By Considering Sets of Subpages

We are hiring! *See our open positions.*