Feature Selection in Website Fingerprinting

Junhua Yan Advisor: Prof. Jasleen Kaur July 24, 2019



Website Fingerprinting

Goal: determine the visited website by inspecting network traffic on client side





Website Fingerprinting

Goal: determine the visited website by inspecting network traffic on client side

Application:

- network manager: protect enterprise networks
- Internet Service Providers: gauge user interests
- malicious entities: exploit private user data

•





Website Fingerprinting

Goal: determine the visited website by inspecting network traffic on client side

Application:

- network manager: protect enterprise networks
- Internet Service Providers: gauge user interests
- malicious entities: exploit private user data

•





Deep Packet Inspection

70	72	74	2e	32	31	36	20	48	54	54	50	2f	31	2e	31	prt.216	HTTP/1.1
ðd	0a	48	6f	73	74	3a	20	77	77	77	2e	67	6f	6f	67	Host:	www.goog
5c	65	2e	63	6f	6d	Ød	Øa	55	73	65	72	2d	41	67	65	le.com	User-Age
5e	74	3a	20	4d	6f	7a	69	6c	6c	61	2f	35	2e	30	20	nt: Mozi	lla/5.0
28	41	6e	64	72	6f	69	64	20	36	2e	30	2e	31	3b	20	(Android	6.0.1;
54	61	62	6c	65	74	3b	20	72	76	3a	35	36	2e	30	29	Tablet;	rv:56.0)
20	47	65	63	6b	6f	2f	35	36	2e	30	20	46	69	72	65	Gecko/5	6.0 Fire
56	6f	78	2f	35	36	2e	30	Ød	Øa	41	63	63	65	70	74	fox/56.0	Accept
3a	20	74	65	78	74	2f	68	74	6d	6c	2c	61	70	70	6c	: text/h	tml,appl
50	67	61	74	60	64	60	34	70	60	74	64	60	24	70	64	ication/	vhtml_vm

Figure: Unencrypted payload over HTTP





Deep Packet Inspection

eb 53 17 03 03 00 ac 00 00 00 00 00 00 00 01 a3 . **S**. 99 71 07 49 71 49 62 49 01 43 f2 9b c4 fe db 37 .g.IgIbI .C....7 da a4 f5 33 0b ab 0f df 55 2c 4f 70 37 ea 8c f7 26 a0 56 f9 05 2e 0c 1b 39 0c 66 db a6 42 ..&.V... ..9.f..B cc 20 5e 60 0a f6 31 3b fc 81 cf 09 7c f5 13 32 0c 2b 54 9a Øc bb ed c2 e1 7b 1e ec 45 d6 2c 4a .+T..... .{..E... bf 64 65 ec a2 2e 42 d5 69 b7 b0 fb 82 b1 47 dc .de...B. i....G. 3c e1 b5 2e 49 ea 1e 1e 71 7b 20 d9 6c 4c 42 .<...I.. .a{ .lLB c3 66 79 bc a4 1a e2 7f a9 fe 2e a9 2c 7e c4 8e .fy.....,~.. 9e 6f ba bc ff 3a 82 a6 d5 5c f3 d3 22 cd fb 1e 4f 31 b2 1f 60 52 18 0e d3 ca db 7a a5 12 d3 7a 01...`R..z...z f8 68 a2 .h.

Figure: Encrypted payload over HTTPS







- TCP/IP signature-based identification

 - Extract features from TCP/IP headers Apply supervised machine learning algorithm •





TCP/IP Header Field	Function
Total Length	Total length of IP datagram
Source	The IP address of the original
Address	source of the IP datagram
Destination	The IP address of the final
Address	destination of the IP datagram
Source Port	TCP port of sending host
Destination Port	TCP port of Destination host

Deep Packet Inspection

2 TCP/IP signature-based identification

Table: Five key fields in TCP/IP header.

