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?

Web servers

ı and store web page

Web archive

Fetch archived page

User

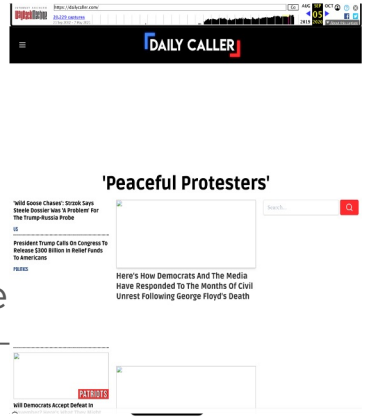
Problems With Web Archives: **Poor Page Fidelity**

Web servers

ı and store web page

Web archive

Fetch archived page



User

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

Web servers

ch and store web page

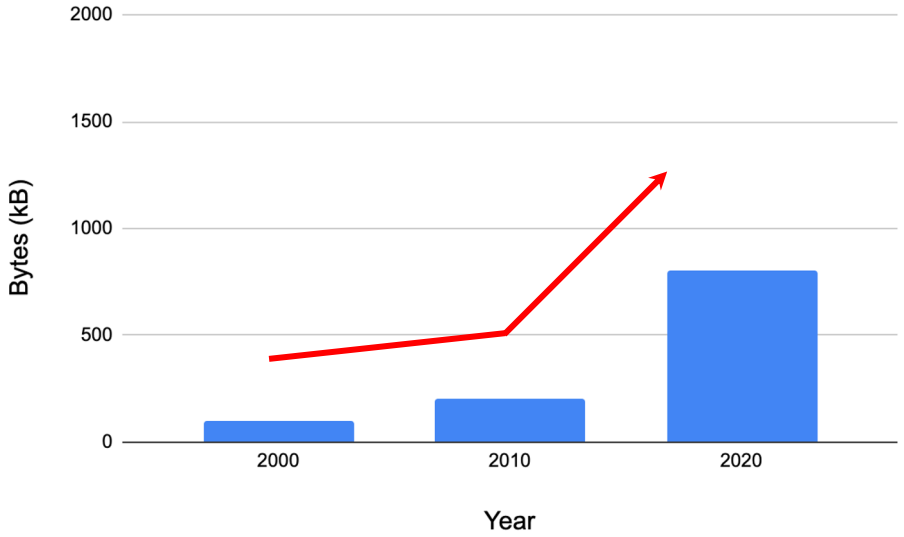
Web archive

Fetch archived page

User

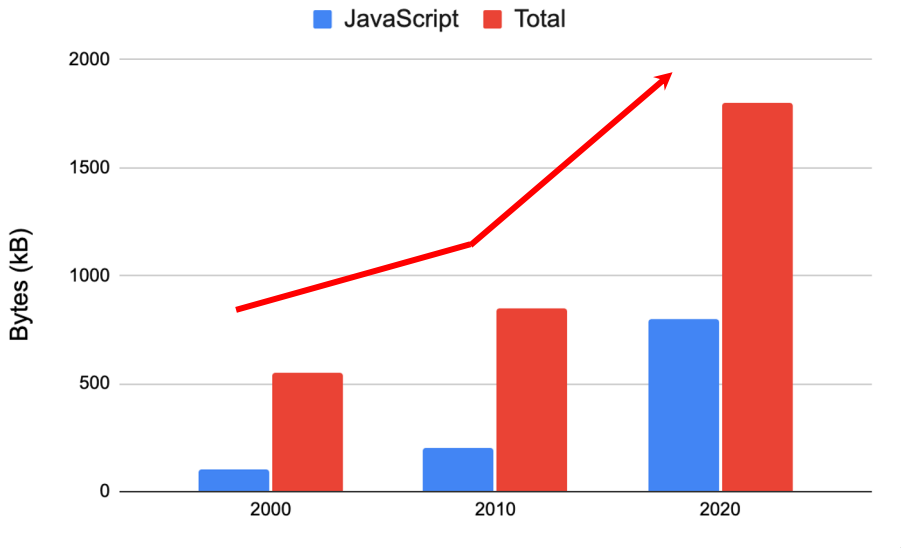
Browser Errors: Missing Resources (404) and Runtime Errors

