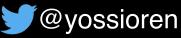
Side-Channel Attacks and Human Secrets

Yossi Oren, BGU

https://iss.oy.ne.ro



CROSSING Conference, TU Darmstadt, Germany September 2019

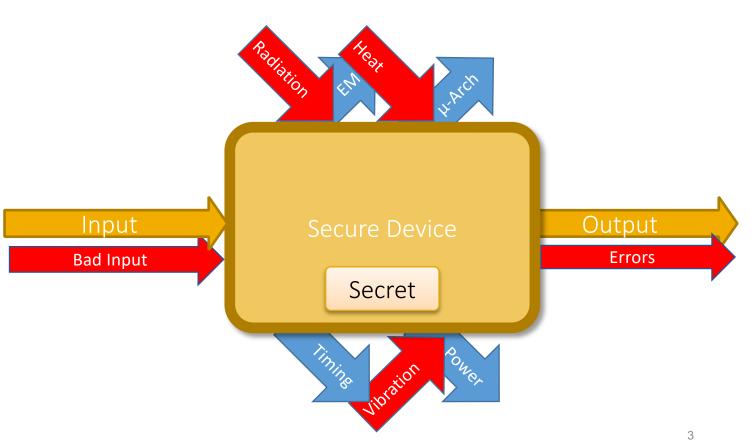
Joint work with Anatoly Shusterman, Lachlan Kang, Yosef Meltser, Yarden Haskal, Prateek Mittal and Yuval Yarom

