# Side-Channel Attacks on Human Secrets

Yossi Oren, BGU

https://iss.oy.ne.ro

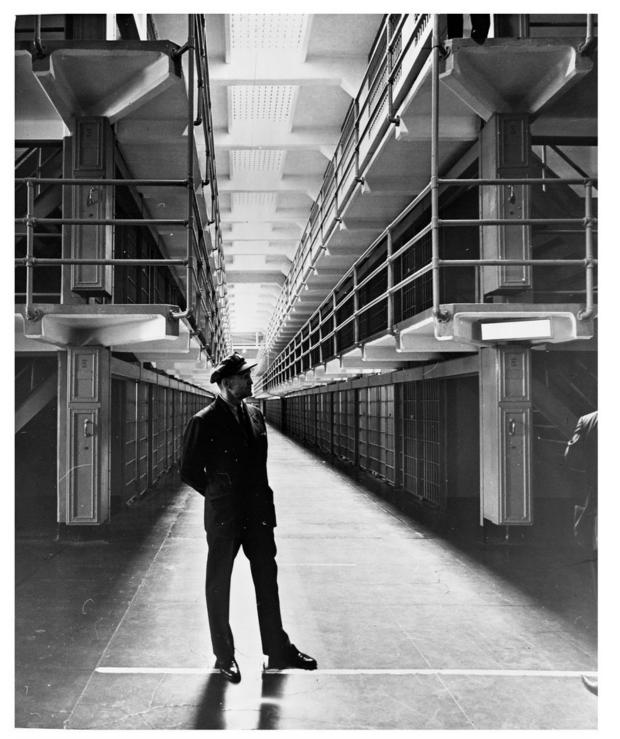


@yossioren

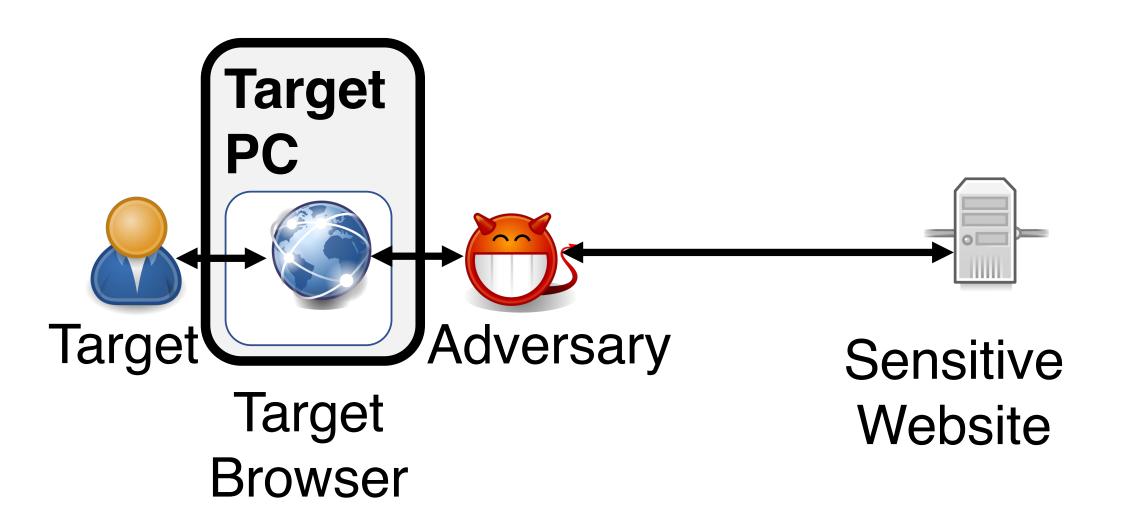
7<sup>th</sup> Technion Summer School, 04/Oct/2018