- Extract features from TCP/IP headers
- Apply supervised machine learning algorithm





Author	Scenario	Features	Classifier
Liberatore et al. 2006 (L)	SSH	packet size count	Naive Bayes
Herrmann et al. 2009 (H)	SSH, Tor	packet size frequency	Multinomial Bayes
Panchenko et al. 2011 (P)	SSH, Tor	burst markers, HTML markers, # of markers, ratio of incoming	
		packets, occurring packet sizes, transmitted bytes, # of packets	SVM
Dyer et al. 2012 (Vng++)	SSH	per-direction bandwidth, transmission time, burst markers	Naive Bayes
Wang et al. 2013 (FLSVM)	Tor	Tor cell instances	Distance-based SVM
Feghhi et al. 2016 (DTW)	SSH	uplink timing information	Dynamic Time Warping
Panchenko et al. 2016	Tor	# of incoming & outgoing packets, sum of incoming	
(CUMUL)		& outgoing packet sizes, interpolant of cumulative packet size	SVM
		# of packets, ratio of incoming & outgoing packets ,	
Hayes et al. 2016 (k-FP)	Tor	packet ordering, concentration of outgoing packets, # of	Random Forests
		packets per second, inter-arrival time, transmission time	
Trevisan et al. 2016 (T)	HTTP	server IP address count, hostname count	*

Table: Summary of prior work evaluated in our work.



Author	Scenario	Features	Classifier
Liberatore et al. 2006 (L)	SSH	packet size count	Naive Bayes
Herrmann et al. 2009 (H)	SSH, Tor	packet size frequency	Multinomial Bayes
Panchenko et al. 2011 (P)	SSH, Tor	burst markers, HTML markers, # of markers, ratio of incoming	
		packets, occurring packet sizes, transmitted bytes, # of packets	SVM
Dyer et al. 2012 (Vng++)	SSH	per-direction bandwidth, transmission time, burst markers	Naive Bayes
Wang et al. 2013 (FLSVM)	Tor	Tor cell instances	Distance-based SVM
Feghhi et al. 2016 (DTW)	SSH	uplink timing information	Dynamic Time Warping
Panchenko et al. 2016	Tor	# of incoming & outgoing packets, sum of incoming	
(CUMUL)		& outgoing packet sizes, interpolant of cumulative packet size	SVM
		# of packets, ratio of incoming & outgoing packets,	
Hayes et al. 2016 (k-FP)	Tor	packet ordering, concentration of outgoing packets, # of	Random Forests
		packets per second, inter-arrival time, transmission time	
Trevisan et al. 2016 (T)	HTTP	server IP address count, hostname count	*

Table: Summary of prior work evaluated in our work.



Author	Scenario	Features	Classifier
Liberatore et al. 2006 (L)	SSH	packet size count	Naive Bayes
Herrmann et al. 2009 (H)	SSH, Tor	packet size frequency	Multinomial Bayes
Panchenko et al. 2011 (P)	SSH, Tor	burst markers, HTML markers, # of markers, ratio of incoming	
		packets, occurring packet sizes, transmitted bytes, # of packets	SVM
Dyer et al. 2012 (Vng++)	SSH	per-direction bandwidth, transmission time, burst markers	Naive Bayes
Wang et al. 2013 (FLSVM)	Tor	Tor cell instances	Distance-based SVM
Feghhi et al. 2016 (DTW)	SSH	uplink timing information	Dynamic Time Warping
Panchenko et al. 2016	Tor	# of incoming & outgoing packets, sum of incoming	
(CUMUL)		& outgoing packet sizes, interpolant of cumulative packet size	SVM
		# of packets, ratio of incoming & outgoing packets ,	
Hayes et al. 2016 (k-FP)	Tor	packet ordering, concentration of outgoing packets, # of	Random Forests
		packets per second, inter-arrival time, transmission time	
Trevisan et al. 2016 (T)	HTTP	server IP address count, hostname count	*

Table: Summary of prior work evaluated in our work.