אוניברסיטת בן-גוריון בנגב Ben-Gurion University of the Negev

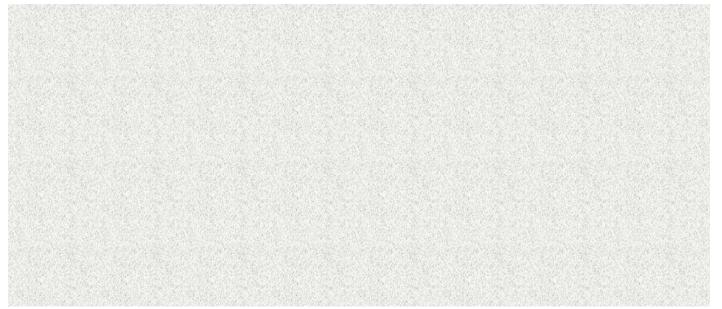


https://orenlab.sise.bgu.ac.il

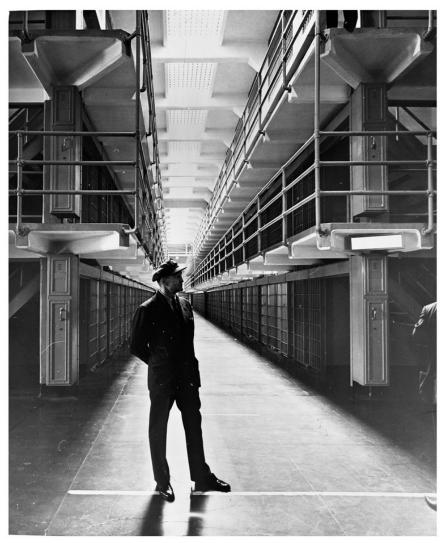
Implementation Attacks



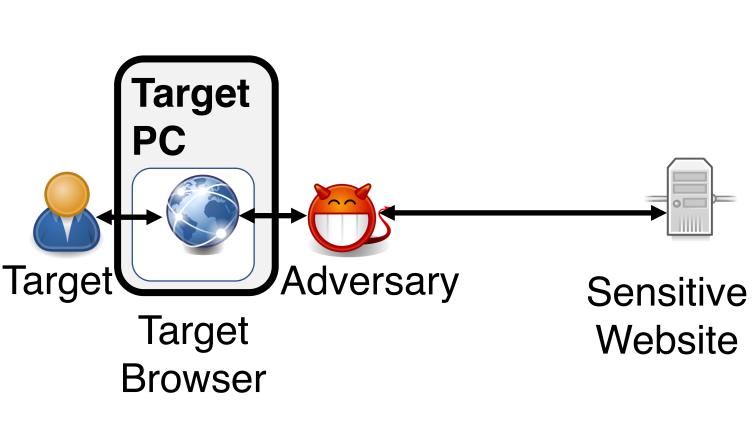
Types of Secrets

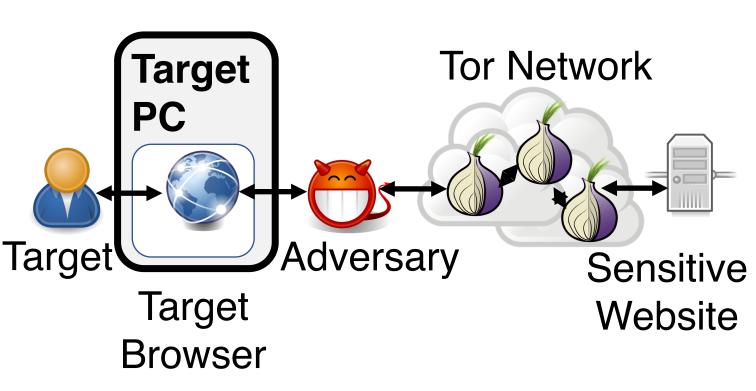


- What if the secret is compromised?
- How do we protect the secret from attack?



Credit: SF Public Library courtesy of Golden Gate NRA, Park Archives, Interpretive Negative Collection, GOGA-2316





Website Fingerprinting

Automated Website Fingerprinting through Deep Learning Vera Rimmer^{*}, Davy Preuveneers^{*}, Marc Juareg⁵, Tom Van Goethem^{*} and Wouter Joosen^{*} in International Strategy and S never the origin and destination of a communication at the Email: marc.juarez@esat.kuleuven.be never me origin and oestination of a communication at me same time. Tor's architecture thus prevents ISPs and local same une, not s arcunceture trus prevents for and near network observerts from identifying the websites users visit. nework observers from identifying the websites users insti-As a result of previous research on Tor privacy, a serious As a result of previous research on 10r privacy, a scinuse side-channel of Tor network traffic was revealed that allowed Abstract-Several studies have shown that the network traffic succentained or for network traffic was revealed that attowed a local adversary to infer which websites were visited by a Abstract—Neveral studies have shown that the network trained that is generated by a visit to a website over Tor reveal to construct the state of the state of the state of the state of the state to the state of the a local adversary to inner writen websites were visited by a particular user [14]. The identifying information leaks from the area between the second by the that is generated by a visit to a website over Tor reveals information specific to the website through the limits and over a manuel, market, the manual terms, the second second particular user (14). The localitying information reass usual the communication's meta-data, more precisely, from the data the communication is meta-data. information specific to the website through the luming and sizes of network packets. By capturing traffic traces between the second data to the trace network of the network of the second seco ue communications insta-tradit, more precisely, ituni une ar-rections and sizes of energyfiel network packets. As this side sizes of network packets. By capturing traffic traces between users and their for entry guard, a network envedropper can be sense that more data to control which under the control of the users and their for entry guard, a network eavedropper can everygate this metadata to reveal which website for users are white the second of each state to the second of t rections and sizes of encrypted network process, as and side. It channel information is often unique for a specific website, it is the structure of the structu leverage this meta-data to reveal which website for users are bibling. The success of such attacks heavily depends on the 201 visiting. The success of such attacks nearly depends on the particular set of traffic features that are used to construct a becommented traffic features are measure environment. particular set of traffic features that are used to construct the ingerprint. Typically, these features are manually engineering ingerprint, typically, these learners are manually engineered and, as such, any change introduced to the for network, can sender these sendents constants for the interaction of a sender

website ingerprintings and find that the performance achieved by our deep learning approaches is comparable to known mechadis

our deep learning approaches is comparable to known methods which include various research efforts spanning over multiple

which include various research efforts spanning over multiple years. The obtained success rate exceeds 96% for a desed work of 100 weature and 0.4% for any kinemet should work of any

une remains ware approach an el technique for website fingerprinting.

rears. The obtained success rate exceeds 96% for a closed world of 100 websites and 94% for our biggest closed world of 900

of 100 websites and 94% for our biggest closed world of 900 dasses. In our open world evaluation, the most performant deen learning model is 26 were accurate these the comparison

