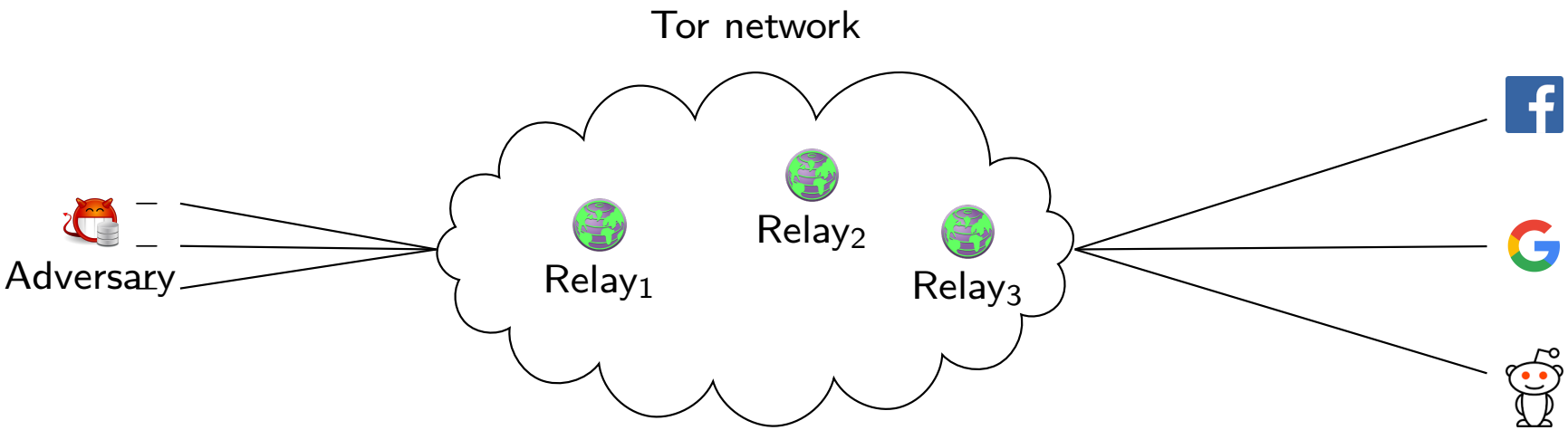
k-fingerprinting: a Robust Scalable Website Fingerprinting Technique

Jamie Hayes George Danezis

University College London

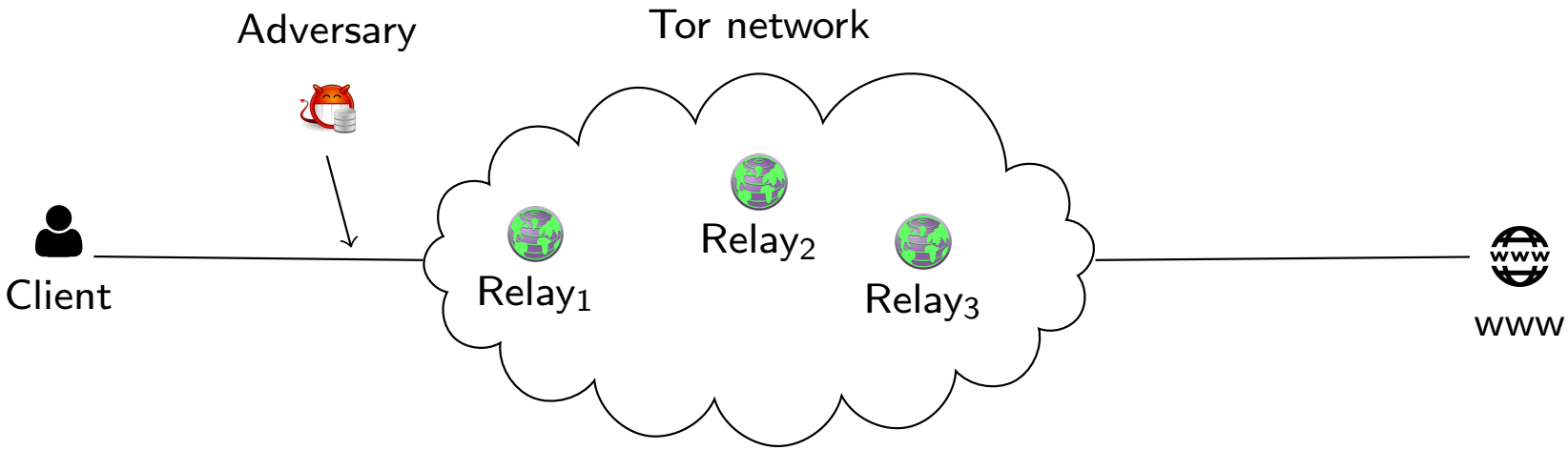
August 12, 2016

How does website fingerprinting work? - Training



Create fingerprints for , and

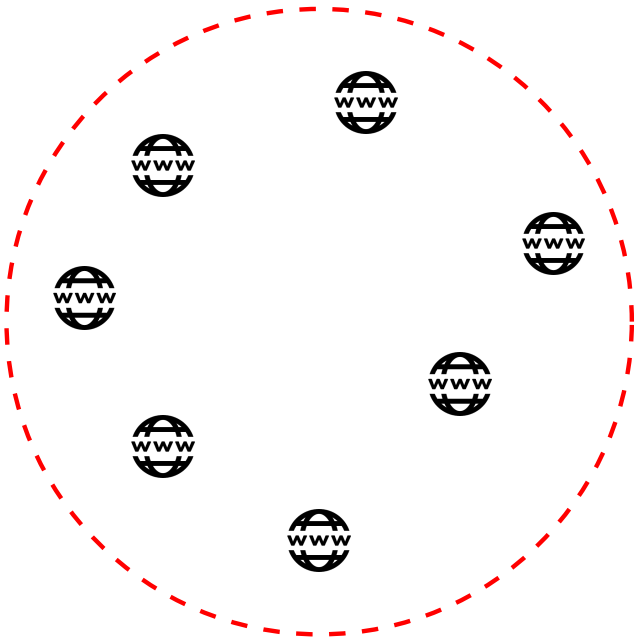
How does website fingerprinting work? - Attack



