#### AdGraph: A Graph-Based Approach to Ad and Tracker Blocking

Umar Iqbal, Peter Snyder, Shitong Zhu, Benjamin Livshits, Zhiyun Qian, and Zubair Shafiq IEEE Symposium on Security and Privacy, 2020



Advertising enables "free" content Publishers show content Earn revenue with ads



Advertising enables "free" content Publishers show content Earn revenue with ads



#### Digital revenues for full year 2018 surpassed \$100 billion for the first time

Interactive Advertising Bureau (IAB) '19

Advertising enables "free" content

Publishers show content

Earn revenue with ads

Problems with online advertising ecosystem

Advertising enables "free" content Publishers show content

rublishers show conten

Earn revenue with ads

Problems with online advertising ecosystem Privacy concerns – Behavioral targeting "I see ads for things I dream about."

"My phone is eavesdropping on me"

Advertising enables "free" content Publishers show content Earn revenue with ads

Problems with online advertising ecosystem Privacy concerns – Behavioral targeting Performance issues – Slow page load

	Seconds to load advertising content			Seconds to load	editorial content
boston.com	30.8 seconds				8.1
₹ <b>`ELAZE</b>	11.9		7.0		
THE DAILY BEAST	11.3		5.0		
😵 INDEPENDENT	10.9	4.	.4		
Chicago Tribune	9.5	3.4			
examiner.com	9.1	2.1			
NEWYORKPOST	8.9			16.9	
SALON	8.8		8.1		
CNN	8.6	5.3			
SFGATE	7.9		10.4		
Los Angeles Times	7.7	3.7			
Mirror	7.2	6.5			
b/r bleacher report	6.7	5.8			
*Chron	6.6	6.1			
Vox	6.3	3.6			
AL	6.1	5.9			
The Telegraph	6.0	3.8			

Advertising enables "free" content

Publishers show content

Earn revenue with ads

Problems with online advertising ecosystem Privacy concerns – Behavioral targeting Performance issues – Slow page load Malvertising

#### Hackers have breached 60 ad servers to load their own malicious ads

Why buy legitimate ad slots to deliver malvertising when you can just hack the server instead.

#### Major sites including New York Times and BBC hit by 'ransomware' malvertising

Adverts hijacked by malicious campaign that demands payment in bitcoin to unlock user computers

Advertising enables "free" content Publishers show content Earn revenue with ads

Problems with online advertising ecosystem Privacy concerns – Behavioral targeting Performance issues – Slow page load Malvertising Intrusive



Advertising enables "free" content Publishers show content Earn revenue with ads

Problems with online advertising ecosystem Privacy concerns – Behavioral targeting Performance issues – Slow page load Malvertising Intrusive

Solution

Ad & tracker blockers



#### Outline

State of Ad/Tracker Blocking Ads & Trackers Filter list blocking Machine learning based blocking

AdGraph

Graph-based representation Machine learning on graph representation Evaluation

#### Outline

State of Ad/Tracker Blocking Ads & Trackers Filter list blocking Machine learning based blocking

AdGraph Graph-based representation Machine learning on graph representation Evaluation



Ads are audio-visual promotional content



Ads are audio-visual promotional content Trackers collect sensitive information



Ads are audio-visual promotional content Trackers collect sensitive information

They are:

Created with JavaScript Requested with HTTP Displayed with HTML

Ads and trackers involve HTML, Network, and JavaScript

```
<script>
    eval("
        var iframe = document.createElement("iframe");
        iframe.src = "adnetwork.com/load_ads";
        document.body.append(iframe);
        ");
</script>
    </script>
        var img = document.createElement("img");
        iframe.src = "adnetwork.com/load_ads";
        </script>
        var img = document.createElement("img");
        iframe.src = "adnetwork.com/load_ads";
        </script>
</iframe>
```



JavaScript

Ads are audio-visual promotional content Trackers collect sensitive information

They are:

Created with JavaScript





Ads are audio-visual promotional content Trackers collect sensitive information

They are:

Created with JavaScript Requested with HTTP





Ads are audio-visual promotional content Trackers collect sensitive information

They are:

Created with JavaScript Requested with HTTP Displayed with HTML





JavaScript

Ads are audio-visual promotional content Trackers collect sensitive information

They are:

Created with JavaScript Requested with HTTP Displayed with HTML

Ads and trackers involve HTML, Network, and JavaScript





Manually curated with crowdsourcing

Manually curated with crowdsourcing Leads to scalability issues



3 months to add new rules [Iqbal et al. '17]

Manually curated with crowdsourcing Leads to scalability issues



3.8 year to remove rules [Snyder et al. '20]

Manually curated with crowdsourcing Leads to scalability issues



90% rules are useless [Snyder et al. '20]

Manually curated with crowdsourcing Leads to scalability issues

Operate at HTML/Network/JS layer in isolation



Manually curated with crowdsourcing Leads to scalability issues

Operate at HTML/Network/JS layer in isolation Leads to accuracy issues

Manually curated with crowdsourcing Leads to scalability issues

Operate at HTML/Network/JS layer in isolation Leads to accuracy issues



If You Have Dark Spo This Immediately (It's

This simple \$50 lighting trick will make your kitchen look stunning. (See pics)

Manually curated with crowdsourcing Leads to scalability issues

Operate at HTML/Network/JS layer in isolation Leads to accuracy issues



This simple \$50 lighting trick will make your kitchen look stunning. (See pics)

Manually curated with crowdsourcing Leads to scalability issues

Operate at HTML/Network/JS layer in isolation Leads to accuracy issues



This simple \$50 lighting trick will make your kitchen look stunning. (See pics)





Network layer [Bhagavatula et al. 14, Gugelmann et al. '15] HTTP header properties as features presence of words like "ad" cookies set by response

Network layer [Bhagavatula et al. 14, Gugelmann et al. '15] HTTP header properties as features presence of words like "ad" cookies set by response

JavaScript layer [Wu et al. '16, Ikram et al. '17] JS API names as features document.cookie element.clientWidth

Network layer [Bhagavatula et al. 14, Gugelmann et al. '15]



JavaScript layer [Wu et al. '16, Ikram et al. '17] JS API names as features document.cookie element.clientWidth

Network layer [Bhagavatula et al. 14, Gugelmann et al. '15]





#### Outline

State of Ad/Tracker Blocking Online advertising Filter list blocking Machine learning based blocking

AdGraph

Graph-based representation Machine learning on graph representation Evaluation

#### AdGraph



#### AdGraph

Graph-based cross-layer representation of ad/tracker behavior


Graph-based cross-layer representation of ad/tracker behavior



Graph-based cross-layer representation of ad/tracker behavior

### ML to automatically learn ad/tracker behavior



Chromium instrumentation

### Graph-based cross-layer representation of ad/tracker behavior



### Graph-based cross-layer representation of ad/tracker behavior



### Graph-based cross-layer representation of ad/tracker behavior



Graph-based cross-layer representation of ad/tracker behavior



<script>
 eval("
 var iframe = document.createElement("iframe");
 iframe.src = "adnetwork.com/load\_ads";
 document.body.append(iframe);
 ");
</script>
 <iframe src = "adnetwork.com/load\_ads">
 <script>
 <script>
 var img = document.createElement("img");
 iframe.src = "adnetwork.com/load\_ads";
 </script>
</iframe>



Network Request



Cross-layer interactions

# <script> eval(" var iframe = document.createElement("iframe"); iframe.src = "adnetwork.com/load\_ads"; document.body.append(iframe); "); </script> <iframe src = "adnetwork.com/load\_ads"> <script> <script> var img = document.createElement("img"); iframe.src = "adnetwork.com/load\_ads"> <script> // iframe.src = "adnetwork.com/load\_ads"> </script> // iframe.src = "adnetwork.com/load\_ads"> // iframe.src = "adnetwork.com/load\_ads"; // iframe>