Author	Scenario	Features	Classifier
Liberatore et al. 2006 (L)	SSH	packet size count	Naive Bayes
Herrmann et al. 2009 (H)	SSH, Tor	packet size frequency	Multinomial Bayes
Panchenko et al. 2011 (P)	SSH, Tor	burst markers, HTML markers, # of markers, ratio of incoming	
		packets, occurring packet sizes, transmitted bytes, # of packets	SVM
Dyer et al. 2012 (Vng++)	SSH	per-direction bandwidth, transmission time, burst markers	Naive Bayes
Wang et al. 2013 (FLSVM)	Tor	Tor cell instances	Distance-based SVM
Feghhi et al. 2016 (DTW)	SSH	uplink timing information	Dynamic Time Warping
Panchenko et al. 2016	Tor	# of incoming & outgoing packets, sum of incoming	
(CUMUL)		& outgoing packet sizes, interpolant of cumulative packet size	SVM
		# of packets, ratio of incoming & outgoing packets ,	
Hayes et al. 2016 (k-FP)	Tor	packet ordering, concentration of outgoing packets, # of	Random Forests
		packets per second, inter-arrival time, transmission time	
Trevisan et al. 2016 (T)	HTTP	server IP address count, hostname count	*

Table: Summary of prior work evaluated in our work.

• Limited set of features studied



Author	Scenario	Features	Classifier
Liberatore et al. 2006 (L)	SSH	packet size count	Naive Bayes
Herrmann et al. 2009 (H)	SSH, Tor	packet size frequency	Multinomial Bayes
Panchenko et al. 2011 (P)	SSH, Tor	burst markers, HTML markers, # of markers, ratio of incoming	
		packets, occurring packet sizes, transmitted bytes, # of packets	SVM
Dyer et al. 2012 (Vng++)	SSH	per-direction bandwidth, transmission time, burst markers	Naive Bayes
Wang et al. 2013 (FLSVM)	Tor	Tor cell instances	Distance-based SVM
Feghhi et al. 2016 (DTW)	SSH	uplink timing information	Dynamic Time Warping
Panchenko et al. 2016	Tor	# of incoming & outgoing packets, sum of incoming	
(CUMUL)		& outgoing packet sizes, interpolant of cumulative packet size	SVM
		# of packets, ratio of incoming & outgoing packets ,	
Hayes et al. 2016 (k-FP)	Tor	packet ordering, concentration of outgoing packets, # of	Random Forests
		packets per second, inter-arrival time, transmission time	
Trevisan et al. 2016 (T)	HTTP	server IP address count, hostname count	*

Table: Summary of prior work evaluated in our work.

• Limited set of features studied

What's the extent of website fingerprint-ability?



• Are there other features can be used to achieve comparable accuracy with state-of-the-art?

• What if we hide some of informative features, e.g., packet size?

• *Can features that are informative in one scenario (e.g., Tor) be used to accurately identify websites in another scenario (e.g., SSH)?*



• *Are there other features can be used to achieve comparable accuracy with state-of-the-art?*

• Extract a comprehensive list of TCP/IP header features

• What if we hide some of informative features, e.g., packet size?

• *Can features that are informative in one scenario (e.g., Tor) be used to accurately identify websites in another scenario (e.g., SSH)?*



• *Are there other features can be used to achieve comparable accuracy with state-of-the-art?*

• Extract a comprehensive list of TCP/IP header features

• What if we hide some of informative features, e.g., packet size?

• Consider eight different communication scenarios

• *Can features that are informative in one scenario (e.g., Tor) be used to accurately identify websites in another scenario (e.g., SSH)?*



• *Are there other features can be used to achieve comparable accuracy with state-of-the-art?*

• Extract a comprehensive list of TCP/IP header features

• What if we hide some of informative features, e.g., packet size?

• Consider eight different communication scenarios

• *Can features that are informative in one scenario (e.g., Tor) be used to accurately identify websites in another scenario (e.g., SSH)?*

• Identify and analyze importance of features in each scenario





- Packet direction
 - Outgoing Incoming
- Packet length: length of rectangle
- Packet timestamp
- TCP connection : (IP address, port number)





- Packet-level
 - *e.g.*, # of incoming packets, packet size count, total incoming bytes, ...