10

S

3

06374

80

channel information is often unique for a specific wersate, a can be leveraged to form a unique fingerprint, thus allowing can oc reveraged to form a unique nugerprint, mus allowing network earesdroppers to reveal which website was visited and, as such, any change introduced to the Tor network can render these carefully constructed features ineffective. In this naseo on the traine that it generated. The feasibility of Website Fingerprinting (WF) attacks on The reasoning of website ringerprinting (wr) arraces on the was assessed in a series of studies [25], [3], [19], [24], [24], [25], [25], [26], [render these carefully constructed features ineffective, in this paper, we show that an adversary can automate the feature sentencies are also and thus automatically demonstrative tree based on the traffic that it generated. paper, we show that an adversary can automate the feature orgineering process, and thus automatically deamonymize for traffic he analytim our norse motion bound to down to main to a engineering process, and thus automatically deamonynize Tor traffic by applying our good method based on deep learning. We ender a A dearest commutant of more them them emition encount traffic by applying our novel method based on deep learning. We work to allocid a dataset comprised of more than three million network terrors, which is the largest dataset of each terrors, one of the largest dataset of each terrors. collect a dataset comprised of more than three million network traces, which is the largest dataset of web traffic ever used for instant second and the dataset of an advector and advector of the dataset of the datase traces, which is the largest dataset of web traffic ever used for respire for the second seco

for was assessed in a series or sources (23), (34), (19), (24), (32), in the related works, the attack is treated as a classi-1241 in the remety works, the marks is treated as a classification problem. This problem is solved by, first, manually nearon proviem. This proviem is sorred of, first- manually engineering features of traffic traces and then classifying these engineering reatures of traine traces and user classifying uses features with state-of-practice machine learning algorithms reaures wan state-of-practice machine learning argoritanis-Proposed approaches have been shown to achieve a classifica-Proposed approaches have been snown to achieve a classification accuracy of 91.96% correctly recognized websites (301.1131) and accuracy of 101.000 metatric material from the formation of the source tion accuracy or 91-90% correctly recognized websites 1-90. [24], [13] in a set of 100 websites with 100 traces per website. [24], [13] in a set of 100 websites with 100 traces per website. Their works show that finding distinctive features is essential ineit works snow ma intuing distinctive reatures is essential for accurate recognition of websites. Moreover, this tasks can dasses. In our open world evaluation, the most performant deep learning model is 2% more accurate than the state-ofor accurate recognition or websites, moreover, this tasks can be costly for the adversary as he has to keep up with changes deep learning model is 2% more accurate than the state-of-the-art attack. Furthermore, we show that the implicit features the art attack. Furthermore, we show that the implicit resulting automatically learned by our approach are far more regilient to a supervise determine of unit control over time we consider the to costly for the adversary as the has to keep up with changes introduced in the network protocol [4], [20], [9]. The WF automatically learned by our approach are far more realismt to dynamic changes of web content over time. We conclude that the additive to extraorationally constrained the neurophysical phonon content. introduced in the network productor [44], [20], [2], the wr research community thus far has not investigated the success dynamic changes of yeb content over time. We conclude mail the ability to automatically construct the most relevant traffic features and endowed memory terms reconstruct memory and research community mus tar nas non unexugance are success of an attacker who automates the feature extraction step for the analy to automatically construct the most relevant traffic features and perform accurate traffic recognition makes our deen learning based assessed an encident formula and encident teatures and perform accurate traffic recognition makes our deep learning based approach an efficient, flexible and robust inductions for exclusion for exclusion for exclusion for exclusion.

or an anascer who automates are resurre extraction step for classification. This is the key problem that we address in this vork. An essential step of traditional machine learning is feature women any or narrowan morenny rearing is reading a reading and the second of the secon o the a communication tool that proa browledge, to find a representation of raw ars It is an actively developed a that are most relevant to the work. d to be even more

- Collect Labeled **Network Traces**
- Extract Features
- Train Classifier (classical/deep)

Classify Unknown **Network Traces**

How is WF Evaluated?