Adversary checks if fingerprint of www is equal to fingerprint of or or

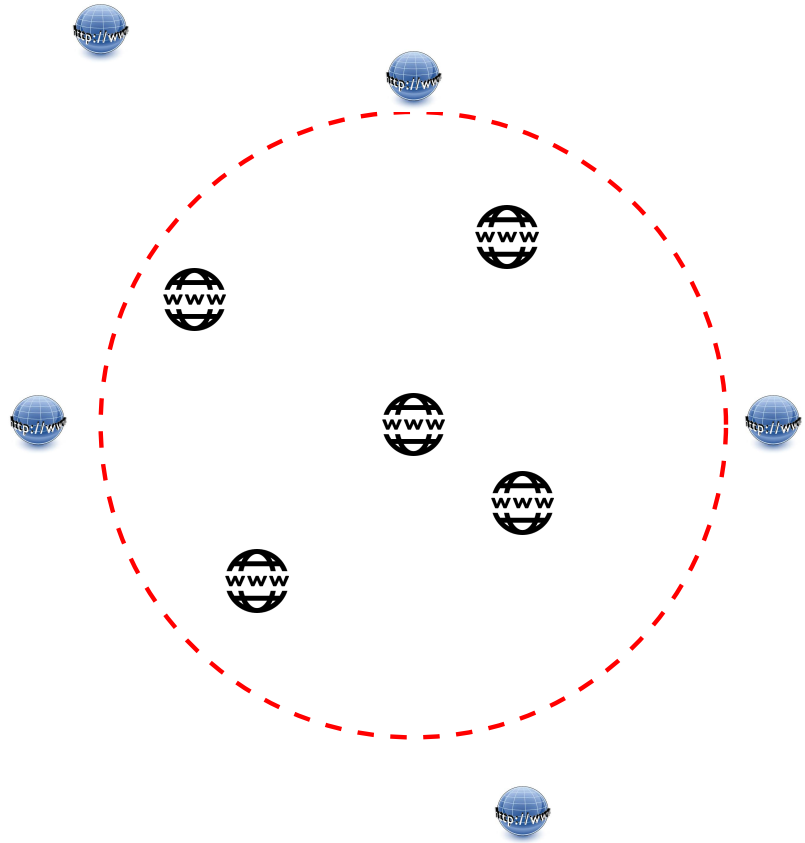
Experimental Attack set-up

Access only:



Closed World

Access any:



Open World

Contributions

k-FP - New attack based on Random Forests and k-NN[1]

An analysis of the features used in this and prior work to determine which yield the most information about an encrypted or anonymized webpage.

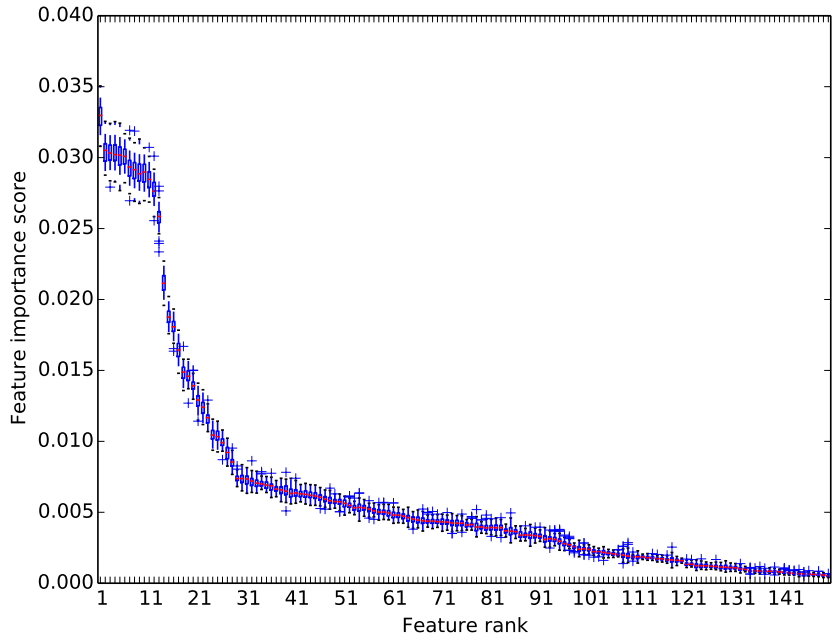
Large open world setting. In total we tested *k*-FP on 101,130 unique webpages.

Experimented with both standard websites and Tor hidden services.

[1] Wang et al. "Effective Attacks and Provable Defenses for Website Fingerprinting" 2014

Feature Analysis

Features need to be drawn from a diverse set to bypass targeted WF defenses.



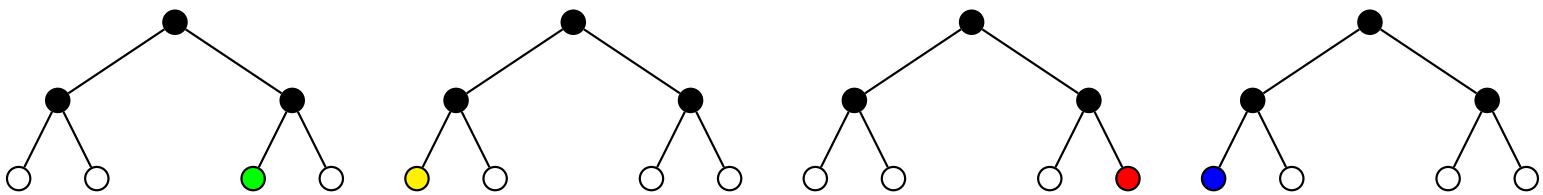
The "best" features were number of packets (incoming/outgoing) and information leaked from the first few seconds of loading a webpage.

k-FP Attack

Train on a classification task with network traffic information as features.

Use Random Forest output as the fingerprint of a website load.

Then use k-NN for classification.