- Packet-level
- Burst-level
 - **Burst**: a sequence of packets sent in one direction between two packets sent in the opposite direction
 - *e.g.*, packet seq.: (10, 10, -10, -10, 10) → burst seq.: (20, -20, 10)
 - *e.g.*, # of incoming bursts, burst size count,...





- Packet-level
- Burst-level
- TCP-level
 - *e.g.*, average # of incoming packets/TCP conn., average incoming bytes/TCP conn., ...





- Packet-level
- Burst-level
- TCP-level
- Port-level
 - TCP connections with different port numbers
 - e.g., average # of incoming packets sent over 443, ...
- IP address-level
 - TCP connections with different IP addresses
 - e.g., average incoming bytes transmitted with 216.58.217, ...





- Packet-level
- Burst-level
- TCP-level
- Port-level
- IP address-level

109 feature categories, \sim 35,683 features

** 61 feature categories have never been considered before











	Packet Direction	Packet Length	Packet Time	IP Address	Port/ TCP
HTTPx	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark
Anonymized IP address	\checkmark	\checkmark	\checkmark		\checkmark
SSH/VPN	\checkmark	\checkmark	\checkmark		



	Packet Direction	Packet Length	Packet Time	IP Address	Port/ TCP
HTTPx	\checkmark	, √	\checkmark	\checkmark	\checkmark
Anonymized IP address	\checkmark	\checkmark	\checkmark		\checkmark
SSH/VPN	\checkmark	\checkmark	\checkmark		
HTTPx + PadToMTU	\checkmark		\checkmark	\checkmark	\checkmark
Tor	\checkmark		\checkmark		



	Packet	Packet	Packet	IP	Port/
	Direction	Length	lime	Address	TCP
HTTPx	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark
Anonymized IP address	\checkmark	\checkmark	\checkmark		\checkmark
SSH/VPN	\checkmark	\checkmark	\checkmark		
HTTPx + PadToMTU	\checkmark		\checkmark	\checkmark	\checkmark
Tor	\checkmark		\checkmark		
Tor + Fixed Inter-arrival Time	\checkmark				



	Packet	Packet	Packet	IP	Port/
	Direction	Length	Time	Address	ТСР
HTTPx	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark
Anonymized IP address	\checkmark	\checkmark	\checkmark		\checkmark
SSH/VPN	\checkmark	\checkmark	\checkmark		
HTTPx + PadToMTU	\checkmark		\checkmark	\checkmark	\checkmark
Tor	\checkmark		\checkmark		
Tor + Fixed Inter-arrival Time	\checkmark				
HTTPx + Incoming Packets Only	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark
HTTPx + Outgoing Packets Only	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark



	Packet	Packet	Packet	IP	Port/
	Direction	Length	Time	Address	TCP
HTTPx	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark
Anonymized IP address	\checkmark	\checkmark	\checkmark		\checkmark
SSH/VPN	\checkmark	\checkmark	\checkmark		
HTTPx + PadToMTU	\checkmark		\checkmark	\checkmark	\checkmark
Tor	\checkmark		\checkmark		
Tor + Fixed Inter-arrival Time	\checkmark				
HTTPx + Incoming Packets Only	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark
HTTPx + Outgoing Packets Only	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark

Table: Information available in each scenario.

• Tor

- Murdoch et al. 2005, Panchenko et al. 2011, Yu et al. 2012, Cai et al. 2012, Wang et al. 2013, Wang et al. 2014, Panchenko et al. 2016, Abe et al. 2016, Rimmer et al. 2017
- SSH/VPN
 - Bissias et al. 2005, Liberatore et al. 2006, Herrmann et al. 2009, Lu et al. 2010, Panchenko et al. 2011, Dyer et al. 2012, Feghhi et al. 2016,

• HTTPx

• Sun et al. 2002, Gong et al. 2010, Maciá-Fernández et al. 2010, Miller et al. 2014, Trevisan et al. 2016,



Goal: select informative features in each scenario *Criterion*: *Mean Decrease Impurity (MDI) Importance* derived from decision tree-based ensemble methods

• *Key Idea*: compute *the average decrease of entropy* of each feature in multiple decision trees to measure their importance



• MDI Importance is biased with correlated features



Figure: Bias with correlated features on MDI importance.



