

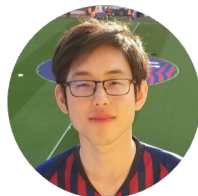
# Addressing the Adverse Impacts of JavaScript on Web Archives

---



**Ayush Goel**

*University of Michigan*



**Jingyuan Zhu**

*University of Michigan*



**Ravi Netravali**

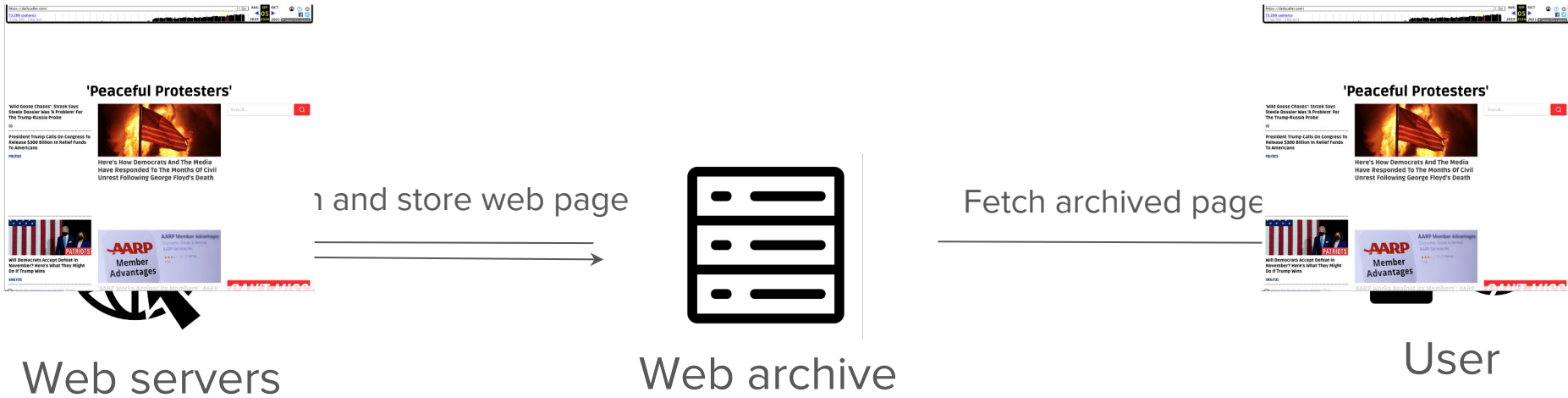
*Princeton University*



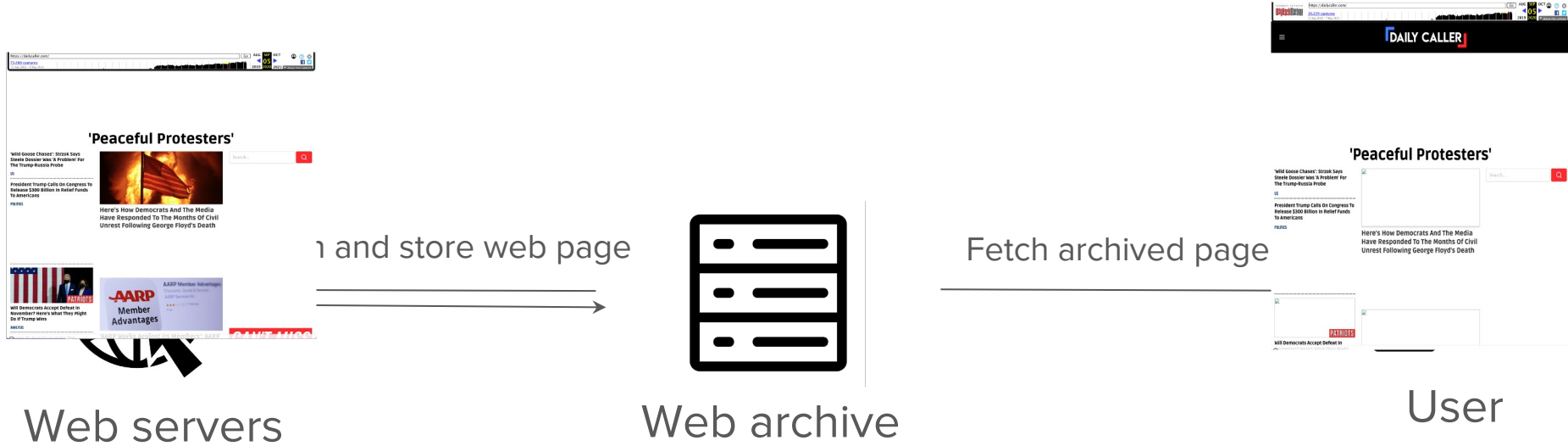
**Harsha V. Madhyastha**

*University of Michigan*

# How Modern Web Archives Operate?



## Problems With Web Archives: **Poor Page Fidelity**





# Problems With Web Archives: **Poor Page Fidelity**



Fetch and store web page



Fetch archived page

Web servers

Web archive



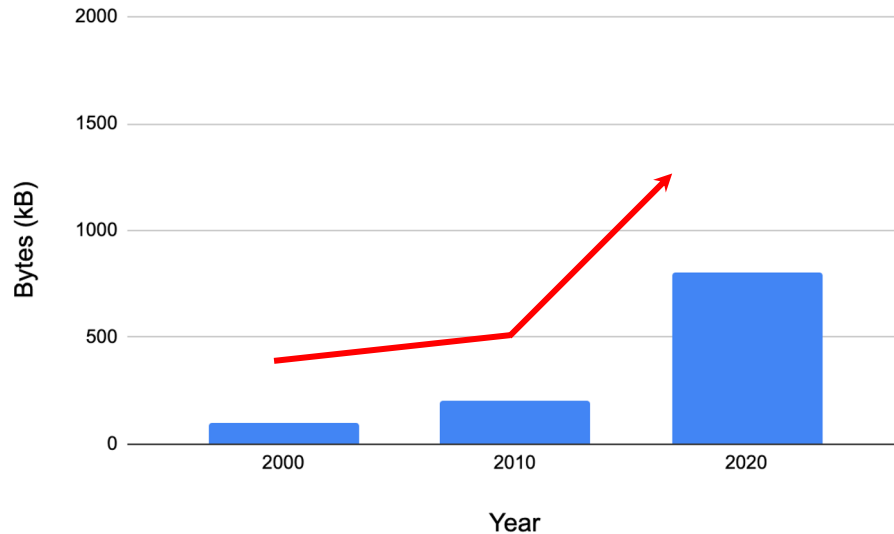
User



# Browser Errors: Missing Resources (404) and Runtime Errors

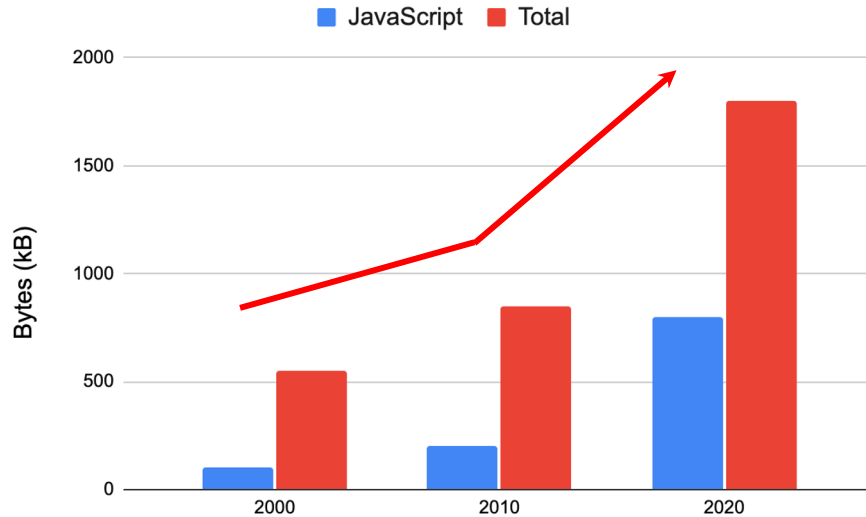
✖	▶ GET https://web.archive.org/web/20200905133311/https://htlb.casalemedia.com/cyg...com%2F%22%7D%2C%22ext%22%3A%7B%22source%22%3A%22prebid%22%7D%7D&ac=j&sd=1 404	(index):117	🔗
✖	▶ POST https://web.archive.org/web/20200905133311/https://hb.undertone.com/hb?pid=2252&domain=dailycaller.com 404	(index):117	🔗