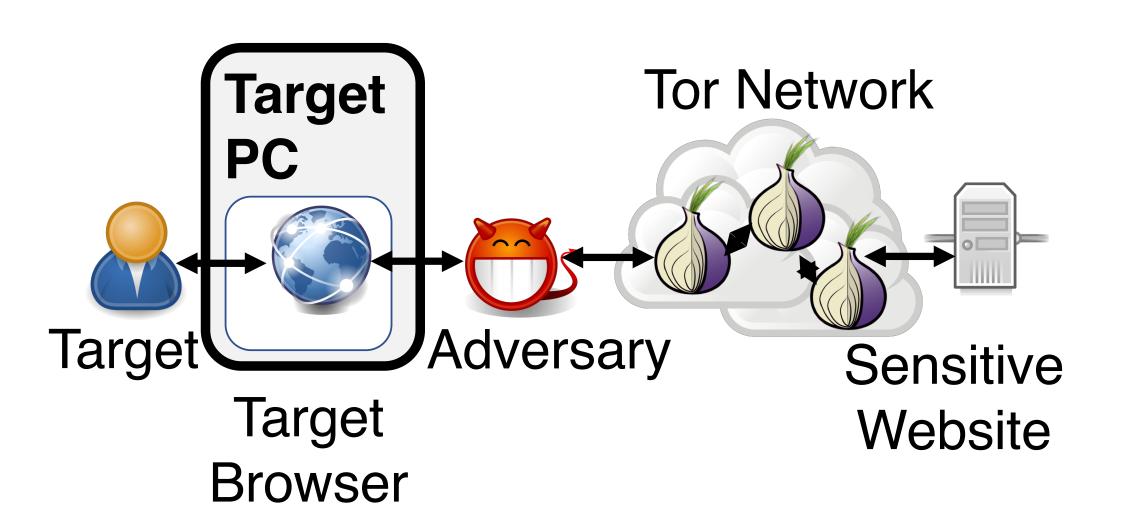
Joint work with Anatoly Shusterman, Lachlan Kang, Yosef Meltser, Yarden Haskal, Prateek Mittal and Yuval Yarom

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



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 Juarez§, Tom Van Goethem\* and Wouter Joosen\*

%imec-COSIC, ESAT, KU Leuven Email: marc.juarez@esat.kuleuven.be

Abstract—Several studies have shown that the network traffic Abstract—Several studies have shown that the network traffic that is generated by a visit to a website over Tor reveals that is generated by a visit to a website over the fining and the state of the s that is generated by a visit to a website over for reveals information specific to the website through the timing and information specific to the website through the timing and sizes of network packets. By capturing traffic traces between sizes of network packets. By capturing traffic accordance continues and their Tor network and a natural continues and their Tor network accordance continues and their torque continues and Sizes of network packets. By capturing traine traces between users and their Tor entry guard, a network eavesdropper can leverage this meta-date to reveal which website Tor users are users and their Tor entry guard, a network eavesdropper can leverage this meta-data to reveal which website Tor users are relative to the standard of the control of the standard of the stand leverage this meta-data to reveal which website Tor users are visiting. The success of such attacks heavily depends on the profit of the success of the footnote that are used to construct the construction of the footnote that are used to construct the construction of the footnote that are used to construct the construction of the footnote that are used to construct the construction of the construction o visiting. The success of such attacks heavily depends on the particular set of traffic features that are used to construct the particular set of traffic features are manually opinionally there features are manually opinionally. particular set of traffic features that are used to construct the ingerprint. Typically, these features are manually engineered manually engineered manually engineered manually engineered any change introduced to the Tor naturally and as each any change introduced to the Tor naturally and as each any change introduced to the Tor naturally and as each any change introduced to the Torinature. ungerprint. Typically, these features are manually engineered and, as such, any change introduced to the Tor network can render those carefully constructed features in offention. In the carefully constructed features in offention. and, as such, any change introduced to the 10r network can render these carefully constructed features ineffective. In this render these carefully constructed features methodize in this paper, we show that an adversary can automate the feature paper, we show that an adversary can automatically demonstrate the paper, we show that an adversary can automatically demonstrate the feature paper. paper, we show that an adversary can automate the feature regineering process, and thus automatically deanonymize useful for the beautiful too the continuous contraction to the contraction of the contrac engineering process, and thus automatically deanonymize Tor traffic by applying our novel method based on deep learning. We collect a detect commerced of more those million not core traffic by applying our novel method based on deep learning. We collect a dataset comprised of more than three million network than three million network than three million network than the learning dataset of most traffic areas made for the traffic areas made for collect a dataset comprised of more than three million network traces, which is the largest dataset of web traffic ever used for probability for and find that the nonformation ordinary to probability for the proportion of the state of the nonformation of the nonformatio traces, which is the largest dataset of web traffic ever used for website fingerprinting, and find that the performance achieved by website fingerprinting, and find that the performance methods we have been approached to known methods and have been proposed to the proposed to the proposed to the performance of the p website ingerprinting, and find that the performance achieved by our deep learning approaches is comparable to known methods which include various seconds offerts enomine over multiple with include various seconds. our deep learning approaches is comparable to known memous 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 world of 100 waheiter and 0.4% for our biomeet decord would or one years. The obtained success rate exceeds 96% for a closed world of 900 of 100 websites and 94% for our biggest closed world one of the most nonformant of the most nonformation of 100 websites and 94% for our biggest closed world of 900 classes. In our open world evaluation, then the state of the classes, the control of the control classes. In our open world evaluation, the most performant deep learning model is 2% more accurate than the state-of-the order attack. Eurobaneans we show that the implicit features deep learning model is 2% more accurate than the state-of-the-art attack. Furthermore, we show that the implicit features the state of the-art attack. Furthermore, we show that the implicit features automatically learned by our approach are far more resilient to automatically learned by our approach are far more resilient to automatically learned by content over time. We conclude that automatically learned by our approach are far more resilient to dynamic changes of web content over time. We conclude that the challest to automatically construct the most colorant traffic dynamic changes of web content over time. We conclude that traffic the ability to automatically construct the most relevant traffic recomittee mokes one feetures and parform accurate traffic recomittee mokes one the ability to automatically construct the most relevant traffic features and perform accurate traffic recognition makes our deen learning based anneach an afficient flexible and robust reatures and perform accurate traffic recognition makes our deep learning based approach an efficient, flexible and robust technique for website fingermining