Automated Website Fingerprinting through Deep Learning Vera Rimmer^{*}, Davy Preuveneers^{*}, Marc Juareg⁵, Tom Van Goethem^{*} and Wouter Joosen^{*} never the origin and destination of a communication at the Email: marc.juarez@esat.kuleuven.be never me origin and oestination of a communication at me same time. Tor's architecture thus prevents ISPs and local same ume. For s arcunceture ruts prevents for and local network observers from identifying the websites users visit. nework observers from identifying the websites users insti-As a result of previous research on Tor privacy, a serious As a result or previous research on 10r privacy, a scinuse side-channel of Tor network traffic was revealed that allowed Abstract-Several studies have shown that the network traffic succentained or for network traffic was revealed that attowed a local adversary to infer which websites were visited by a Abstract—Several studies have shown that the network traffic that is generated by a visit to a website over for reveal to the several second of the two sets the second to the transformed a local adversary to inter which websites were visited by a particular user [14]. The identifying information leaks from the set that is generated by a visit to a website over Tor reveals information specific to the website through the limits and over a manuel, market, the manual terms, the second second particular user [14]. The localitying minimum cases usual be communication's meta-data, more precisely, from the data the communication's meta-data. information specific to the website through the lumang and stars of network packets. By capturing traffic traces between the second data to the trace network of the network of the second seco the communication's meta-tasta, more precisely, roun the ar-rections and sizes of encrypted network packets. As this side sizes of network packets. By capturing traffic traces between users and their for cutty guard, a network cavedropper can become the meta-date to evolve which website for traces are teers and their for entry guard, a network eavedropper can leverage this metadata to reveal which website for users are solution of the second of each state to the second of the second rections and sizes of encrypted network process, as and side. It channel information is often unique for a specific website, it is the structure of the structu leverage this meta-data to reveal which velsite Tor users are thisting. The success of such attacks heavily depends on the channel information is often unique for a specific website, it can be leveraged to form a unique fingerprint, thus allowing 201 visiting. The success of such attacks heavily depends on me particular set of traffic features that are used to construct a proceeding traffic features for measurements and an and the set of the set can be reveraged to form a unique nugerprint, thus autowing network earesdroppers to reveal which website was visited particular set of traffic features that are used to construct the interprint. Typically, these features are manually engineered and as each one Annue introduced to the two encounts and fingerprint. Typically, these features are manually engineered and as such any change introduced to the for network, can be sender these events there exists a such as the sender the sende nased on the traine that it generated. The feasibility of Website Fingerprinting (WF) attacks on and, as such, any change introduced to the Tor network can render these carefully constructed features ineffective. In this source we show that an observance can extremely the frames render these carefully constructed features ineffective, in this paper, we show that an adversary can automate the feature sentencies are also and thus automatically demonstrative tree based on the traffic that it generated. paper, we show that an adversary can automate the feature orgineering process, and thus automatically deamonymize for traffic by analyting our named modeled based on down beaution. We engineering process, and thus automatically deamonymize for traffic by applying our provel method based on deep learning. We ender a A densed constructed of source theor three minime reserved. traffic by applying our novel method based on deep learning. We work to allocid a dataset comprised of more than three million network terrors, which is the largest dataset of each terrors, one of the largest dataset of each terrors. 10 collect a dataset comprised of more than three million network traces, which is the largest dataset of web traffic ever used for instant second and the dataset of an advector and advector of the dataset of the datase traces, which is the largest dataset of web traffic ever used for respire for experimental states of the performance achieved by our dword between association to the state of the state of

website ingerprintings and find that the performance achieved by our deep learning approaches is comparable to known mechadis

cs

3

06374

80

our deep learning approaches is comparable to known methods which include various research efforts spanning over multiple

which include various research efforts spanning over multiple years. The obtained success rate exceeds 96% for a closed work of 100 weature and 0.000 for a travel strengt strengt strengt strengt

and remains waves approach an el

rears. The obtained success rate exceeds 96% for a closed world of 100 websites and 94% for our biggest closed world of 900

of 100 websites and 34% for our biggsst closed world of 900 dasses. In our open world evaluation, the most performant deen learning model is 1% more commute the the standard

dasses. In our open world evaluation, the most performant deep learning model is 2% more accurate than the state-of-

