Duumviri: Detecting Trackers and Mixed Trackers with a Breakage Detector

He (Shawn) Shuang, University of Toronto

Lianying Zhao, University of Toronto/Carleton University

David Lie, University of Toronto

Tracker Blockers

Adblock Plus

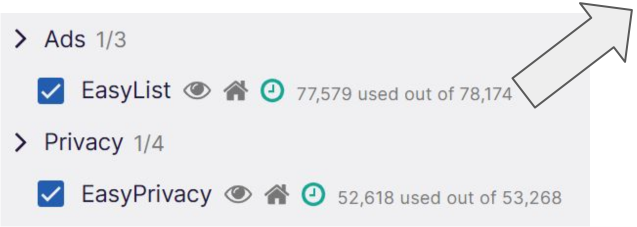
uBlock Origin

AdGuard

Privacy Badger

Filter Lists

Adblock Plus



> Ads 1/3

☑ EasyList 77,579 used out of 78,174

> Privacy 1/4

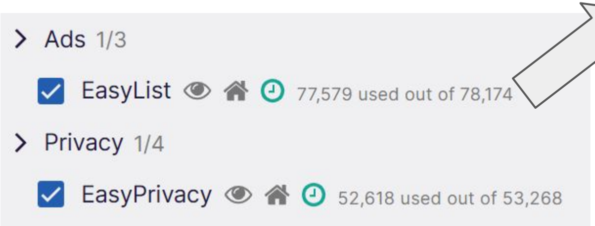
☑ EasyPrivacy 52,618 used out of 53,268

Human-written rules

```
_adobe_analytics.js
_track_pixel.gif?
://analytics.*/collect
/clicklog.js
&action=js_stats&
&ev=PageView&
```

Filter Lists Problem

Adblock Plus



› Ads 1/3

☑ EasyList 77,579 used out of 78,174

› Privacy 1/4

☑ EasyPrivacy 52,618 used out of 53,268

Human-written rules

```
_adobe_analytics.js
_track_pixel.gif?
://analytics.*/collect
/clicklog.js
&action=js_stats&
&ev=PageView&
```

Problems:

1. Scalability
2. Accuracy

Previous Works

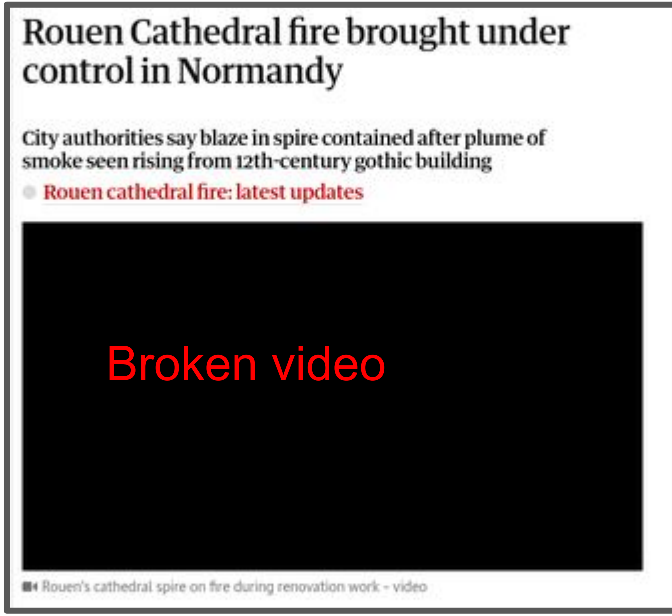
Feature sources

1. **HTML features**
    a. E.g., whether iframe element is the ancestor of a third party script element
2. **Network layer features**
    a. E.g., whether the request contains over three query string parameters
3. **JavaScript execution features**
    a. E.g., whether a JavaScript called *Canvas* API before sending a request
4. **Graph features** involves the interactions between HTML elements, network requests and JavaScript executions
    a. E.g., flow of information from cookiejar to requests
    b. [AdGraph SP 2020], [WebGraph USENIX 2022]