✖	▶ POST https://web.archive.org/web/20200905133311/https://prebid.a-mo.net/a/c net::ERR_BLOCKED_BY_CLIENT	(index):117	🔗
✖	GET https://web.archive.org/web/20200905133311js_/https://www.googletagservices.com/tag/js/gpt.js net::ERR_BLOCKED_BY_CLIENT	web.archive.org/:95	🔗
✖	▶ GET https://web.archive.org/web/20200905143156/https://basketballbelieve.com/v2/0/dqjeWxSDEoEN7Rcv_wombat.js?v=txqj7nKC:21Hq1V3IXf9ltPp42IefE2tP04CRTbHa1odGic45rvuiJnc0HpWUmAYLr net::ERR_HTTP2_SERVER_REFUSED_STREAM		🔗
✖	▶ GET https://web.archive.org/web/20200905133311im_/https://images.dailycaller.com/p_/web/20200905133311i...1599262824744.jpg:1 _content/uploads/2020/09/GettyImages-1268412082-scaled-e1599262824744.jpg net::ERR_HTTP2_SERVER_REFUSED_STREAM		🔗
✖	▶ GET https://web.archive.org/web/20200905133311/https://c.amazon-adsystem.com/e/_9-1274-40b0-8c1c-1f4c8a36cb_14&gdprl=%7B%22status%22%3A%22cmp-timeout%22%7D 404	apstag.ts:4455	🔗
✖	▶ SyntaxError: Unexpected token '<' at t (apstag.ts:4455:18)	apstag.ts:4455	