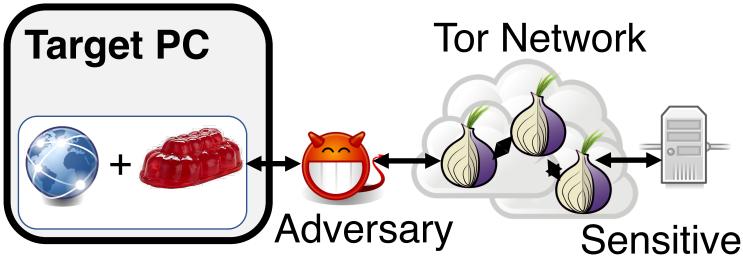
deep learning model is 2% more accurate than the state-of-beart attack Furthermore, we show that the implicit features suscentiated beamed to success the termination of the state-ofthe art attack. Furthermore, we show that the implicit resulting automatically learned by our approach are far more regilient to a supervise determine of unit events are not to be a supervised way and the supervised state.

The reastbility of website ringerprinting (wr) atraces on for was assessed in a series of studies [25], [3], [19], [24], for was assessed in a series or sources (23), (34), (19), (24), (32), in the related works, the attack is treated as a classi-[52]. In the related works, the attack is treated as a cassification problem. This problem is solved by, first, manually nearon proviem. This proviem is sorred of, first- manually engineering features of traffic traces and then classifying these engineering realities or traine traces and men classifying discrete features with state-of-practice machine learning algorithms teaures with state-of-practice machine learning argoritanis-Proposed approaches have been shown to achieve a classifica-Proposed approaches have been snown to achieve a classification accuracy of 91.96% correctly recognized websites (301.1131) and accuracy of 101.000 metatric material from the formation of the source tion accuracy or 91-90% correctly recognized websites 1-90. [24], [13] in a set of 100 websites with 100 traces per website. [24], [15] in a set of 100 websnes with 100 traces per websne. Their works show that finding distinctive features is essential fuelt works snow mai infund distinctive readines is essential for accurate recognition of websites. Moreover, this tasks can or accurate recognition or websites, moreover, this tasks can be costly for the adversary as he has to keep up with changes ve costry for the autorstary as the task to keep up with changes introduced in the network protocol [4], [20], [9]. The WF automatically learned by our approach are far more realismt to dynamic changes of web content over time. We conclude that the additive to extraorationally constrained the neurophysical phonon content. introduced in the network productor [44], [20], [2], the wr research community thus far has not investigated the success dynamic changes of web content over time. We conclude that the ability to automatically construct the most relevant traffic for the start surface contents traffic committee and research community mus tar nas non unexugance are success of an attacker who automates the feature extraction step for the analytic to automatically construct the most relevant traffic recognition makes our density and perform accurate traffic recognition makes and encoded reatures and perform accurate traffic recognition makes our deep learning based approach an efficient, flexible and robust inductions for exclusion for exclusion for exclusion for exclusion.

or an analysic who automates the realize extraction step for classification. This is the key problem that we address in this $\sum_{i=1}^{n} d_{i}$ vork. An essential step of traditional machine learning is feature wanna any or naumona macany canne ia caany paring Feature engineering is a manual process, based on o the a communication tool that proa boouledge to find a representation of raw ars It is an actively developed a that are most relevant to the work. d to be even more

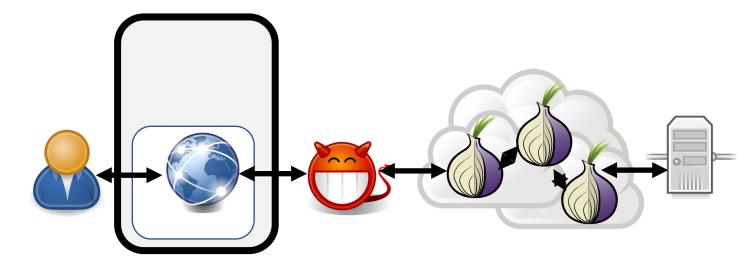
- Main metric is accuracy
- Closed World vs Open World
- Base rate is important!
- Network based WF has >90% accuracy