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



Corpus: Landing pages of 300 Alexa sites

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

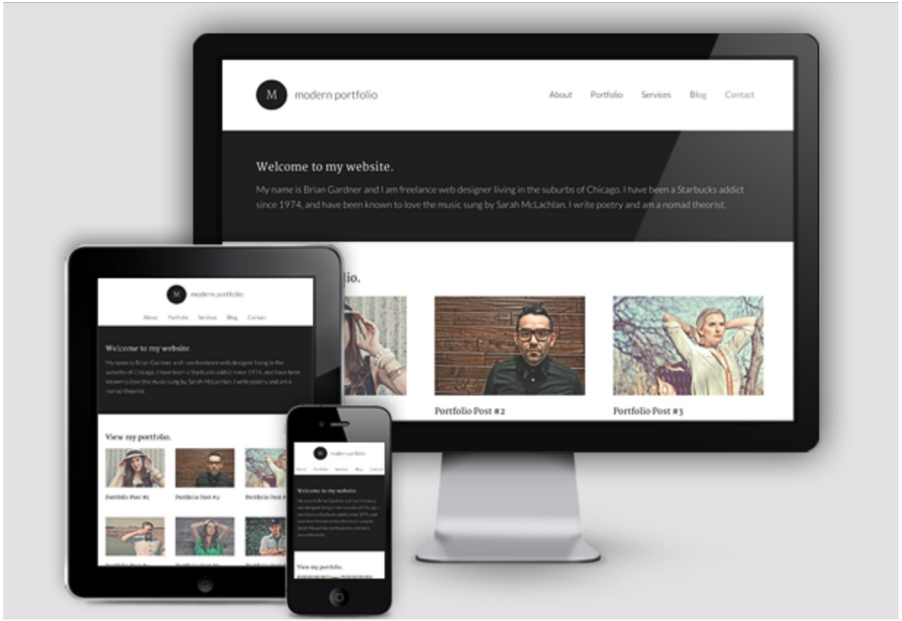


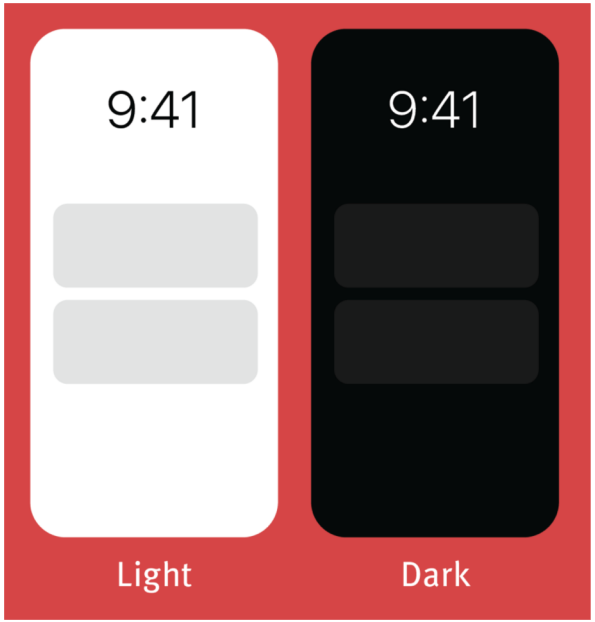
Corpus: Landing pages of 300 Alexa sites

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

Vendor: Chrome
OS: Linux
Time: 7:01pm, 1/1/1970
Random number: 7878567

Fetch and store web page

Web servers

Web archive

Vendor: Chrome
OS: Linux
Time: 7:01pm, 1/1/1970
Random number: 7878567

Fetch archived page

User

Non-determinism Critical to Page Functionality

Original page

JavaScript Tetris

Level: 1 Lines: 0 Start Pause

Goal: fill as many rows as possible!

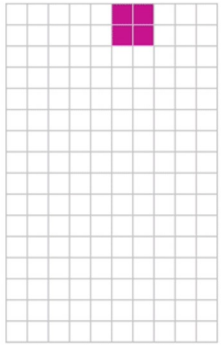
Control Keys

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
Copyright © 1999 Alexei Kourbatov



Deterministic page (broken)

JavaScript Tetris

Level: 1 Lines: 0 Start Pause

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 ↓

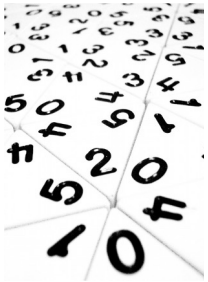
Mouse Control

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
Copyright © 1999 Alexei Kourbatov