• MDI Importance is biased with correlated features



Figure: Bias with correlated features on MDI importance.



• MDI Importance is biased with correlated features



Figure: Bias with correlated features on MDI importance.

- Solution
 - Cluster correlated features
 - 2 Choose one from each cluster as a representative
 - **③** Calculate *MDI Importance*



• MDI Importance is biased with correlated features



Figure: Bias with correlated features on MDI importance.

- Solution
 - 1 Cluster correlated features

Complexity: $O(n^2)$ *HTTPx*: $n \approx 36,000$

- 2 Choose one from each cluster as a representative
- **6** Calculate *MDI Importance*



Feature Selection Methodology



1. Reduce number of features

- Calculate MDI importance
- Filter out less important features
 - consider top *n* features that contribute to 99% of the total *MDI importance*
 - e.g., 35,711 → 5,852



Feature Selection Methodology



1. Reduce number of features

2. Remove correlated features

- Perform hierarchical clustering based on *Euclidean distance*
- Determine number of clusters based on silhouette scores
- Select one feature from each cluster
 - *e.g.*, $5,852 \rightarrow 2,512$



Feature Selection Methodology





Dataset

• Our Dataset

- visit 3,000 websites listed in Alexa each 20 times with *Google Chrome Version* 61.0.3163.100
- 2,032 websites, each with at least 16 visits



Dataset

• Our Dataset

- visit 3,000 websites listed in Alexa each 20 times with *Google Chrome Version* 61.0.3163.100
- 2,032 websites, each with at least 16 visits
- Two other public datasets
 - SSH2000 Dataset [Liberatore et al. 2006]
 - 2,000 websites, each is visited 51 time over SSH
 - Tor Dataset [Wang et al. 2013]
 - 100 websites, each is visited 90 times with Tor browser



Evaluation Methodology

- 1 Select informative features in each scenario
- Output Compare classification accuracy with feature sets proposed in previous work



Selected Features in *Tor* Overview

- TCP/IP header information:
 - packet direction and timestamp
- *No.* of feature categories:
 - new/all: 15/38
- *Sum.* of importance:
 - new/all: 22.18/100

1	preposition of first 300 incoming packets	24.039
2	concentration of outgoing packets in first 2,000 packets	7.417
3	initial 30 incoming packets	5.906
4	alternative concentration of outgoing packets	5.673
5	** cumulative size with direction of first 100 packets	5.65
6	initial 30 packets	5.611
7	position of first 300 outgoing packets	5.424
8	position of first 300 incoming packets	4.413
9	initial 30 outgoing packets	4.197
10	preposition of first 300 outgoing packets	4.196
11	** inter-arrival time of first 20 packets	2.38
12	unique burst size	1.978
13	** inter-arrival time of first 20 incoming packets	1.896
14	** inter-arrival time of first 20 outgoing packets	1.824
15	** initial 30 outgoing bursts	1.761
16	** initial 30 bursts	1.3
17	number of outgoing packets per second	1.205
18	** # of packets in incoming burst count	1.163
19	** # of packets in a burst count	1.108
20	alternative outgoing packets per second	0.934
21	** outgoing burst duration	0.878
22	# of outgoing packets per TCP conn.	0.864
23	** initial 30 incoming bursts	0.862
24	ratio of incoming packets # per TCP conn.	0.842
25	concentration of first 30 outgoing packets	0.815
26	** burst duration	0.812
27	burst size count	0.785
28	** # of packets in outgoing burst	0.65
29	size of incoming bursts	0.591
30	alternative packets per second	0.558
31	concentration of last 30 incoming packets	0.463
32	interpolant of cumulative packet size	0.438
33	** # of packets in each burst	0.432
34	concentration of last 30 outgoing packets	0.428
35	number of packets per second	0.428
36	number of incoming packets per second	0.372
37	** # of packets in outgoing burst count	0.358
38	** incoming burst duration	0.34