High Breakage Rate

Problem: current tracker detector can cause high breakage

- Previous work (e.g., [AdGraph SP 2020]) breaks 15% of pages



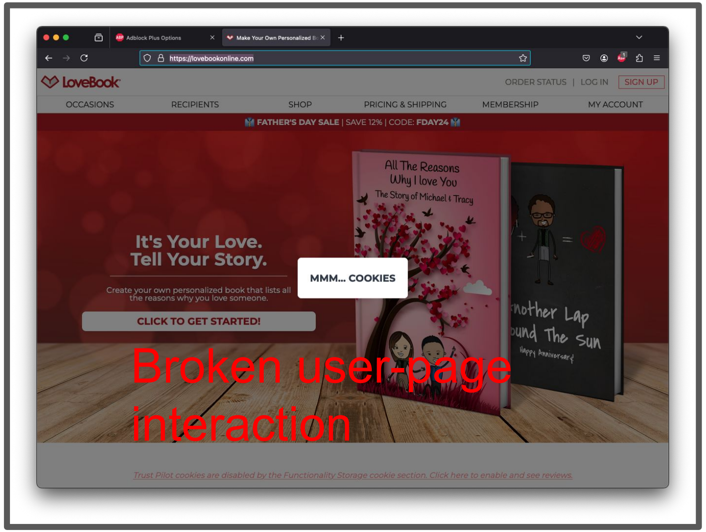
Rouen Cathedral fire brought under control in Normandy

City authorities say blaze in spire contained after plume of smoke seen rising from 12th-century gothic building

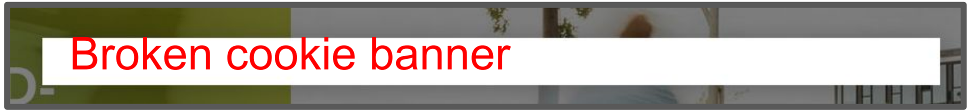
Rouen cathedral fire: latest updates

Broken video

Rouen's cathedral spire on fire during renovation work - video


Adblock Plus Options
Make Your Own Personalized B
https://lovebookonline.com
LoveBook
ORDER STATUS | LOG IN
SIGN UP
OCCASIONS
RECIPIENTS
SHOP
PRICING & SHIPPING
MEMBERSHIP
MY ACCOUNT
FATHER'S DAY SALE | SAVE 12% | CODE: FDAY24
It's Your Love. Tell Your Story.
Create your own personalized book that lists all the reasons why you love someone.
CLICK TO GET STARTED!
MMM... COOKIES
All The Reasons Why I love You
The Story of Michael & Tracy
Another Lap Around The Sun
Happy Anniversary
Trust Pilot cookies are disabled by the Functionality Storage cookie section. Click here to enable and see reviews.

Broken user-page interaction



Broken cookie banner

Reasons for Breakage

Reason #1: functional request is misidentified as tracking request

- Trained from an imperfect data source
- Hard to improve the quality as developers working on it

Reason #2: functional information cannot be separated from tracking requests

- Mixed tracker

For functionality

For Tracking

Mixed Trackers Example

Example mixed tracker

URL: https://www.EXAMPLE.com/landingpage-a-psurl.html?
goods_id=601099526089385&
sku_id=17592258865022&
_x_ns_msclkid=eeec99c83e911b00583ffc4bc3e34060

Consequences

- Blocking the request causes the redirect to fail
- Allowing the request hurts privacy

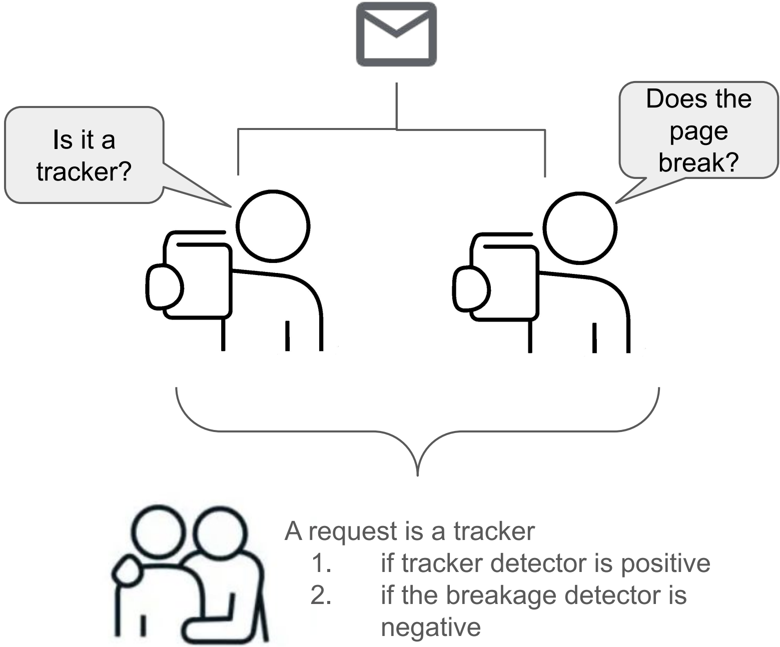
Duumviri

Duumrivi: addressing the high breakage rate of previous works

- Maximizing privacy: identifying trackers
- Minimizing breakage: not breaking web pages