Cross-layer interactions

JS (element)  $\rightarrow$  Network (request)

```
<script>
    eval("
        var iframe = document.createElement("iframe");
        iframe.src = "adnetwork.com/load_ads";
        document.body.append(iframe);
        ");
</script>
    </script>
        <script>
            <script>
            var img = document.createElement("img");
            iframe.src = "adnetwork.com/load_ads";
            <script>
            var img = document.createElement("img");
            iframe.src = "adnetwork.com/load_ads";
            </script>
        <//iframe>
```



**Cross-layer** interactions

JS (element)  $\rightarrow$  Network (request) Network (request)  $\rightarrow$  HTML (response)

```
<script>
    eval("
        var iframe = document.createElement("iframe");
        iframe.src = "adnetwork.com/load_ads";
        document.body.append(iframe);
        ");
</script>
    </script>
        <script>
            <script>
            var img = document.createElement("img");
            iframe.src = "adnetwork.com/load_ads";
            <script>
            var img = document.createElement("img");
            iframe.src = "adnetwork.com/load_ads";
            </script>
        <//iframe>
```



Cross-layer interactions

JS (element)  $\rightarrow$  Network (request) Network (request)  $\rightarrow$  HTML (response)

Building cross-layer context





Cross-layer interactions

JS (element)  $\rightarrow$  Network (request) Network (request)  $\rightarrow$  HTML (response)

Building cross-layer context Easy to link Network with HTML





Cross-layer interactions

JS (element)  $\rightarrow$  Network (request) Network (request)  $\rightarrow$  HTML (response)

Building cross-layer context Easy to link Network with HTML JavaScript activity attribution is tricky





### JavaScript Attribution

No API to attribute JavaScript to HTML and Network requests

# JavaScript Attribution

No API to attribute JavaScript to HTML and Network requests

Stack Walking [Privacy Badger, OpenWPM] Look at stack at points of interest Incomplete and evadable e.g. eval, inline scripts



### JavaScript Attribution

No API to attribute JavaScript to HTML and Network requests

Stack Walking [Privacy Badger, OpenWPM] Look at stack at points of interest Incomplete and evadable e.g. eval, inline scripts Image: Source of the second secon

Browser Instrumentation [JSGraph '18] Capture events as scripts execute Detailed cross-layer interaction

#### JSgraph: Enabling Reconstruction of Web Attacks via Efficient Tracking of Live In-Browser JavaScript Executions

Display Market M

Bo Li, Phani Vadrevu, Kyu Hyung Lee, and Roberto Perdisci Department of Computer Science, University of Georgia {bo,vadrevu,khlee,perdisci}@cs.uga.edu

Abstract—In this paper, we propose JSgraph, a forensic engine that is able to efficiently record fine-grained details pertaining to the execution of JavaScript (JS) programs within the browser,

on JS-driven DOM modifications. Ultimately, our goal is to enable a detailed, post-mortem reconstruction of ephemeral JS-based web attacks experienced by real network users.

### Chromium Instrumentation

Instrument rendering (Blink) and JavaScript (V8) engines Build cross-layer context as a graph HTML modifications, Network requests, JS attributions

### Chromium Instrumentation

Instrument rendering (Blink) and JavaScript (V8) engines

- Build cross-layer context as a graph
- HTML modifications, Network requests, JS attributions