Traffic Moulding Defenses against WF



Website

Source: lakeland.co.uk





Adversary

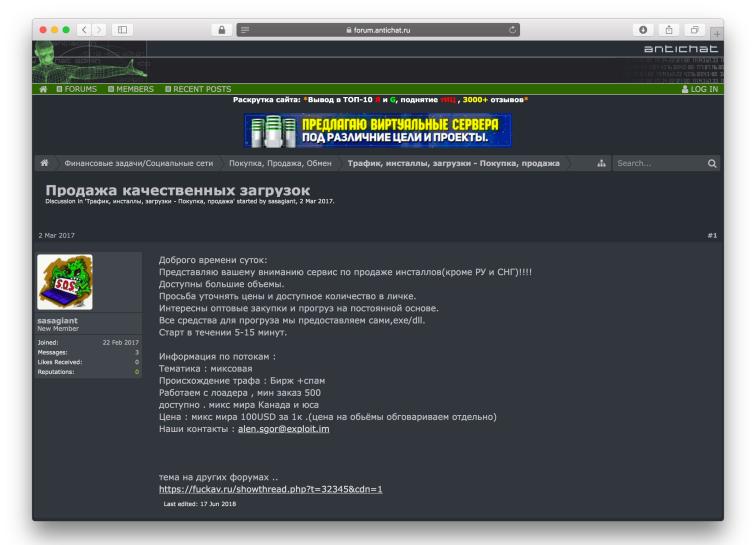
Architectural Boundary

Tor Network

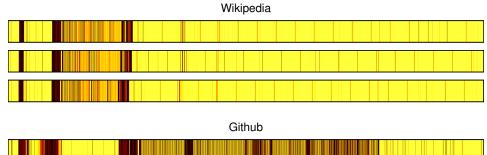
Target Browser

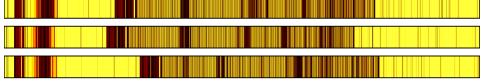
Target

Sensitive Website

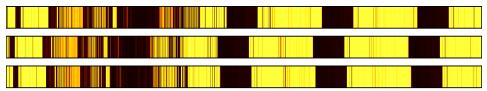


Memorygrams





Oracle



Cache-Based WF

Prateek Mittal

Princeton University

Robust Website Fingerprinting Through the Cache Occupancy Channel Anatoly Shusterman Ben-Gurion University of the Negev shustera@post.bgu.ac.il Yosef Meltser Ben-Gurion Univ. of the Negev yosefmel@post.bgu.ac.il Yarden Haskal Ben-Gurion Univ. of the Negev yardenha@post.bgu.ac.il Ben-Gurion Univ. of the Negev Yossi Oren yos@bgu.ac.il

Website fingerprinting attacks, which use statistical analysis on network traffic to compromise user privacy, have ysis on network mane to compromise user privacy, nave been shown to be effective even if the traffic is sent over anonymity-preserving networks such as Tor. The classical assumes an on-path adversary, who can observe all traffic

traveling between the user's computer and the secure net-In this work we investigate these attacks under a different attack model, in which the adversary is capable of sending a small amount of malicious JavaScript code to the target

a sman answin or maneurus ravasenta cone io me inger user's computer. The malicious code mounts a cache sidechannel attack, which exploits the effects of contention on the CPU's cache, to identify other websites being browsed. The effectiveness of this attack scenario has never been systematically analyzed, especially in the open-world model which assumes that the user is visiting a mix of both sen-

show that eache website fingerprinting attacks in sitive and non-sensitive sites. traces of cache activ-

pmittal@princeton.edu University of Adelaide and Data61 Yuval Yarom yval@cs.adelaide.edu.au

activity reduces the effectiveness of the attack and completely eliminates it when used in the Tor Browser.