Base Rate

Previous attacks had very high True Positive Rate (TPR) and very low False Positive Rate (FPR), but as the number of samples rises so too will the false alarms.

As the number of samples grows, the vast majority of alarms will be false positives.

Base Rate

FPR needs to be very low for an accurate attack as more fingerprints are tested.

Suppose we have a FPR of 1%.

If a client loads 100 unmonitored webpages. Then the attacker will mark 1 webpages incorrectly as monitored.

If a client load 1,000,000 unmonitored webpages. Then the attacker will mark 10,000 webpages incorrectly as monitored.

Accuracy metrics

TPR - The probability that a monitored page is classified as the correct monitored page.

FPR - The probability that an unmonitored page is incorrectly classified as a monitored page.

BDR - The probability that a page corresponds to the correct monitored page given that the classifier recognized it as that monitored page.

Assuming a uniform distribution of pages BDR can be found from TPR and FPR using the formula

$$\frac{TPR \cdot \Pr(M)}{(TPR \cdot \Pr(M) + FPR \cdot \Pr(U))}$$

where

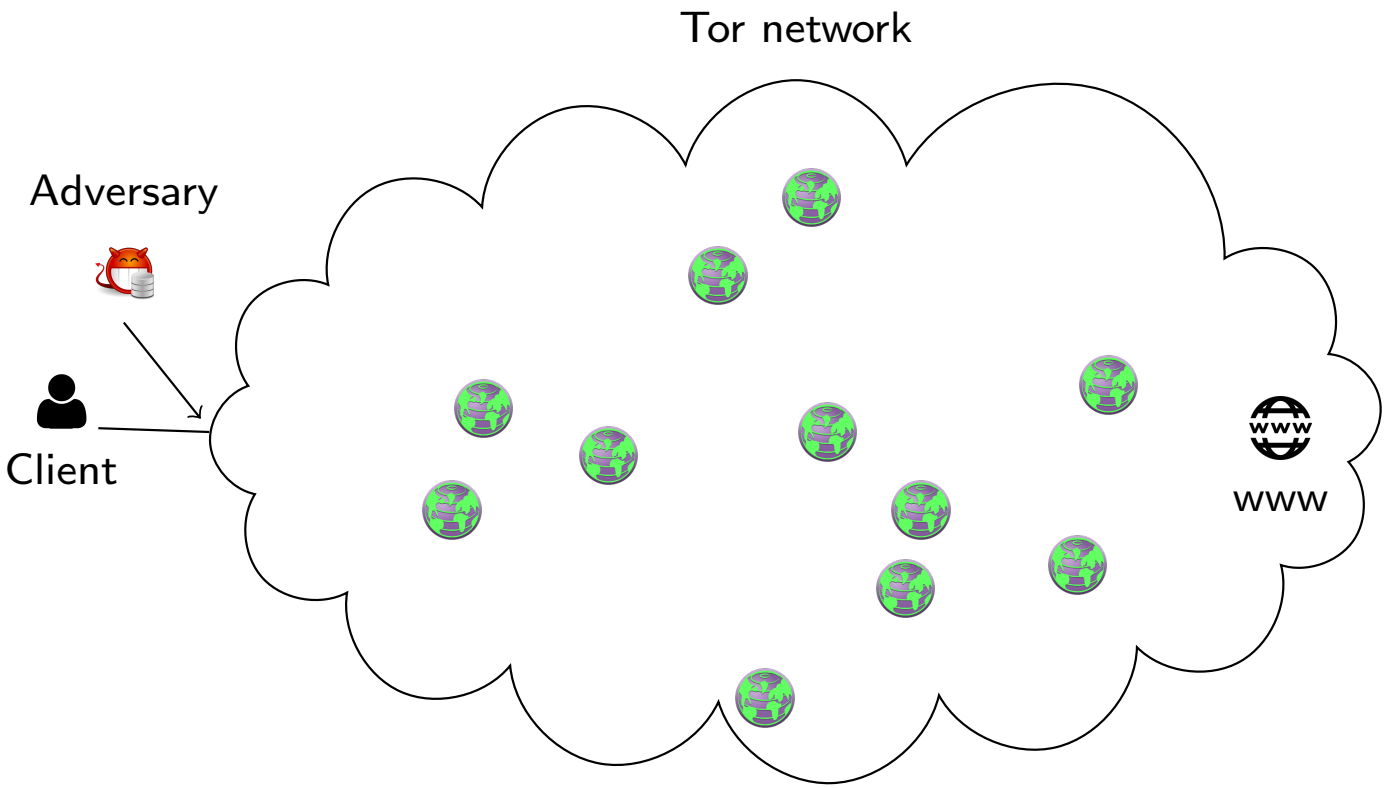
$$\Pr(M) = \frac{|\text{Monitored}|}{|\text{Total Pages}|}, \quad \Pr(U) = 1 - P(M).$$