Extract two types of features Structural & Content

Extract two types of features Structural & Content

Structural features capture graph properties



Extract two types of features Structural & Content

Structural features capture graph properties Average degree connectivity



Extract two types of features Structural & Content

Structural features capture graph properties Average degree connectivity

Content features capture node properties

https://events.bouncex.net/track.gif/bid\_selected?partner=i
ndex&deployment=masthead&deal\_id=106202001&price=3.50000&au
ction\_number=1&ad\_unit\_id=26&source=ads&campaignid=917423&a
gent=user&mode=0&websiteid=340&visitid=1588398576368654&dev
iceid=2799665660403664656&pageviewid=1&sequenceid=17&client
timestamp=1588398589360&clientapiversion=tag3&device=d

https://maxcdn.bootstrapcdn.com/bootstrap/3.3.7/css/bootstr ap.min.css

Extract two types of features Structural & Content

Structural features capture graph properties Average degree connectivity

**Content features** capture node properties length of URL

https://events.bouncex.net/track.gif/bid\_selected?partner=i
ndex&deployment=masthead&deal\_id=106202001&price=3.50000&au
ction\_number=1&ad\_unit\_id=26&source=ads&campaignid=917423&a
gent=user&mode=0&websiteid=340&visitid=1588398576368654&dev
iceid=2799665660403664656&pageviewid=1&sequenceid=17&client
timestamp=1588398589360&clientapiversion=tag3&device=d

https://maxcdn.bootstrapcdn.com/bootstrap/3.3.7/css/bootstr ap.min.css



# Machine Learning

Ground truth

Filter lists – despite shortcomings [Iqbal et al. '17, Snyder et al. '20] Manual evaluation of disagreements with classifier



# Machine Learning

Ground truth

Filter lists – despite shortcomings [Iqbal et al. '17, Snyder et al. '20] Manual evaluation of disagreements with classifier



Random forest classifier 10-fold cross validation

Accuracy is more than 95.33%

Recall 86.6% – Precision 89.1%

**Evaluation:** Accuracy

Accuracy is more than 95.33%

Recall 86.6% – Precision 89.1%





**Evaluation:** Accuracy

Accuracy is more than 95.33% Recall 86.6% – Precision 89.1%

Disagreement analysis with filter lists

Accuracy is more than 95.33% Recall 86.6% – Precision 89.1%

Disagreement analysis with filter lists

Filter lists under block due to unknown Ad/Trackers AdGraph detects 43.1% new ad/tackers