Analyze How Non-determinism Impacts Resources Fetched

Date, Math.random, Performance (DRP)



Client characteristics

Analyze impact on resources fetched

No impact on resource fetches

➔ retain the non-determinism

Influences resources fetched

➔ Eliminate non-determinism

Strawman Fix: Remove All Non-determinism

Vendor: Chrome
OS: Linux
~~**Time**: 7:01pm, 1/1/1970~~
~~**Random number**: 7878567~~

Web servers

Fetch and store web page

Web archive

Fetch archived page

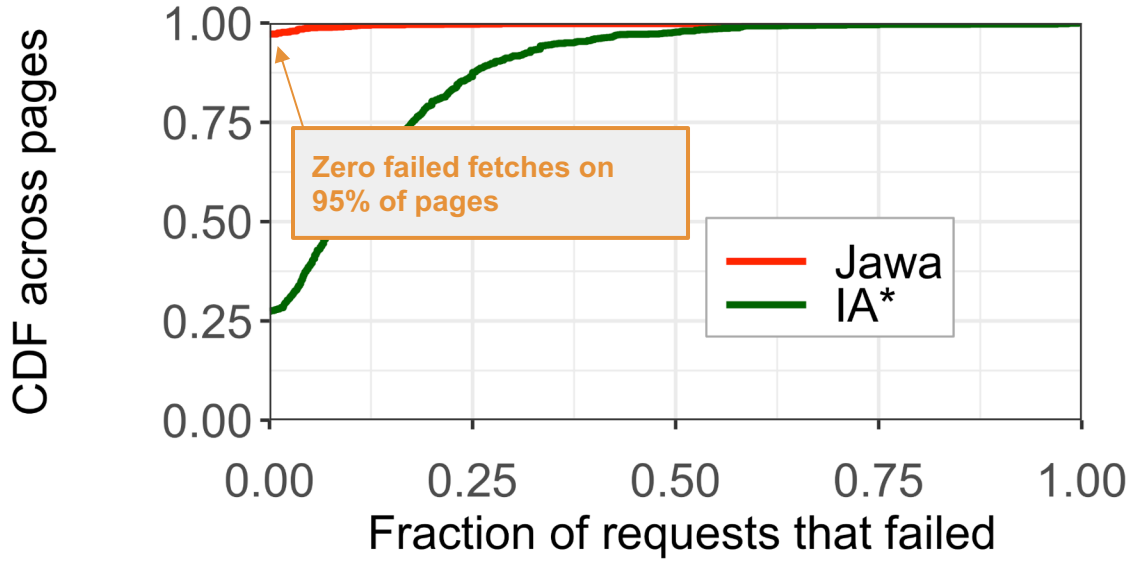
Vendor: Chrome
OS: Linux
~~**Time**: 7:01pm, 1/1/1970~~
~~**Random number**: 7878567~~

User

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

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

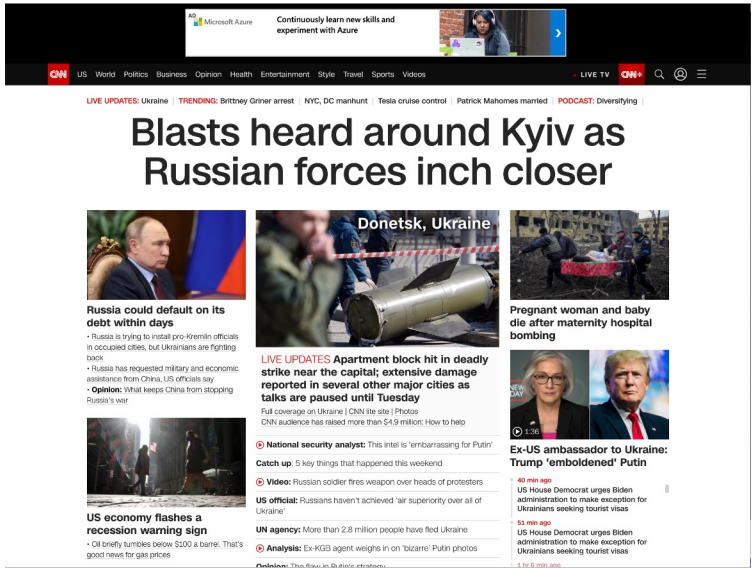
DRP APIs: Use

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



Loaded using IE

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

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