Tor hidden services

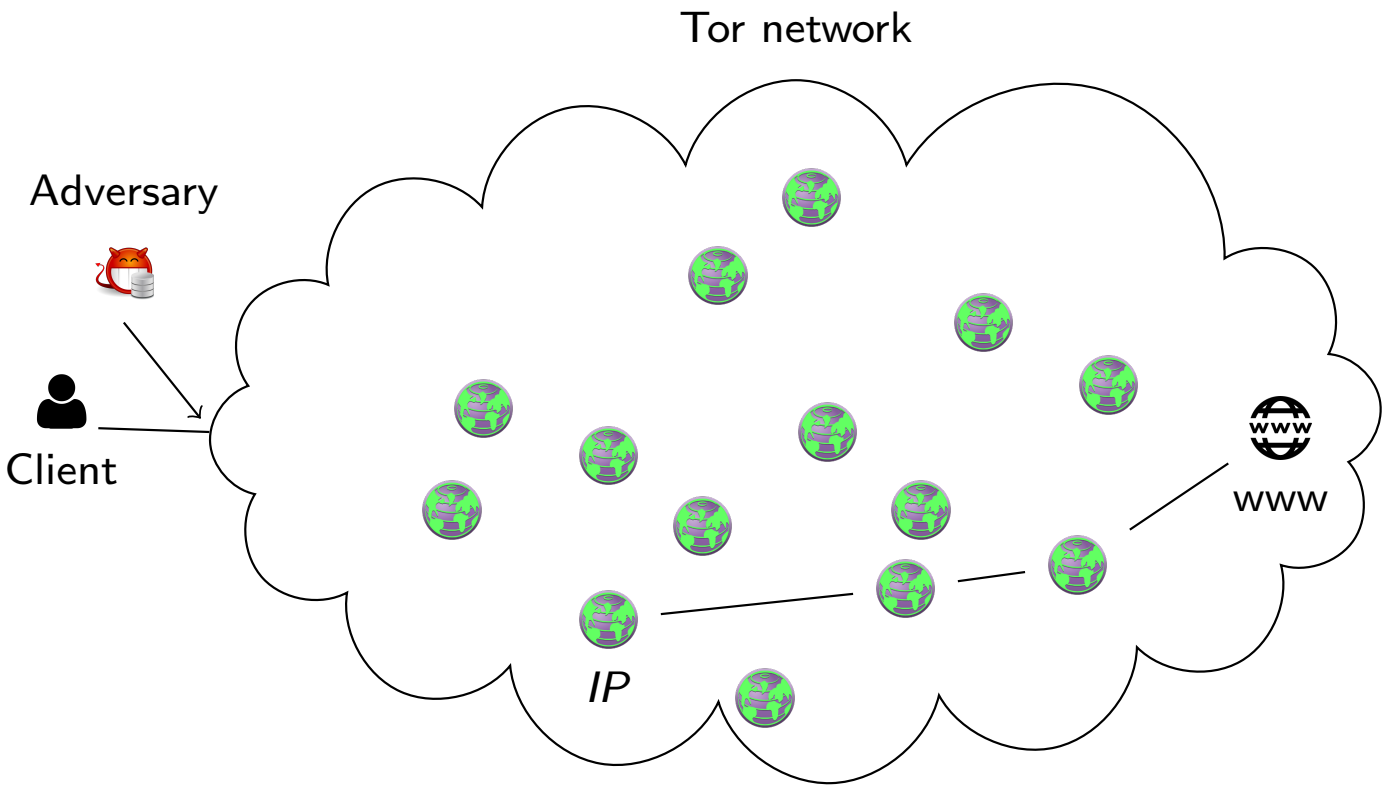
Protects receiver anonymity in addition to sender anonymity.

Sensitive servers such as SecureDrop use Tor hidden services.