		For the latent updates on COVID-19, sic	in up for the Reuters Now newsletter Here	×	_
	EDITION. UNITED STATES 🐱	🖉 RE	UTERS		
_	<u>↑</u>	Business Markets World Politics Tech Bi	vaikinguiews Wealth Life 😰 Pictures I	∎n Q	_
	If your payched is O Brake these for More Brake Theory Interest	with the second secon	Kyur Accurt, A tyu Salit         Image: Constraint of the second sec	dianomi stock could be the tervinere of the market and and Commerce	
	THE WIRE	Cuomo warns against 'blindy' reopening	Exclusive: U.S. con parties stimuliar were to 9 me healthcare provide 5 facing criminal inquiries ACPM EDT	Arcor Payshape to Consume to Consumeto to Consume to Consume to Consume to Consume to Consume	
	c 3m ago Americans begin to surface from isolation as states case	New York's governor pushed back against what he called premature demands that he open the state, even as Georgia, Texas, and other states partially	Buffett's Berkshire posts nearly \$50 billion luss as coronavirus causes pain	REJ	
	clamp-downs	Work during the coronavirus pandemic.     Mospitals premise new safety measures     Back to work advocates lack testion many	12.17PM EDT		
	million deal for Telefonica's Costa Rica unit	Graphic: The COVID-19 testing challenge	mocks U.S. coronavirus response in Lego-Uke animation		
	@ 35m ago				
		Filter	LISTS		
		Filter	LISTS		
Pinancia ×		ГШЕГ	LISTS		
Financia × tant	ESTINGAL UNITES STATES V	Earthe bilder (gelden or CO201 1), eq	LISTS region for the Renders New consendence or Maximum TERS Mathematical Section 10 (1997) (1	× ≯fin ∎rv q.	
Pleance × Secure   https://www.reuters.com		For the based casister or COVER 19, or	LISTS projective Reconvertience age UTERS undergrieven With Life @ Peteren	× ≯fin ∎rv Q	
Planci ×	Elanou, curres sonts + 10003 Elanos Comerco 10003 Elanos Comerco 10003 Elanos Comerco 10004 Elanos Comerco	Construction of Construction o	LISTS are the the bases become the the the the bases become the the the the the the the the the th	× ≯fin 17 Q. Nabal225 19,019.5 -2,846	
Plancis X	EXPROR LATES SATE ↓ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■	Contraction of the second seco	LISTS (11)	ж эfin 17 Q Нам 225 19,0935 - 2,846	
risanci x 🛌	Extranse Garding Lateral V Topology Carbon Construction Safe Station Construction Safe Station Safe Station Construction Saf	CONTRACTOR AND	LISSS Production Non-Analysis Production Non	ж 9°f fa 10° Q Natei25 17,03 % 2,846	
Franco: X Secure   Itzga)/www.ruders.com	Concerc Carrier District V     Concerc Carrier District Distrindistrict District District District District District District Di	CONTRACTOR CONTRA	LISSUS are porter headen between ease UTERS Water on the fully of a porter TSE too 5,26,26 - 2,346 TSE too 5,26,26 -	ж Э́fin 17 Q NBM225 10,6035 -2846	
Pando x	CONTINUE LANGED SUSCEL V	Control and a control and control and a control and a control and a c	LISSUS (1) (1) (1) (1) (1) (1) (1) (1) (1) (1)	ж 9 fin 17 Q Нам 225 19,093.5 - 2,846	
Franco x	Image: control contro control control contro control control control control control co	CURRENT OF CONTRACT ON CONTRACT         CURRENT ON CONTRACT         CURRE	LISSUS Production Networkshow Production Networkshow	ж 9°f (h 10°7 Q Natei 225 17,00 ж 2,846	
Thato: X	COLOR COLOR OF COLOR  COLOR COLOR OF COLOR  COLOR OF COLOR  COLOR OF COLOR  COLOR OF COLOR  COLOR OF COLOR  COLOR OF COL	<text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text>		× ⊮ Г іл № 17 Q. Nativi 225 10,010.5 - 2,86%	

AdGraph

Accuracy is more than 95.33% Recall 86.6% – Precision 89.1%

Disagreement analysis with filter lists

Filter lists under block due to unknown Ad/Trackers AdGraph detects 43.1% new ad/tackers

Filter lists over block due to generic rules AdGraph identifies 28.7% over blocked functional content





Accuracy is more than 95.33%

Recall 86.6% – Precision 89.1%



#### D

#### AdGraph outperforms the current state-of-the-art

Filter lists under block due to unknown Ad/Trackers AdGraph detects 43.1% new ad/tackers

Filter lists over block due to generic rules AdGraph identifies 28.7% over blocked functional content



Real time ad and tracker blocking with ML Instrumentation overhead Classification overhead

Page load time comparison (Stock Chromium and AdBlock Plus)

Makes up by request blocking & less rendering

Real time ad and tracker blocking with ML Instrumentation overhead Classification overhead

Page load time comparison (Stock Chromium and AdBlock Plus)

Makes up by request blocking & less rendering

Faster than Chromium on 42% websites Faster when blocks more



Real time ad and tracker blocking with ML Instrumentation overhead Classification overhead

Page load time comparison (Stock Chromium and AdBlock Plus)

Makes up by request blocking & less rendering

Faster than Chromium on 42% websites Faster when blocks more

Faster than Adblock Plus on 78% websites Avoids rendering overhead



Real time ad and tracker blocking with ML Instrumentation overhead Classification overhead

Page load time comparison (Stock Chromium and AdBlock Plus)

Makes up by request blocking & less rendering

Faster than Chromium on 42% websites Faster when blocks more

Faster than Adblock Plus on 78% websites Avoids rendering overhead

Minor overhead on most websites


# **Evaluation: Performance**

Real time ad and tracker blocking with ML Instrumentation overhead Classification overhead