✖	▶ GET https://web.archive.org/web/20200908065246/https://sb.scorecardresearch.com/ns_c=UTF-8&c8=The%20Daily%20Caller&c7=https%3A%2F%2Fdailycaller.com%2F&c9= 404	b:1	🔗
✖	▶ POST https://web.archive.org/web/20200905133308/https://www.google-analytics.com...97&tid=UA-12159302-1&gid=1562094415.1599312797&r=1&cd6=Rambo&z=190629678 404	analytics.js:48	🔗
✖	▶ GET https://web.archive.org/web/20200905133308/https://www.google-analytics.com...9312797&tid=UA-12159302-1&gid=1562094415.1599312797&cd6=Rambo&z=132763242 404	collect:1	🔗
✖	▶ GET https://web.archive.org/web/20200905133308/https://www.google-analytics.com...9312797&tid=UA-12159302-1&gid=1562094415.1599312797&cd6=Rambo&z=935436807 404	collect:1	🔗
✖	▶ GET https://web.archive.org/web/20200905133308/https://www.google-analytics.com...312797&tid=UA-12159302-1&gid=1562094415.1599312797&cd6=Rambo&z=1949102849 404	collect:1	🔗
✖	▶ GET https://web.archive.org/web/20200905133308/https://www.google-analytics.com...9312797&tid=UA-12159302-1&gid=1562094415.1599312797&cd6=Rambo&z=156266630 404	collect:1	🔗