1 Introduction

University of Adelaide lachlan.kang@adelaide.edu.au

Over the last decades the World Wide Web has grown from an academic exercise to a communication tool that encoman acagemic exercise to a communication toot mat encom-passes all aspects of modern life. Users use the web to acquire information, manage their finances, conduct their social life, and more. This shift to the so called virtual life has resulted in new challenges to users' privacy. Monitoring nus resurveu in new similarity o users privacy aromonant the online behavior of users may reveal personal or sensitive information about the users, including information such as sexual orientation or political beliefs and affiliations. Several tools have been developed to protect the online privacy of users and hide information about the websites

privacy or users and note mitoritation about the weisites they visit [18, 20, 71]. Prime amongst these is the Tor network [20], an overlay network of collaborating servers, alled relays, an overlay network or conneutrating servers, called relays, that anonymously forward Internet traffic beween users and web servers. Tor encrypts the network traffic tween users and web servers. For encrypts the network traine of all of the users, and transmits it between relays in a way that prevents external observers from identifying the traffic vite users. In addition to the network itself, the Tor as that further protects

- Collect Labeled Memorygrams
- Extract Features
- Train Classifier (classical/deep)
- Classify Unknown Memorygrams
- >90% accuracy

Cache-based vs Net-based WF

Cache beats Net	Net beats Cache
Resists net countermeasures	Can be detected by victim
Robust to response caching	Depends on hardware config
Works across NICs	
Lighter attack model	

Countermeasures

Power Analysis Attacks Revealing the Secrets of Smart Cards Stefan Mangard Elisabeth Oswald Thomas Popp

• Hiding

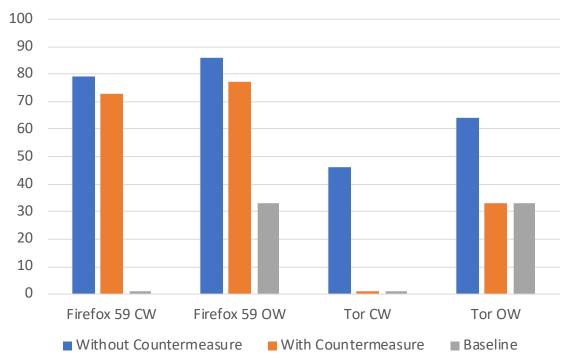
- Lowering the SNR
- Hiding in Time
- Hiding in Amplitude
- Masking
 - Secret Invariance
 - Separation in Time
 - Separation in Space

Hiding in amplitude

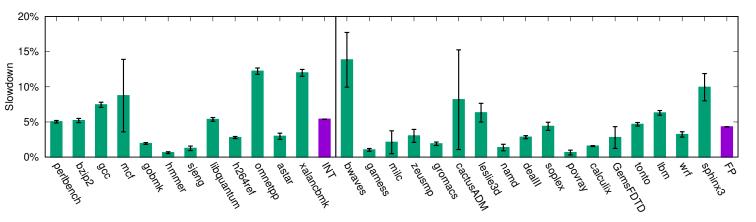
- Idea: run a dummy prime and probe in the background
- What is the effect on WF accuracy?
- What is the effect on performance?

Effect on Accuracy

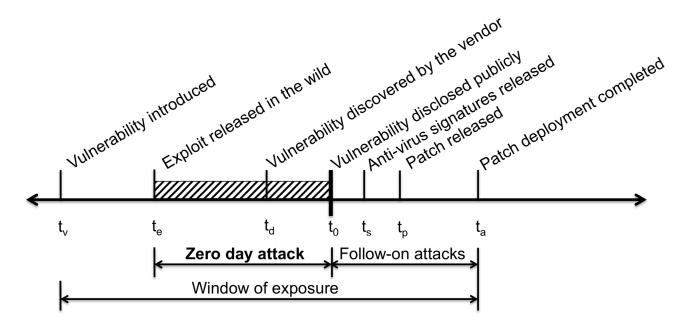
ML with Cache Activity Masking



Effect on Performance



Sustainability



Source: Bilge and Dumitras, CCS 2012

Humans

- Non upgradable, difficult to patch
- Fight security with all their might
- Semi Rational

- 00

Thank you!

- Dataset freely available under CC-BY 4.0 license
- Contains:
 - Thousands of memorygrams in multiple settings
 - Associated network traces
 - Deep learning classifiers in Python

https://orenlab.sise.bgu.ac.il/p/RobustFingerprinting