Contributions

1. A two-modeled approach for tracker detection with a dedicated breakage detector
2. Detecting mixed trackers automatically



A request is a tracker

1. if tracker detector is positive
2. if the breakage detector is negative

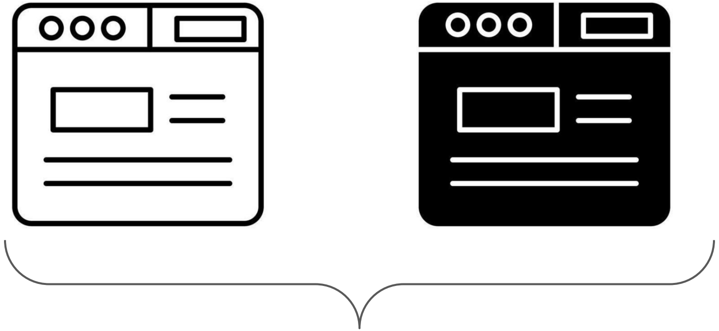
Breakage Detector

Goal: a model that can predict when a web page is broken due to tracker misidentification

1. What is a broken page in the context of tracker detection? How do we determine breakage?
   a. We define a breakage to be changes to web pages from the origin page subjectively determined by the human user
2. How do we collect samples of breakages?

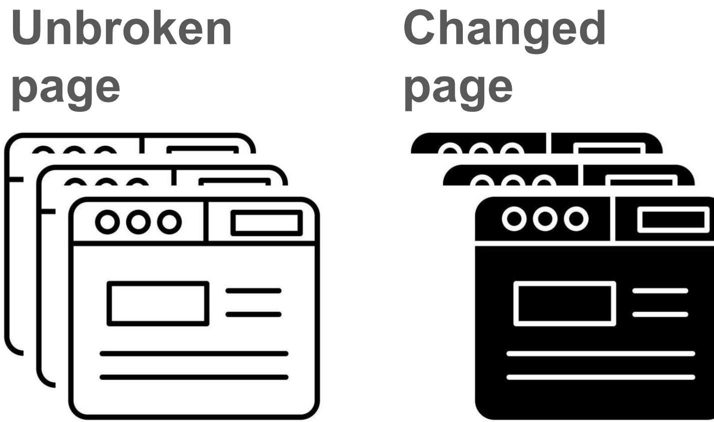
Unbroken page

Changed page



Compare and draw **differential features**

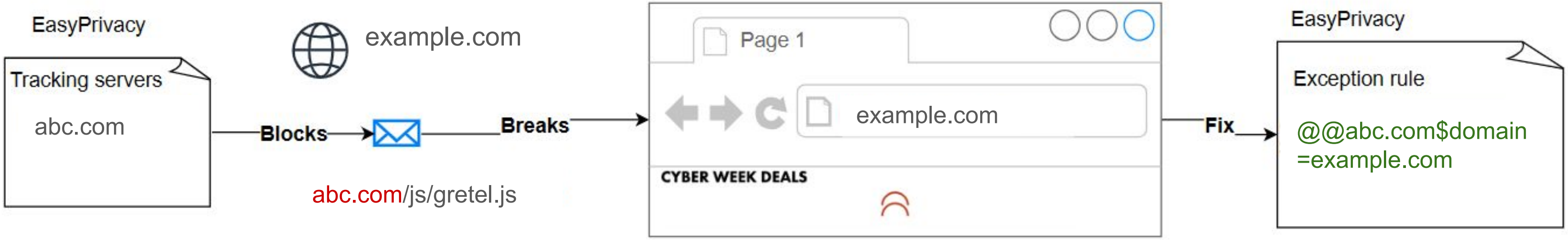
Breakage Detector - Data Samples



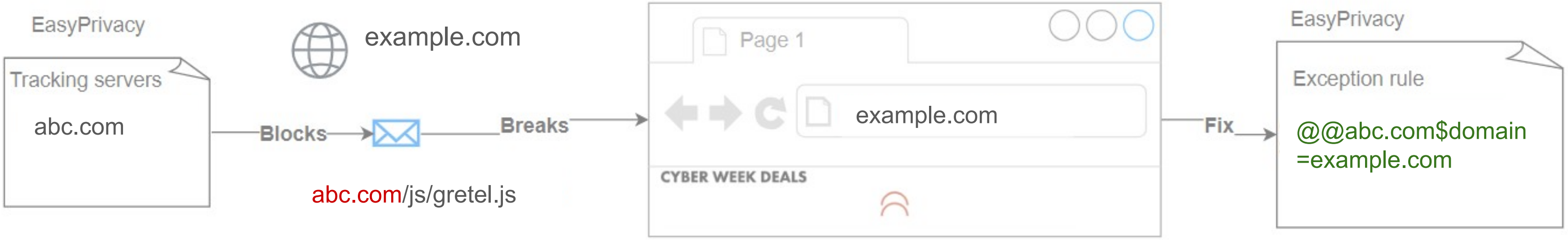
Problem: need instances of unbroken page and changed pages to draw differential features

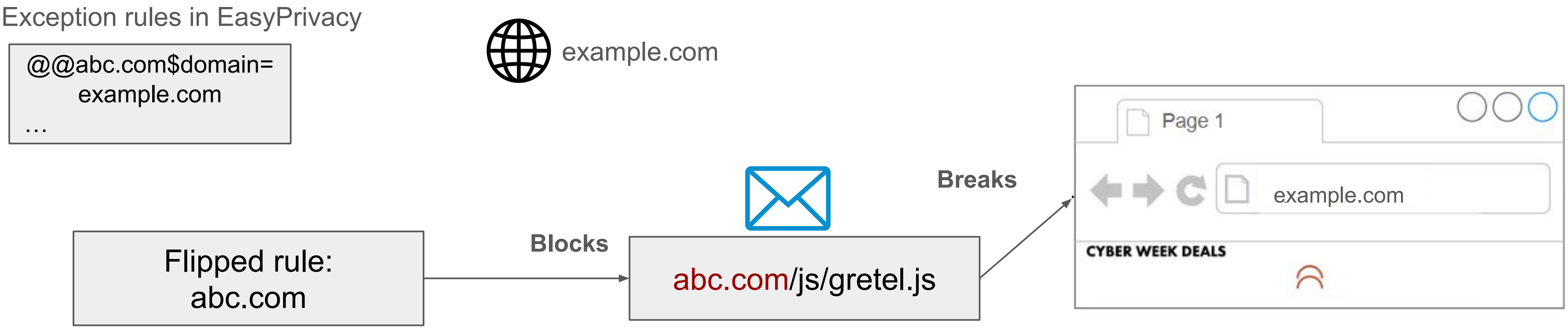
- We cannot crawl the web for breakages
  - Breakage reports are often fixed quickly, live sites with breakages are rare
  - Even if we could find live breakages, we do not have an unbroken version of the page (as the page is currently broken)
- We do not want crafted breakages