# Root Cause: **Increasing JavaScript** on Web Pages



*Corpus:* Landing pages of 300 Alexa sites

# Root Cause: **Increasing JavaScript** on Web Pages

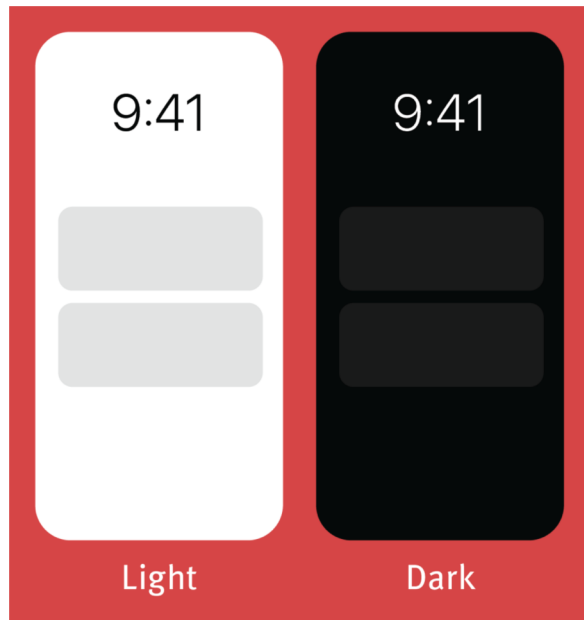
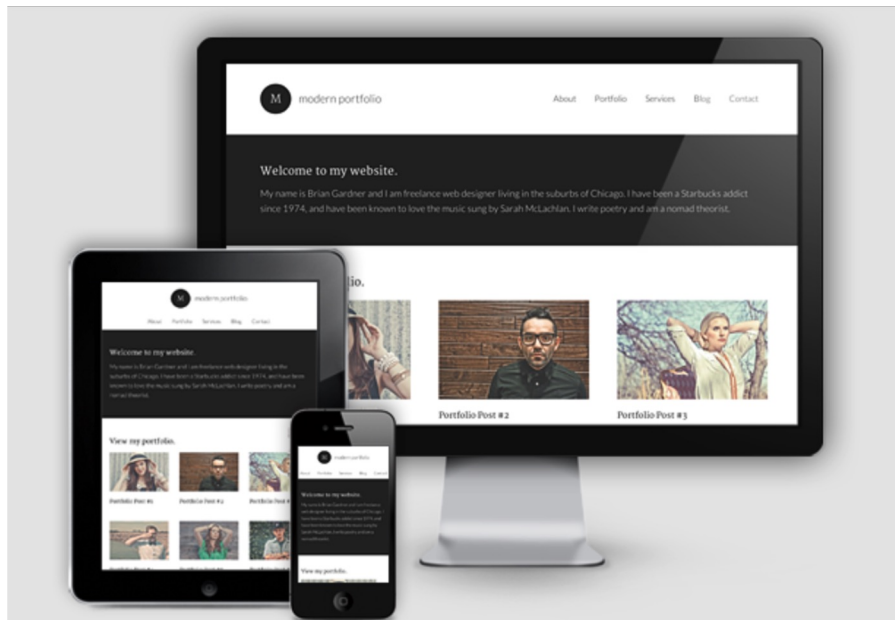