Table: Most informative features in *Tor*. 13/23



Selected Features in Tor **Example** features



Cumulative packet size with direction

captures incoming/outgoing packet ordering •

	1	preposition of first 300 incoming packets	24.039
	2	concentration of outgoing packets in first 2,000 packets	7.417
	3	initial 30 incoming packets	5.906
	4	alternative concentration of outgoing packets	5.673
	5	** cumulative size with direction of first 100 packets	5.65
	6	initial 30 packets	5.611
	7	position of first 300 outgoing packets	5.424
	8	position of first 300 incoming packets	4.413
	9	initial 30 outgoing packets	4.197
	10	preposition of first 300 outgoing packets	4.196
	11	** inter-arrival time of first 20 packets	2.38
	12	unique burst size	1.978
	13	** inter-arrival time of first 20 incoming packets	1.896
	14	** inter-arrival time of first 20 outgoing packets	1.824
	15	** initial 30 outgoing bursts	1.761
	16	** initial 30 bursts	1.3
	17	number of outgoing packets per second	1.205
	18	** # of packets in incoming burst count	1.163
	19	** # of packets in a burst count	1.108
	20	alternative outgoing packets per second	0.934
	21	** outgoing burst duration	0.878
-	22	# of outgoing packets per TCP conn.	0.864
,	23	** initial 30 incoming bursts	0.862
	24	ratio of incoming packets # per TCP conn.	0.842
	25	concentration of first 30 outgoing packets	0.815
	26	** burst duration	0.812
	27	burst size count	0.785
	28	** # of packets in outgoing burst	0.65
	29	size of incoming bursts	0.591
	30	alternative packets per second	0.558
	31	concentration of last 30 incoming packets	0.463
	32	interpolant of cumulative packet size	0.438
	33	** # of packets in each burst	0.432
	34	concentration of last 30 outgoing packets	0.428
	35	number of packets per second	0.428
	36	number of incoming packets per second	0.372
	37	** # of packets in outgoing burst count	0.358
	38	** incoming burst duration	0.34

Table: Most informative features in *Tor*.



Selected Features in Tor **Example** features



Inter-arrival time between packets

interleaving of packets from parallel TCP connections

1	preposition of first 300 incoming packets	24.039
2	concentration of outgoing packets in first 2,000 packets	7.417
3	initial 30 incoming packets	5.906
4	alternative concentration of outgoing packets	5.673
5	** cumulative size with direction of first 100 packets	5.65
6	initial 30 packets	5.611
7	position of first 300 outgoing packets	5.424
8	position of first 300 incoming packets	4.413
9	initial 30 outgoing packets	4.197
10	preposition of first 300 outgoing packets	4.196
11	** inter-arrival time of first 20 packets	2.38
12	unique burst size	1.978
13	** inter-arrival time of first 20 incoming packets	1.896
14	** inter-arrival time of first 20 outgoing packets	1.824
15	** initial 30 outgoing bursts	1.761
16	** initial 30 bursts	1.3
17	number of outgoing packets per second	1.205
18	** # of packets in incoming burst count	1.163
19	** # of packets in a burst count	1.108
20	alternative outgoing packets per second	0.934
21	** outgoing burst duration	0.878
22	# of outgoing packets per TCP conn.	0.864
23	** initial 30 incoming bursts	0.862
24	ratio of incoming packets # per TCP conn.	0.842
25	concentration of first 30 outgoing packets	0.815
26	** burst duration	0.812
27	burst size count	0.785
28	** # of packets in outgoing burst	0.65
29	size of incoming bursts	0.591
30	alternative packets per second	0.558
31	concentration of last 30 incoming packets	0.463
32	interpolant of cumulative packet size	0.438
33	** # of packets in each burst	0.432
34	concentration of last 30 outgoing packets	0.428
35	number of packets per second	0.428
36	number of incoming packets per second	0.372
37	** # of packets in outgoing burst count	0.358
38	** incoming burst duration	0.34

Table: Most informative features in *Tor*.