technique for website fingerprinting. Souter (Tor) is a communication tool that proors. It is an actively developed

never the origin and destination of a communication at the never the origin and destination of a communication at the same time. Tor's architecture thus prevents ISPs and local same time. for s architecture thus prevents 151's and local network observers from identifying the websites users visit. As a result of previous research on Tor privacy, a serious As a result of previous research on 101 privacy, a sections side-channel of Tor network traffic was revealed that allowed a local adversary to infer which websites were visited by a a local adversary to liner which websites were visited by a particular user [14]. The identifying information leaks from the state of t particular user [14]. The menutying miorination reaso noil the communication's meta-data, more precisely, from the directions and sizes of encrypted network packets. As this side rections and sizes of encrypted network packets. As this successful to the channel information is often unique for a specific website, if channel information is often unique for a specific website, it can be leveraged to form a unique fingerprint, thus allowing can be reveraged to rorm a unique ingerprint, mus allowing network eavesdroppers to reveal which website was visited

based on the traine that it generated.

The feasibility of Website Fingerprinting (WF) attacks on based on the traffic that it generated. The reasonary of weosite rangerprining (Wr) anacks on Tor was assessed in a series of studies [25], [31], [19], [24], for was assessed in a series of studies [23], [31], [19], [24], [32]. In the related works, the attack is treated as a classification of the state o [32]. In the related works, the attack is treated as a classification problem. This problem is solved by, first, manually ncauon problem. This problem is solved by, first, manually engineering features of traffic traces and then classifying these engineering reatures or trainc traces and then classifying these features with state-of-practice machine learning algorithms. reaures with state-of-practice machine learning algorithms.

Proposed approaches have been shown to achieve a classificarroposed approaches have been shown to achieve a classification accuracy of 91-96% correctly recognized websites [30], non accuracy of 91-90% correctly recognized websites 1501, [24], [13] in a set of 100 websites with 100 traces per website. Their works show that finding distinctive features is essential Ineir works snow that finding distinctive features is essential for accurate recognition of websites. Moreover, this tasks can for accurate recognition of weosites, professor, this tasks can be costly for the adversary as he has to keep up with changes oe cosuy tot the auversary as ne mas to keep up with changes introduced in the network protocol [4], [20], [9]. The WF introduced in the network protocol [4], [20], [7]. The WF research community thus far has not investigated the success research community thus far has not investigated the success of an attacker who automates the feature extraction step for or an anacker who automates the feature extraction step for classification. This is the key problem that we address in this

An essential step of traditional machine learning is feature an essential step of traditional maxime reating is a manual process, based on removed to a manual process, and of raw knowledge, to find a representation of raw 4 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 Juarez<sup>§</sup>, Tom Van Goethem\* and Wouter Joosen\*