*Corpus:* Landing pages of 300 Alexa sites

Root Cause: **JS Execution Varies Across Loads of Same Page**

**Non-determinism**

**Resources fetched** different from **crawled** → **Poor page fidelity**







# Sources of JavaScript Non-Determinism

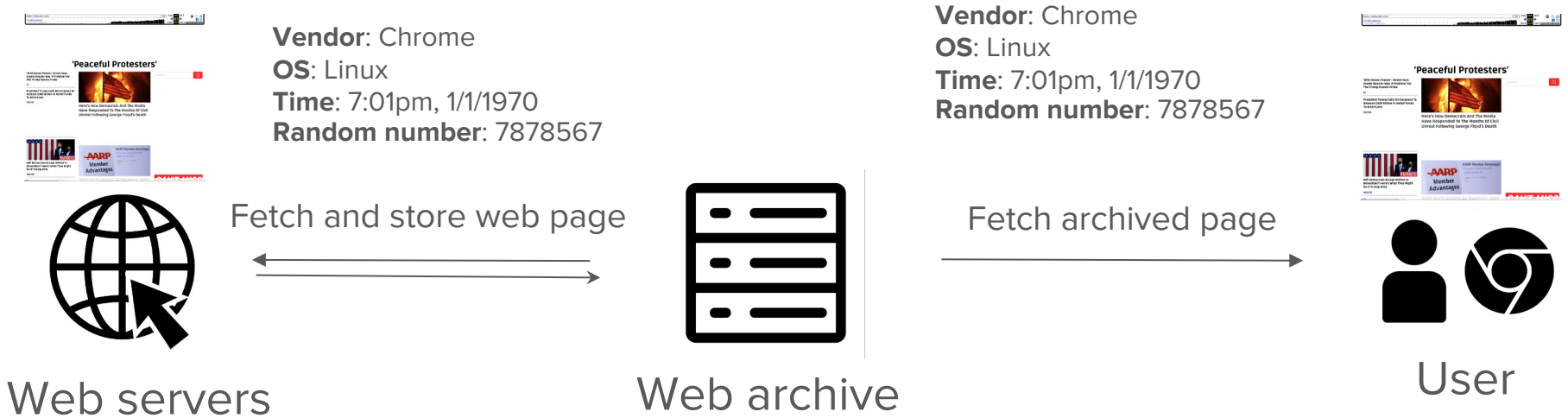
- Browser vendor (Chrome, Firefox, Edge)
- Operating system (Linux, Windows, Mac)
- Screen dimensions (height, width)
- Type of internet connection (wifi, cellular)
- Geolocation
- Hardware specs: # CPUs, memory

Client characteristics

- Current time
- Random number
- Performance

DRP APIs

# Strawman Fix: Remove All Non-determinism



# Non-determinism Critical to Page Functionality

Original page

Deterministic page (broken)

## JavaScript Tetris

Goal: fill as many rows as possible!

### Control Keys

- To move left press 4 or ←
- To move right press 6 or →
- To rotate press 5 or 8 or ↑
- To drop faster press space or ↓

### Mouse Control

- To move left click to the left of the piece
- To move right click to the right of the piece
- To rotate click on/above the piece
- To drop faster click below the piece

Featured at [JavaScripter.net](http://JavaScripter.net)  
Copyright © 1999 Alexei Kourbatov

Level:  Lines:



## JavaScript Tetris

Goal: fill as many rows as possible!

### Control Keys

- To move left press 4 or ←
- To move right press 6 or →
- To rotate press 5 or 8 or ↑
- To drop faster press space or ↓

### Mouse Control

- To move left click to the left of the piece
- To move right click to the right of the piece
- To rotate click on/above the piece
- To drop faster click below the piece

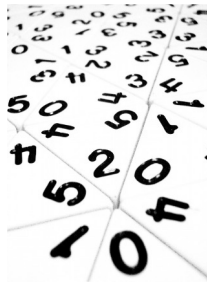
Featured at [JavaScripter.net](http://JavaScripter.net)  
Copyright © 1999 Alexei Kourbatov

Level:  Lines:

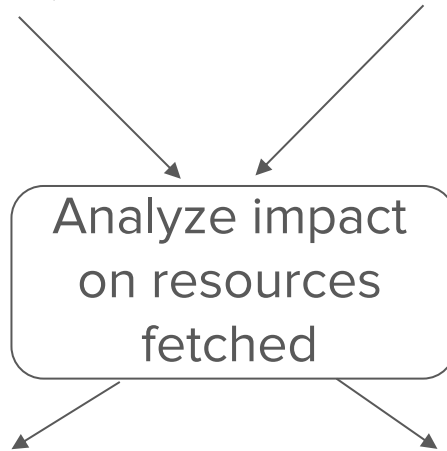


# Analyze How Non-determinism Impacts Resources Fetched

Date, Math.random,  
Performance (DRP)