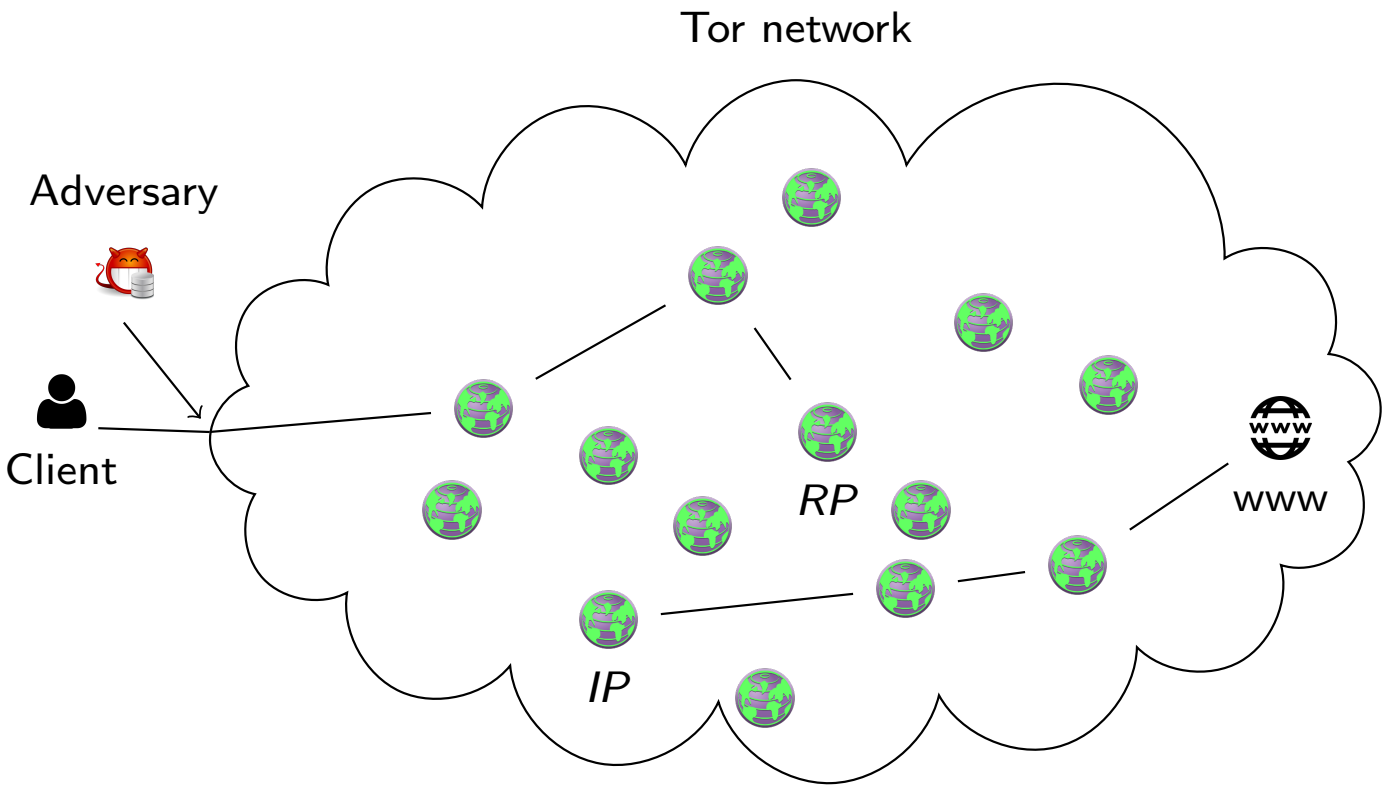
Tor hidden services



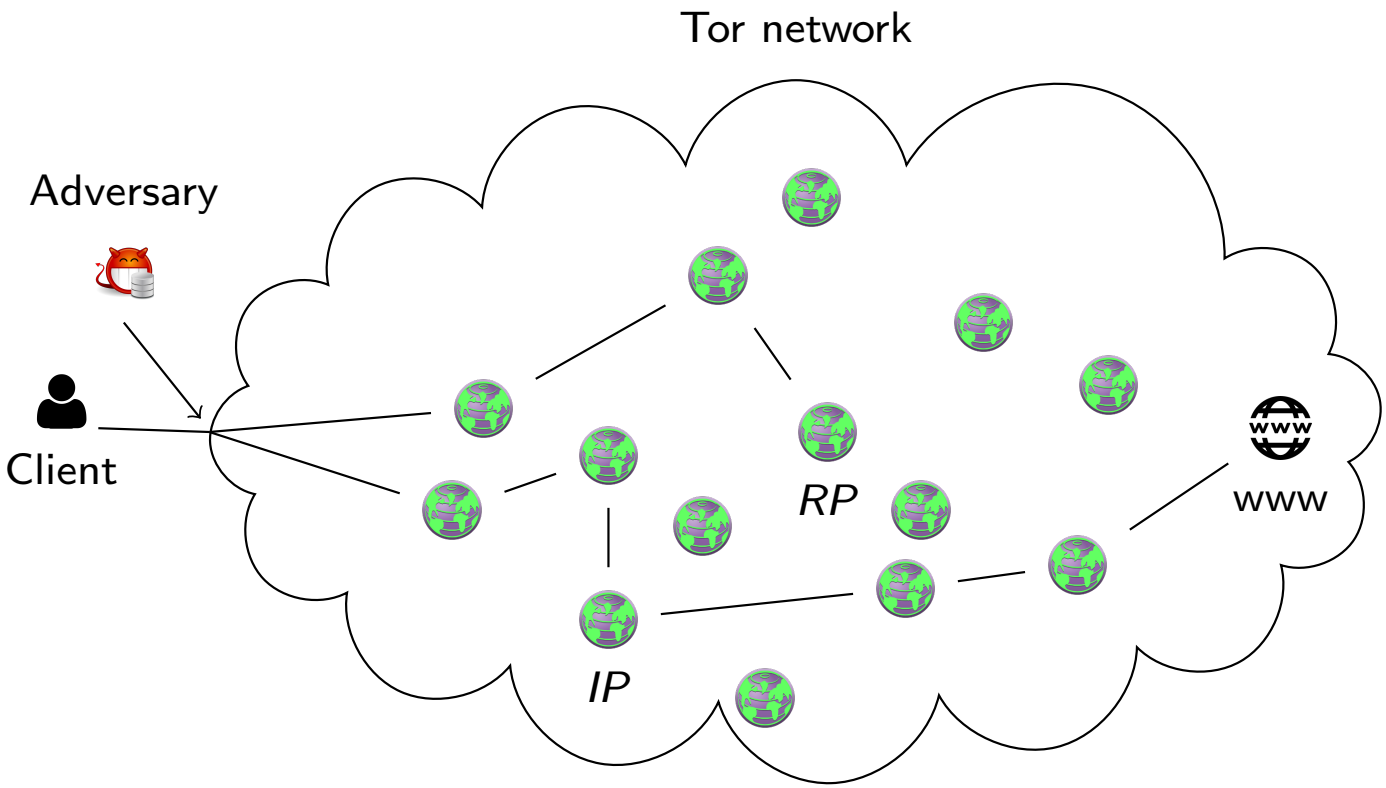
Tor hidden services



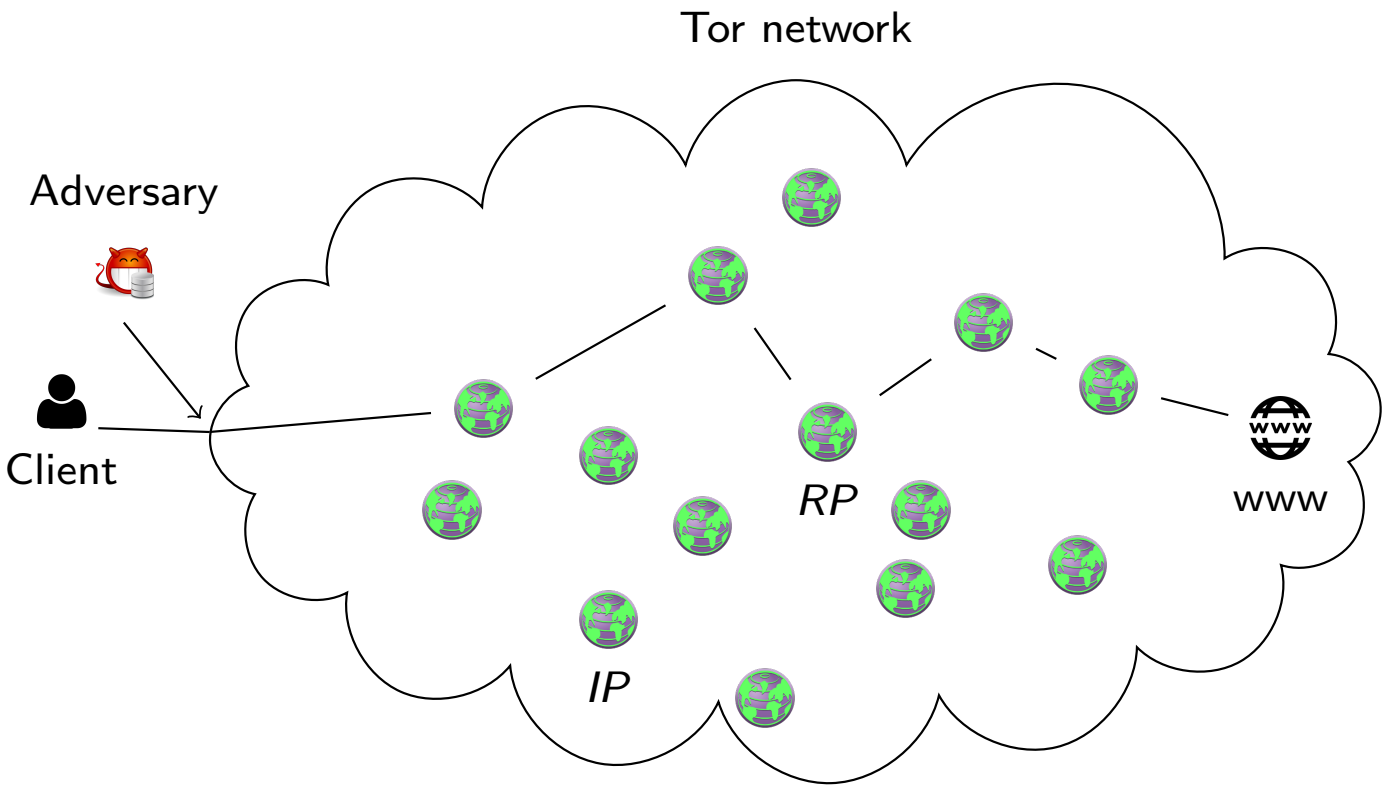
Tor hidden services



Tor hidden services



Tor hidden services



Prelims

All traffic was collected via Tor.

Monitored websites by the Adversary - Alexa Sites (Google, Facebook, Wikipedia etc.) & popular Tor Hidden Services