§imec-COSIC, ESAT, KU Leuven Email: marc.juarez@esat.kuleuven.be

Abstract—Several studies have shown that the network traffic Abstract—Several studies have shown that the network traffic that is generated by a visit to a website over Tor reveals that is generated by a visit to a website over the fining and the state of the s that is generated by a visit to a website over for reveals information specific to the website through the timing and information specific to the website through the timing and sizes of network packets. By capturing traffic traces between sizes of network packets. By capturing traffic encountries and their for network another control of network connections and their for network another control of the packets. Sizes of network packets. By capturing traine traces between users and their Tor entry guard, a network eavesdropper can leverage this meta-date to reveal which website Tor users are users and their Tor entry guard, a network eavesdropper can leverage this meta-data to reveal which website Tor users are relative to the standard of the control of the standard of the stand leverage this meta-data to reveal which website Tor users are visiting. The success of such attacks heavily depends on the profit of the success of the footnote that are used to construct the construction of the footnote that are used to construct the construction of the footnote that are used to construct the construction of the footnote that are used to construct the construction of the construction o visiting. The success of such attacks heavily depends on the particular set of traffic features that are used to construct the particular set of traffic features are manually opinionally there features are manually opinionally. particular set of traffic features that are used to construct the ingerprint. Typically, these features are manually engineered manually engineered manually engineered manually engineered any change introduced to the Tor naturally and as each any change introduced to the Tor naturally and as each any change introduced to the Tor naturally and as each any change introduced to the Torinature. nngerprint. Typically, these features are manually engineered and, as such, any change introduced to the Tor network can and, as such, any change introduced for the construction of the c and, as such, any change introduced to the 10r network can render these carefully constructed features ineffective. In this render these carefully constructed features methodize in this paper, we show that an adversary can automate the feature paper, we show that an adversary can automatically demonstrate the paper, we show that an adversary can automatically demonstrate the feature paper. paper, we show that an adversary can automate the feature regineering process, and thus automatically deanonymize useful for the beautiful too the continuous contraction to the contraction of the contrac engineering process, and thus automatically deanonymize Tor traffic by applying our novel method based on deep learning. We collect a detect commerced of more those million not core traffic by applying our novel method based on deep learning. We collect a dataset comprised of more than three million network than three million network than three million network than the learning dataset of most traffic areas made for the traffic areas made for collect a dataset comprised of more than three million network traces, which is the largest dataset of web traffic ever used for probability for and find that the nonformation ordinary to probability for the proportion of the state of the nonformation of the nonformatio traces, which is the largest dataset of web traffic ever used for website fingerprinting, and find that the performance achieved by website fingerprinting, and find that the performance methods we have been approached to known methods. website fingerprinting, and find that the performance achieved by our deep learning approaches is comparable to known methods united include various sections of forte enomine over multiple with include various sections. our deep learning approaches is comparable to known memous 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 world from the properties and 0.4% for our binness should upon the properties and 0.4%. years. The obtained success rate exceeds 96% for a closed world of 900 of 100 websites and 94% for our higgest closed world on 900 of 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 94% for our higgest closed world on 100 websites and 100 websites of 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world of 900 to 100 websites and 94% for our biggest closed world on 900 websites and 94% for our biggest closed world on 900 websites and 94% for our biggest closed world on 900 websites and 94% for our biggest closed world on 900 websites and 94% for our biggest closed world on 900 websites and 94% for our biggest closed world on 900 websites and 94% for our biggest closed world on 900 websites and 900 websites an classes. In our open world evaluation, the most performant deep learning model is 2% more accurate than the state-of-the ort attack. Eurobasement we chan that the implicit features 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 features automatically learned by our approach are far more resilient to automatically learned by our approach are far more resilient to automatically learned by content over time. We conclude that automatically learned by our approach are far more resilient to dynamic changes of web content over time. We conclude that the challest to automatically construct the most volumest traffic dynamic changes of web content over time. We conclude that traffic the ability to automatically construct the most relevant traffic recomittee mokes one feetures and parform accurate traffic recomittee mokes one the ability to automatically construct the most relevant traffic features and perform accurate traffic recognition makes our features and perform accurate traffic recognition makes on a feature to the reatures and perform accurate traffic recognition makes our deep learning based approach an efficient, flexible and robust technique for website financiarism