Client characteristics



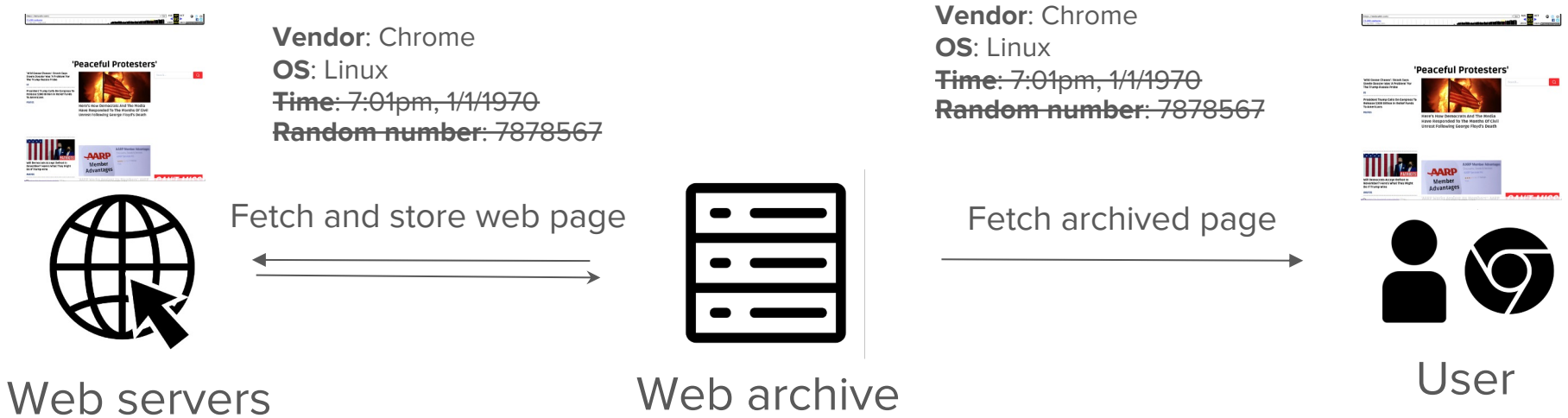
**No impact on resource fetches**

→ retain the non-determinism

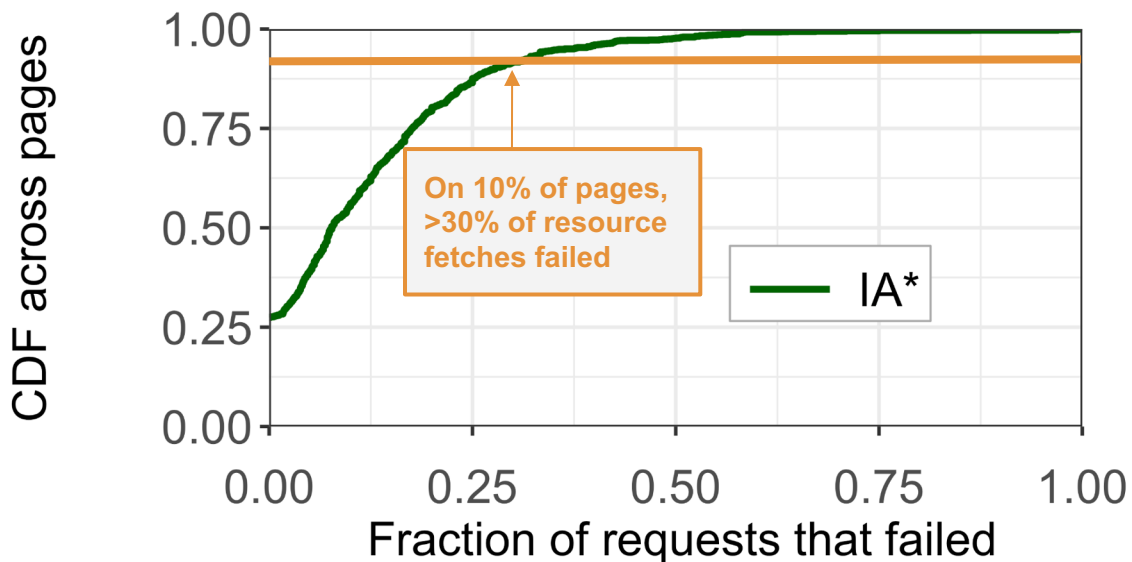
**Influences resources fetched**

→ Eliminate non-determinism

# Strawman Fix: Remove All Non-determinism

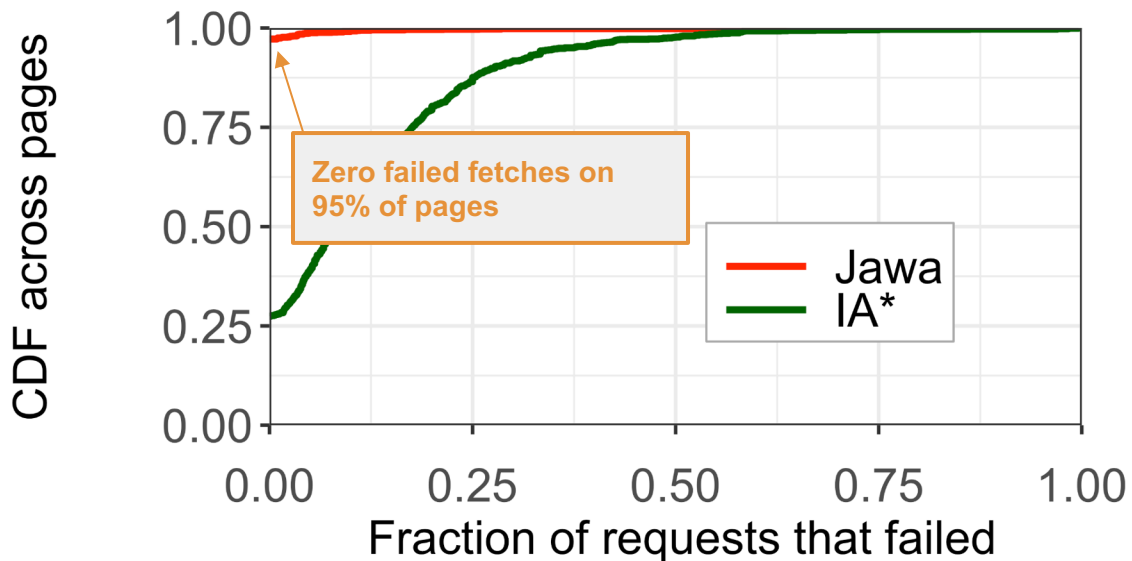


# Eliminated Almost All Failed Resource Fetches



*Corpus*: 3000 web pages from 300 sites

# Eliminated Almost All Failed Resource Fetches



*Corpus*: 3000 web pages from 300 sites

# Summary

- 1) JavaScript's non-determinism negatively impacts web archival
  - a) Missing resources, runtime errors
- 2) Key to fixing this is to selectively remove non-determinism
- 3) Store and reuse values of all client-characteristic APIs



<https://gist.github.com/goelayu/bbc1348f4b5913ce9a304f3cfd708bb1>



[goelayu@umich.edu](mailto:goelayu@umich.edu)





Backup slides

## Client-characteristics: **Remove Non-Determinism**

- While archiving page, store values of client characteristic APIs
  - ◆ Vendor: Chrome
  - ◆ OS: Linux
  - ◆ Chrome version: 110
  - ◆ Screen dimensions: 1200x800
  
- While loading archived page, reuse the stored client characteristic values

# Questions

# DRP APIs: Use

# Handle

- Browser vendor (Chrome, Firefox, Edge)
- Operating system (Linux, Windows, Mac)
- Screen dimensions (height, width)
- Current time
- Random number
- Type of internet connection (wifi, cellular)
- Geolocation
- Hardware specs: # CPUs, memory