Only collected landing page of each website.

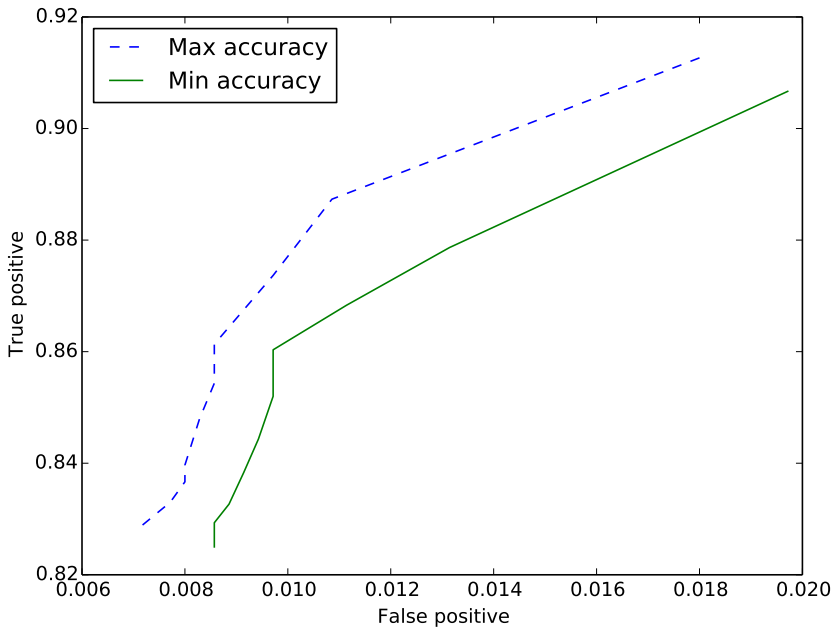
Alexa monitored set consisted of 100 samples for each of 55 websites.

Hidden Services monitored set consisted of 80 samples for each of 30 Hidden Services.

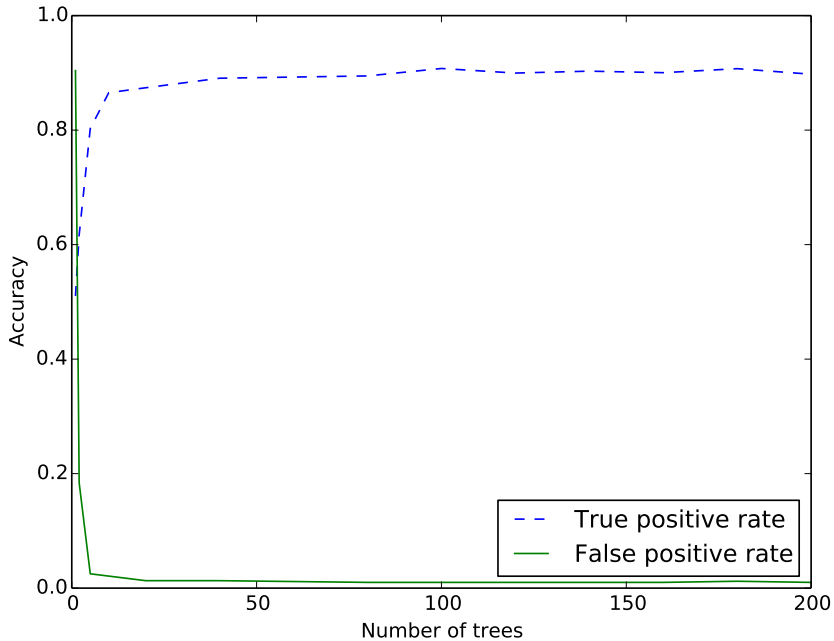
Extra sites for testing purposes - 100,000 websites (chosen from top Alexa list).