technique for website fingerprinting. Souter (Tor) is a communication tool that proors. It is an actively developed

never the origin and destination of a communication at the never the origin and destination of a communication at the same time. Tor's architecture thus prevents ISPs and local same time. For s architecture thus prevents 1575 and local network observers from identifying the websites users visit. As a result of previous research on Tor privacy, a serious As a result of previous research on 101 privacy, a serious side-channel of Tor network traffic was revealed that allowed succenanies of for network traine was revealed that allowed a local adversary to infer which websites were visited by a a local adversary to liner which websites were visited by a particular user [14]. The identifying information leaks from the state of t particular user [14]. The menutying morniation reaso non-the communication's meta-data, more precisely, from the directions and sizes of encrypted network packets. As this side recuons and sizes of encrypted network packets. As this side-channel information is often unique for a specific website, it channel miormation is often unique for a specific website, it can be leveraged to form a unique fingerprint, thus allowing can be reveraged to rorm a unique ingerprint, mus allowing network eavesdroppers to reveal which website was visited

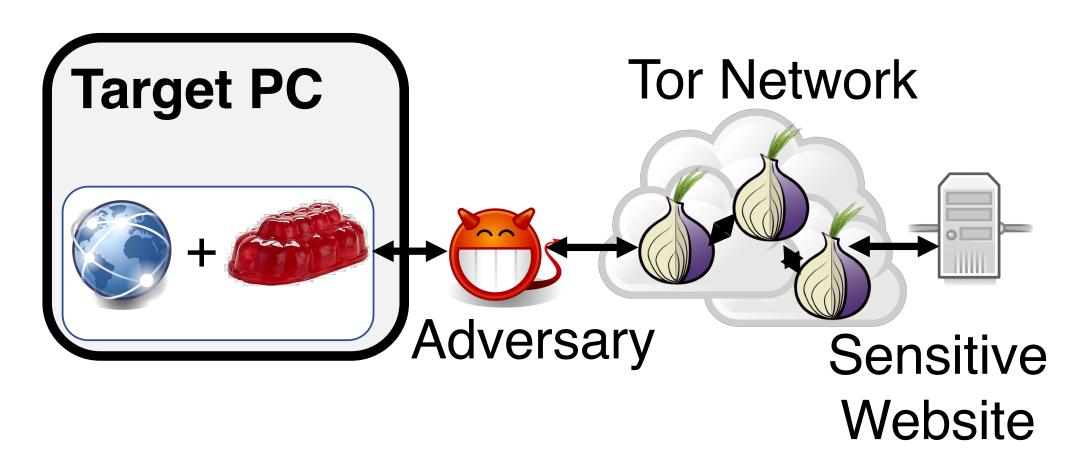
based on the traine that it generated.

The feasibility of Website Fingerprinting (WF) attacks on based on the traffic that it generated. The reasibility of website ringerprining (wr) anacks on Tor was assessed in a series of studies [25], [31], [19], [24], for was assessed in a series of studies [23], [31], [19], [24], [32]. In the related works, the attack is treated as a classification of the state o [32]. In the related works, the attack is treated as a classification problem. This problem is solved by, first, manually ncauon problem. This problem is solved by, first, manually engineering features of traffic traces and then classifying these engineering reatures or trainc traces and then classifying these features with state-of-practice machine learning algorithms. Proposed approaches have been shown to achieve a classificarroposed approaches have been shown to achieve a classification accuracy of 91-96% correctly recognized websites [30], non accuracy of 91-90% correctly recognized websites 1501, [24], [13] in a set of 100 websites with 100 traces per website. Their works show that finding distinctive features is essential Ineir works snow that finding distinctive features is essential for accurate recognition of websites. Moreover, this tasks can for accurate recognition of weosites, professional transfer in the save sail be costly for the adversary as he has to keep up with changes oe costly for the adversary as ne has to keep up with changes introduced in the network protocol [4], [20], [9]. The WF introduced in the network protocol [4], [20], [9]. The Wresearch community thus far has not investigated the success research community mus far has not investigated the success of an attacker who automates the feature extraction step for or an anacker with automates the reading extraction step for classification. This is the key problem that we address in this An essential step of traditional machine learning is feature