Classification Performance in Tor

- Features:
 - Informative features identified in Tor (Ours)
 - Eight feature sets proposed in previous research
- Classifier: Extra-Trees, 10-fold validation
- 2,000 websites, each with 16 instances



Classification Performance in Tor

- Features:
 - Informative features identified in Tor (Ours)
 - Eight feature sets proposed in previous research
- Classifier: Extra-Trees, 10-fold validation
- 2,000 websites, each with 16 instances
 - Our dataset: 82.78 vs. 96.83





Classification Performance in Tor

- Features:
 - Informative features identified in Tor (Ours)
 - Eight feature sets proposed in previous research
- Classifier: Extra-Trees, 10-fold validation
- 2,000 websites, each with 16 instances
 - Our dataset: 82.78 vs. 96.83
 - SSH2000: 63.13 vs. 80.29





Classification Performance in Other Communication Scenarios







Tor+Fixed Inter-arrival Time







Conclusion

- Extract a comprehensive list of features from TCP/IP headers for website fingerprinting
- Study eight different communication scenarios
- Identify and select informative features in each scenario



Limitation & Future Work

Practical issues in website fingerprinting

- impact of caching
- geographic location
- client browser platform
- network segmentation

• ...



Thanks



Feature Selection Approaches

Filters

- Select features based on their correlation with the predict
- Pearson correlation coefficient, mutual information, ...

Wrappers & Embedded

- Measure the relative usefulness of feature subsets
- Wrappers
 - search the space of all feature subsets
 - forward selection, backward selection, ...
- Embedded
 - search guided by the learning process
 - Decision tree

	Computation Efficiency	Feature Correlation
Filters	\checkmark	X
Wrappers	X	\checkmark
Embedded	\checkmark	\checkmark

 Table: Comparison of feature selection approaches.



• Goal: select informative features in each scenario



- Goal: select informative features in each scenario
- *Criterion: Mean Decrease Impurity (MDI) Importance* derived from decision tree-based ensemble methods



- Goal: select informative features in each scenario
- *Criterion: Mean Decrease Impurity (MDI) Importance* derived from decision tree-based ensemble methods
 - Decision Tree



Figure: A decision tree to differ website *A*, *B* and *C*.



- Goal: select informative features in each scenario
- *Criterion: Mean Decrease Impurity (MDI) Importance* derived from decision tree-based ensemble methods
 - Decision Tree
 - Entropy
 - measure *impurity* based on probability of each possible output

A	B	С	Entropy	
10	0	0	0	
5	5	0	$\frac{1}{2}\log 2 + \frac{1}{2}\log 2 + 0 \approx 0.301$	

Table: Entropy with different probabilities.



Figure: A decision tree to differ website *A*, *B* and *C*.



- Goal: select informative features in each scenario
- *Criterion: Mean Decrease Impurity (MDI) Importance* derived from decision tree-based ensemble methods
 - Decision Tree
 - Entropy
 - Information Gain
 - measure *the total decrease of impurity* when consider a feature as a split node

A	В	Entropy	
5	5	0.301	

	A	B	Entropy
≤ 20	5	0	0
> 20	0	5	0

Information Gain =
$$0.301 - (\frac{5}{10} \times 0 + \frac{5}{10} \times 0) = 0.301$$
 (1)



Figure: A decision tree to differ website *A*, *B* and *C*.



Mutual Information & Gini Index

In Extra-Trees, using mutual information/entropy or gini index as impurity measure has been demonstrated to achieve comparable stability score and performance (Haralampieva and Brown 2016).



Hierachical Clustering

• Average Linkage



Figure: Hierarchical Clustering