Exception Rules



Breakage Reconstruction



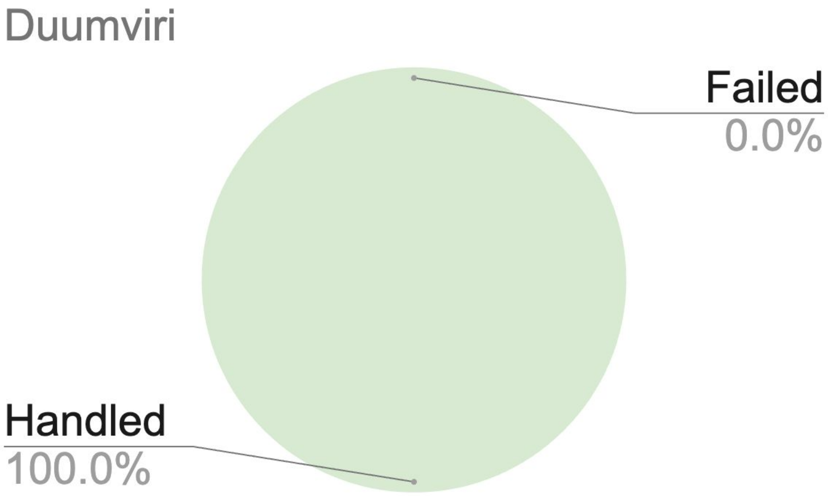


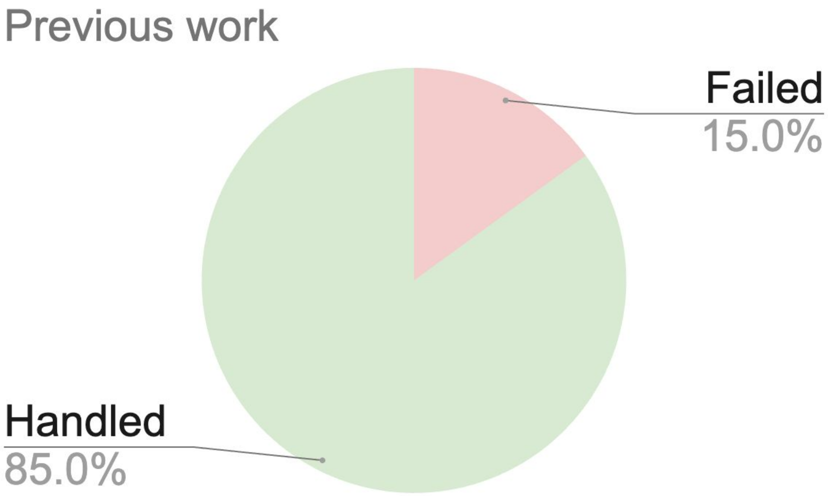
Breakage Reconstruction Evaluation

Our breakage reconstruction is accurate and reliable.

We manually look at 40 user reports of web breakages.

- Compare the reconstructed breakage to user reports

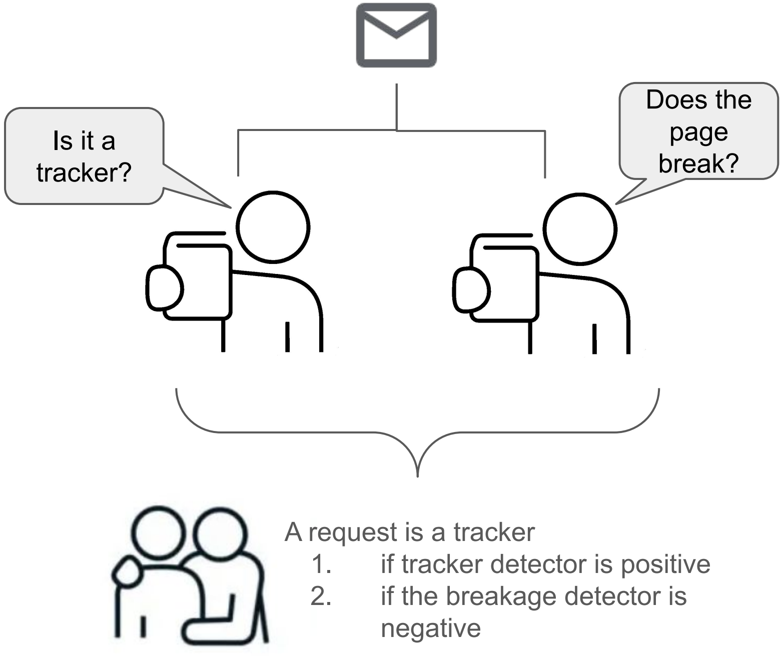

Duumviri


Previous work

Tracker Detection Evaluation

A large scale tracker evaluation on 15K pages

- Accuracy 96.53% compared to filter lists
- Disagreement analysis shows
  - Duumviri is correct 55% of cases where Duumviri is positive