# Impact of Browser Vendor On Archived Page

## Archived using Chrome

Microsoft Azure Continuously learn new skills and experiment with Azure

US World Politics Business Opinion Health Entertainment Style Travel Sports Videos LIVE TV

LIVE UPDATES: Ukraine | TRENDING: Brittany Griner arrest | NYC, DC manhunt | Tesla cruise control | Patrick Mahomes married | PODCAST: Diversifying

## Blasts heard around Kyiv as Russian forces inch closer

**Donetsk, Ukraine**

**Russia could default on its debt within days**

**Ex-US ambassador to Ukraine: Trump 'emboldened' Putin**

**UN agency: More than 2.8 million people have fled Ukraine**

## Loaded using IE

```
SCRIPT5007: Invalid descriptor for property 'responseURL'  
File: bundle-playback.js, Line: 2, Column: 3284  
SCRIPT1010: Expected identifier  
File: wombat.js, Line: 21, Column: 81692  
SCRIPT5009: '__wm' is undefined  
File: www.cnn.com, Line: 6, Column: 3  
SCRIPT87: Invalid argument.  
File: www.cnn.com, Line: 177, Column: 421  
SCRIPT5007: Unable to get property 'desktopSSID' of undefined or null reference  
File: header.e5ea7a852706fffd34dc.bundle.js, Line: 265, Column: 107679  
SCRIPT5007: Unable to get property 'analytics' of undefined or null reference  
File: www.cnn.com, Line: 177, Column: 170250  
SCRIPT5009: '__wm' is undefined  
File: www.cnn.com, Line: 186, Column: 9  
SCRIPT5009: '__wm' is undefined  
File: www.cnn.com, Line: 300, Column: 1
```

# Problems with Web Archives: **Poor Page Fidelity**



**'Peaceful Protesters'**

**'Wild Goose Chases': Sirzok Says Steele Dossier Was 'A Problem' For The Trump-Russia Probe**

US

**President Trump Calls On Congress To Release \$300 Billion In Relief Funds To Americans**

POLITICS

Here's How Democrats And The Media Have Responded To The Months Of Civil Unrest Following George Floyd's Death

Will Democrats Accept Defeat In

PATRIOTS

Search...

**The Boston Globe**

TRENDING: BOSTON MARATHON BOMBER | DR. FAUCI | 'BLACK IS KING' | JOE BIDEN | BOSTON SCHOOLS

ON STATNEWS.COM

How the world made so much progress on a Covid-19 vaccine so fast

The first U.S. Covid-19 case was found in her district. A congressman reflects on the last six months

CORONAVIRUS

The latest COVID-19 numbers from Massachusetts

Town-by-town COVID-19 data in Massachusetts

11 coronavirus vaccines to keep an eye on

Graphic: US shows little signs of being able to control coronavirus spread

EDITORIAL

**End the federal death penalty now**

From the Archives: To end the anguish, drop the death penalty

OPINION

MOST READ ON BOSTONGLOBE.COM

- Melwan's police chief is one of the highest paid in the country — and he says he deserves more
- It was all a lie
- As airlines lose billions, they're cleaning and sanitizing as if their lives depend upon it. But is it?

CORONAVIRUS

Fauci Coronavirus Fleeces



# Sources of JavaScript non-determinism

- **Browser vendor (Chrome, Firefox, Edge)**
- **Operating system (Linux, Windows, Mac)**
- **Screen dimensions (height, width)**
- **Current time**
- **Random number**
- Type of internet connection (wifi, cellular)
- Geolocation
- Hardware specs: # CPUs, memory

# Outline

- Web page trend (increasing JS)
- What is the utility of increasing JS
- Web archives suffer detrimentally
- variation can break the page – runtime errors; missing resources
- simple solution: remove all non-determinism
- our analysis