Parameter tuning - number of neighbours and number of trees

Number of neighbours



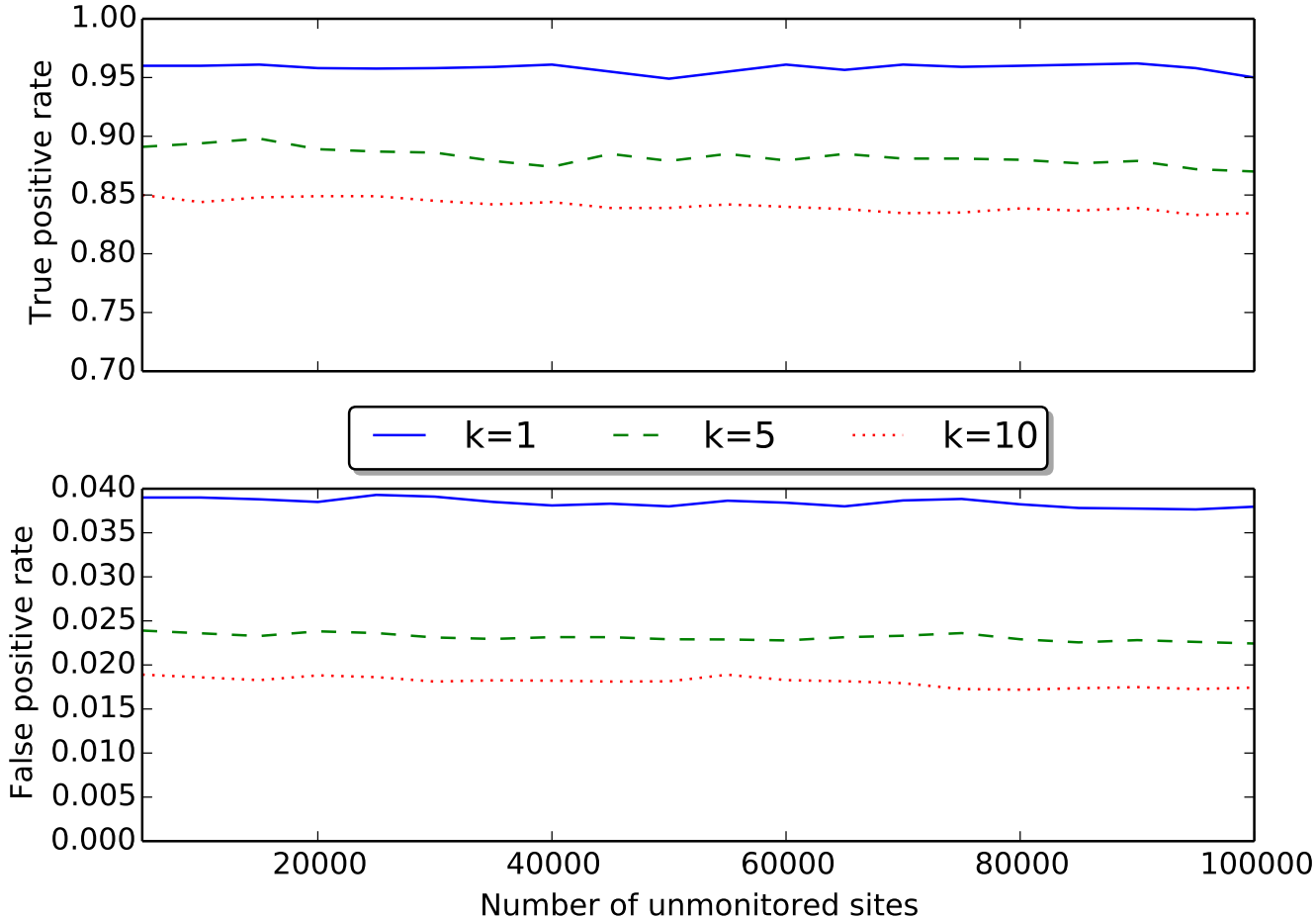
Number of Trees



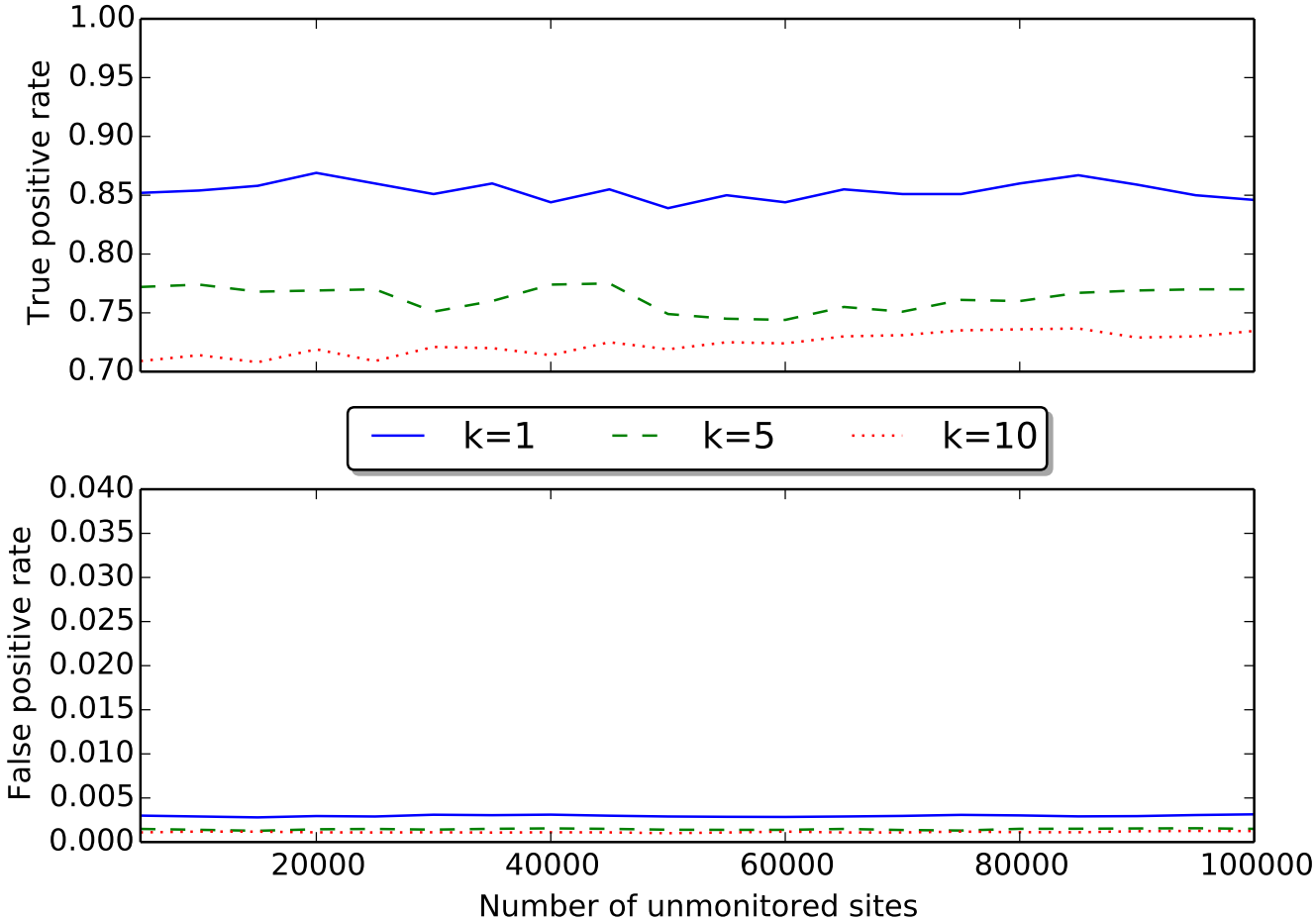
Using different k, the number of neighbours allows us to tune the TPR and FPR.