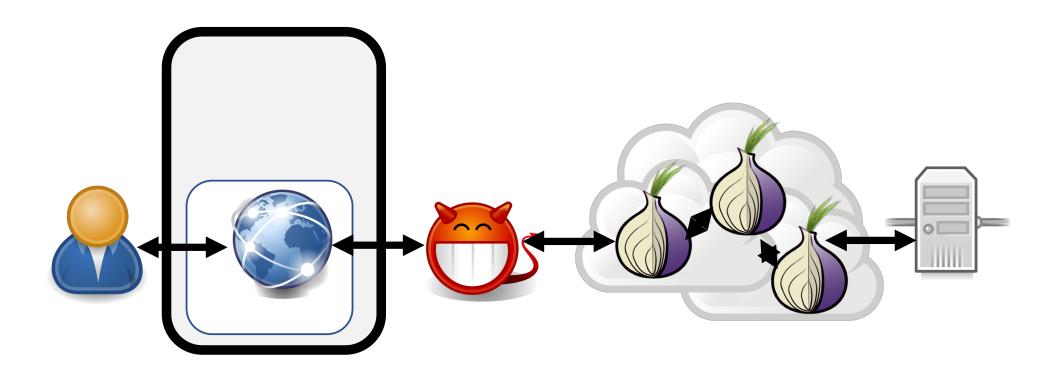
an essential step of traditional maxime reating is a manual process, based on removed to a manual process, and of raw knowledge, to find a representation of raw 4 to be even more

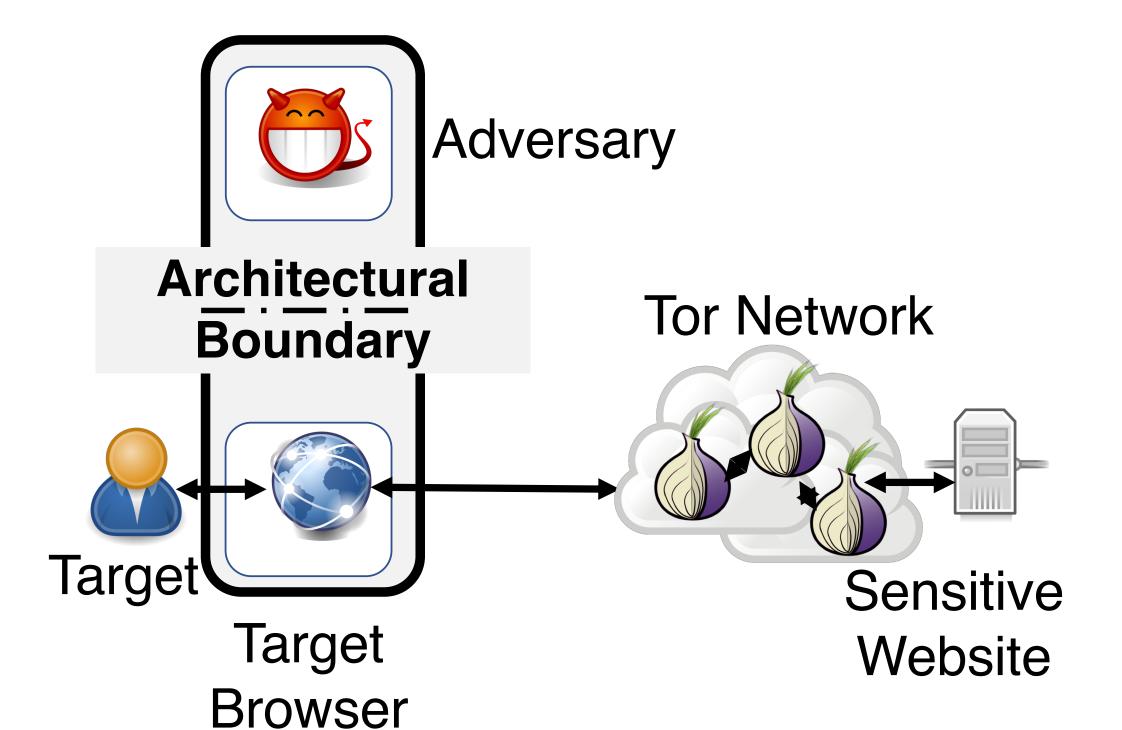
- Main metric is accuracy
- Closed World vs Open World
- Base rate is important!
- Network based WF has >90% accuracy