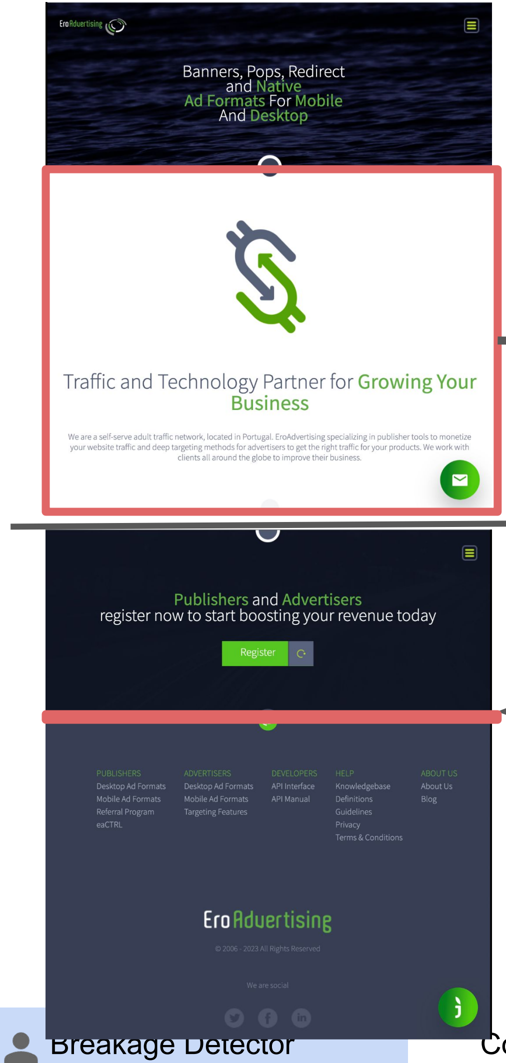
A request is a tracker

1. if tracker detector is positive
2. if the breakage detector is negative

Tracker Detection Evaluation

A large scale tracker evaluation on 15K pages

- Accuracy 96.53% compared to filter lists
- Disagreement analysis shows
  - Duumviri is correct 55% of cases where Duumviri is positive
  - EasyPrivacy-caused breakages



EroAdvertising

Banners, Pops, Redirect and Native Ad Formats For Mobile And Desktop

Traffic and Technology Partner for Growing Your Business

We are a self-serve adult traffic network, located in Portugal. EroAdvertising specializing in publisher tools to monetize your website traffic and deep targeting methods for advertisers to get the right traffic for your products. We work with clients all around the globe to improve their business.

Publishers and Advertisers register now to start boosting your revenue today

Register

PUBLISHERS
Desktop Ad Formats
Mobile Ad Formats
Referral Program
eaCTRL

ADVERTISERS
Desktop Ad Formats
Mobile Ad Formats
Targeting Features

DEVELOPERS
API Interface
API Manual

HELP
Knowledgebase
Definitions
Guidelines
Privacy
Terms & Conditions

ABOUT US
About Us
Blog

EroAdvertising

© 2006 - 2023 All Rights Reserved

We are social

Unbroken page

Web page body

Broken page

Web page body is missing

https://github.com/easylist/easylist/issues/17829

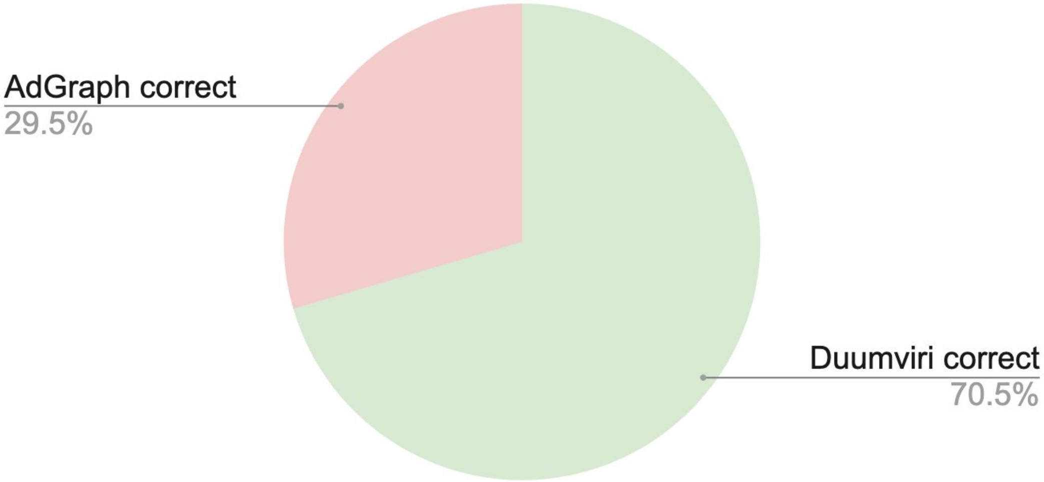
Comparison with previous work

	AdGraph	**Duumviri**
Accuracy	93.51	93.85
Precision	89.46	88.97
Recall	67.74	83.13
AuROC	0.9669	0.9682

Comparison with [AdGraph SP 2020]

- Similar accuracy using filter lists as ground truth
- Duumviri is correct in majority of times

Disagreement analysis



Duumviri

UNIVERSITY OF TORONTO

Carleton University

- Proposes a two-modelled approach for tracker detection
  - A tracker detector
  - A breakage detector
    - Trained from exception rules
- Detects mixed trackers automatically

Code, data, models:
github.com/dlgroupuoft/Duumviri-NDSS25

Artifact Evaluated
NDSS SYMPOSIUM
Available
Functional
Reproduced