After adding 15 decision trees only incremental benefit in adding more.

Alexa monitored set results

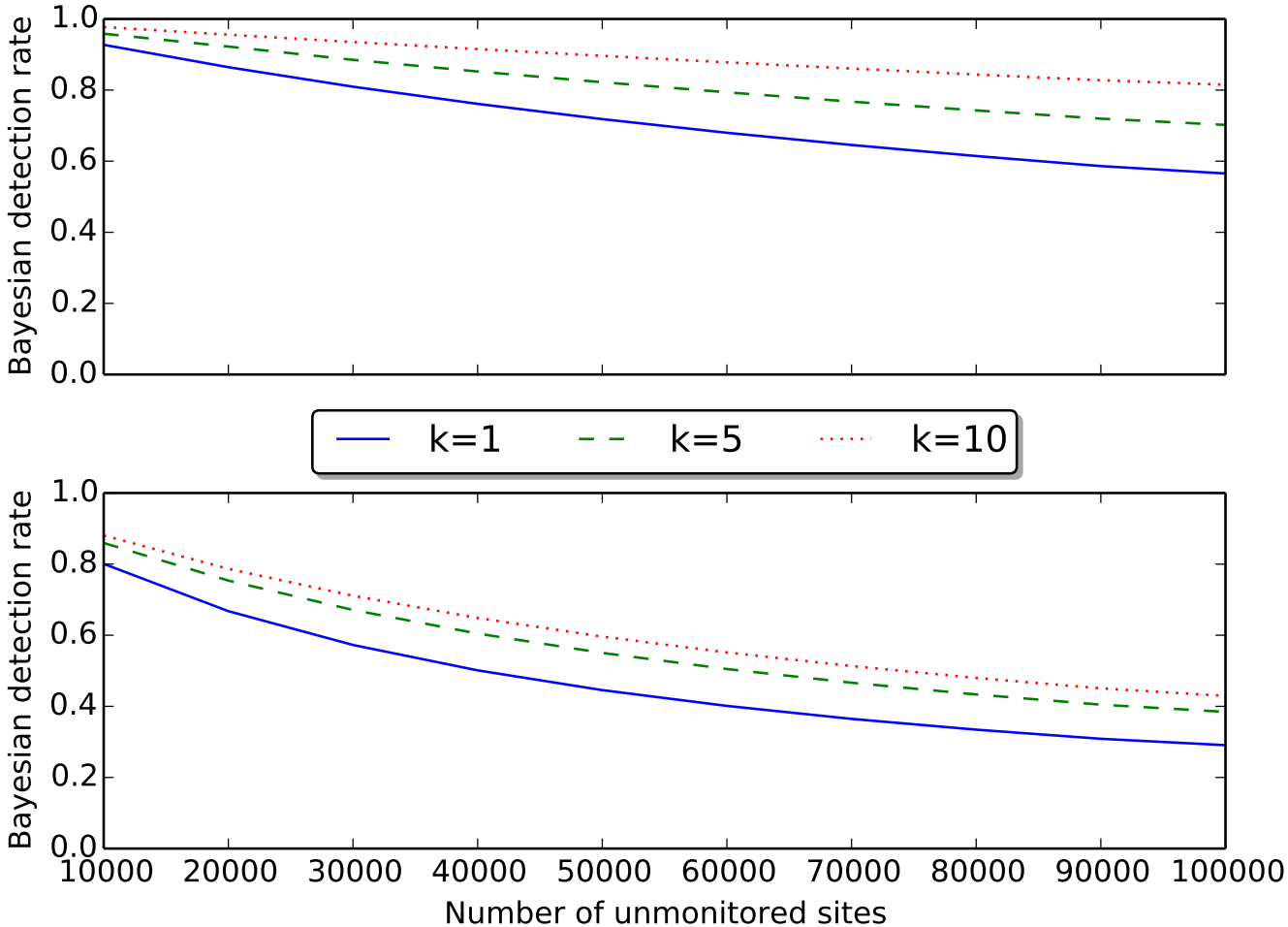


Tor hidden service monitored set results



BDR

Tor Hidden Services Monitored set.



Alexa Monitored set.

Limitations

"The BDR implicitly assumes a base rate, with no particular backing in reality." - We assume uniform expectation of visiting a webpage.

"I would like to better understand how these techniques would work if the attacker did not know the start/stop time that the user visits each website." - Website fingerprinting evaluation may not reflect practical risks.

Conclusion

The open world is not as much of a problem as we had thought, and using state-of-the-art machine learning we expect to be able to tackle other obstacles such as start-stop time identification and multiple tabs.

Attack is highly accurate over a large number of webpages.

Distiguishability between Tor Hidden Services and Non Tor Hidden Services.

Thanks

Questions?

j.hayes@cs.ucl.ac.uk

@_jamiedh

http://www.homepages.ucl.ac.uk/~ucabaye/