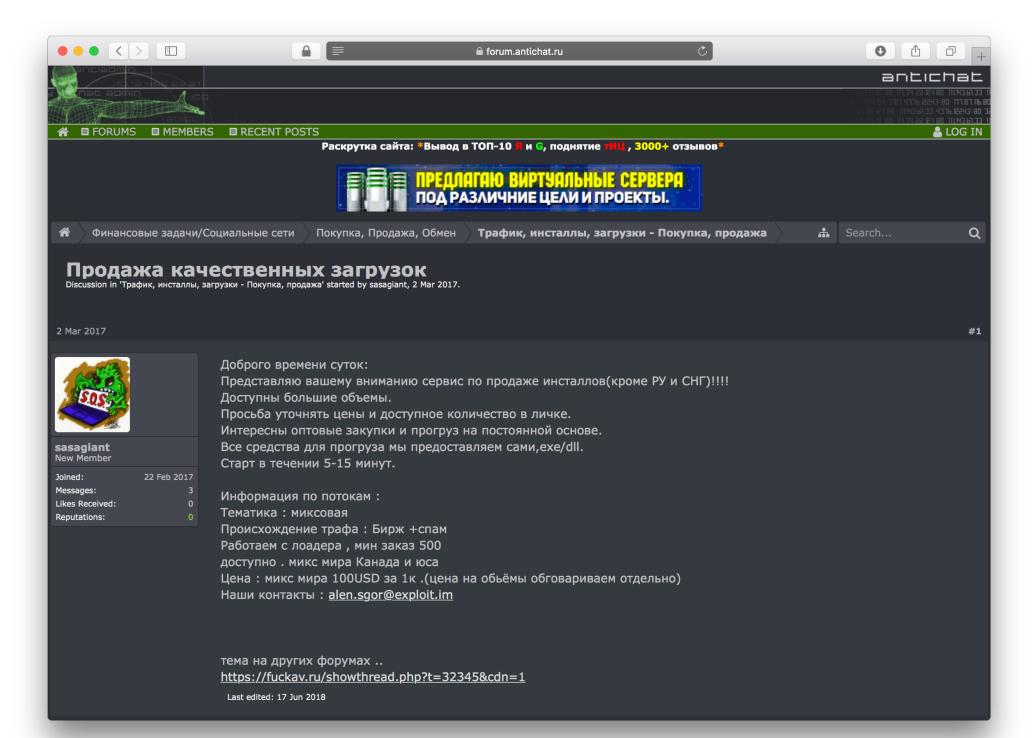
## Traffic Moulding Defenses against WF



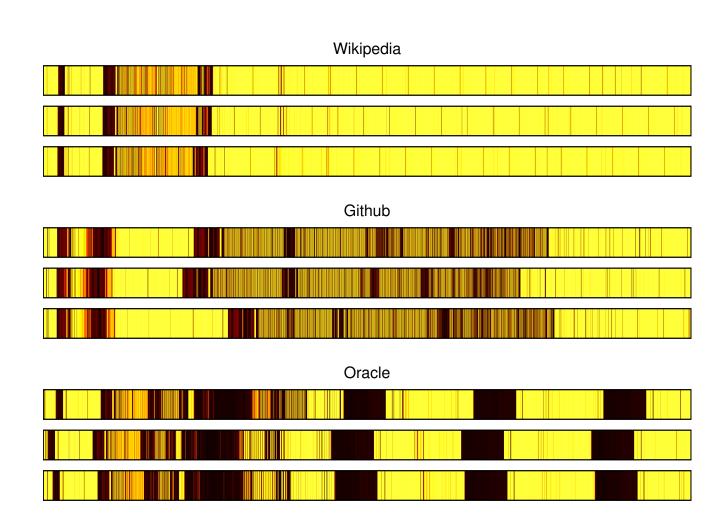
Source: lakeland.co.uk







### Memorygrams



### Cache-Based WF

- 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	

#### Conclusions

- Side-channel attacks can attack human secrets, not just cryptographic secrets
- Specifically, cache-based website fingerprinting is feasible and very dangerous to user privacy
- What other secrets can we attack?
- What kind of countermeasures apply here?
- Why aren't you joining my lab?

https://iss.oy.ne.ro/Lab