## AdGraph improves page load time

0.8

Makes up by request blocking & less rendering

Faster than Chromium on 42% websites Faster when blocks more

P

Faster than Adblock Plus on 78% websites Avoids rendering overhead

Minor overhead on most websites



Use cross-layer context to address **accuracy** issues

Use machine learning address **scalability** issues

Use cross-layer context to address **accuracy** issues

Use machine learning address scalability issues

Open source implementation

#### AdGraph

Code release for our IEEE Symposium on Security and Privacy 2020 paper entitled AdGraph: A Graph-Based Approach to Ad and Tracker Blocking



Use cross-layer context to address **accuracy** issues

Use machine learning address scalability issues

Open source implementation

Maintained by Brave as PageGraph

# AGGraph Sode release for our IEEE Symposium on Security and Privacy 2020 paper www on GitHul www on GitHu

Code Issues 2,205 Pull requests 29 Projects 15 Actions Wiki

### PageGraph

Anton Lazarev edited this page on Dec 16, 2019  $\cdot$  24 revisions

Use cross-layer context to address **accuracy** issues

Use machine learning address scalability issues

Open source implementation

Maintained by Brave as PageGraph

Filter list generation

#### **Filter List Generation for Underserved Regions**

Alexander Sjösten Chalmers University of Technology

Antonio Pastor Universidad Carlos III de Madrid

Panagiotis Papadopoulos Brave Software Benjamin Livshits Brave Software Imperial College London

#### ABSTRACT

Filter lists play a large and growing role in protecting and assisting web users. The vast majority of popular filter lists are crowd-sourced, where a large number of people manually label resources related 20-24, 2020, Taipei, Taiwan. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3366423.3380239

**1** INTRODUCTION

Peter Snvder

Brave Software



Code Issues 2,205 Pull requests 29 Projects 15 Actions Wiki

#### PageGraph

Anton Lazarev edited this page on Dec 16, 2019 · 24 revisions



## References

- 1. Advertising revenue https://www.iab.com/wp-content/uploads/2019/05/Full-Year-2018-IAB-Internet-Advertising-Revenue-Report.pdf
- 2. Malvertising <u>https://www.zdnet.com/article/hackers-have-breached-60-ad-servers-to-load-their-own-malicious-ads/</u>
- 3. Malvertising https://www.theguardian.com/technology/2016/mar/16/major-sites-new-york-times-bbc-ransomware-malvertising
- 4. Slow page load <u>https://www.nytimes.com/interactive/2015/10/01/business/cost-of-mobile-ads.html</u>
- 5. OpenWPM <u>https://github.com/mozilla/OpenWPM</u>
- 6. Privacy Badger <u>https://github.com/EFForg/privacybadger</u>
- 7. Iqbal, Umar et al. "The ad wars: retrospective measurement and analysis of anti-adblock filter lists." Proceedings of the 2017 Internet Measurement Conference. 2017.
- 8. Snyder, Peter et al. "Who filters the filters: Understanding the growth, usefulness and efficiency of crowdsourced ad blocking", SIGMETRICS. 2020.
- 9. Bhagavatula, Sruti, et al. "Leveraging machine learning to improve unwanted resource filtering." Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop. 2014.
- 10. Gugelmann, David, et al. "An automated approach for complementing ad blockers' blacklists." Proceedings on Privacy Enhancing Technologies 2015.
- 11. Ikram, Muhammad, et al. "Towards seamless tracking-free web: Improved detection of trackers via one-class learning." Proceedings on Privacy Enhancing Technologies 2017.
- 12. Wu, Qianru, et al. "A machine learning approach for detecting third-party trackers on the web." European Symposium on Research in Computer Security. Springer, Cham, 2016.
- 13. Li, Bo, et al. "JSgraph: Enabling Reconstruction of Web Attacks via Efficient Tracking of Live In-Browser JavaScript Executions." NDSS. 2018.
- 14. Icon made by Pixel perfect from www.flaticon.com