A graph-based approach to tracker detection

Sandra Siby

(In collaboration with: Umar Iqbal, Zubair Shafiq, Steven Englehardt, Carmela Troncoso)

Mozilla Security Research Summit, 8 November 2019

SPRING
SECURITY AND PRIVACY ENGINEERING LABORATORY

EPFL

THE UNIVERSITY OF IOWA

moz://a

Motivation

Problem

Trackers collect information about a user’s activity on the Internet

Motivation

Problem

Trackers collect information about a user's activity on the Internet

Tracker blocking solutions mainly rely on Filter Lists

Motivation

Problem

Trackers collect information about a user's activity on the Internet

Tracker blocking solutions mainly rely on Filter Lists

What's wrong with filter lists?

- **Manual curation**
- **Keeping up with changes**

Motivation

Problem

Trackers collect information about a user's activity on the Internet

Tracker blocking solutions mainly rely on Filter Lists

What's wrong with filter lists?

- **Manual curation**
- **Keeping up with changes**

How can we solve this? *Build an automated tracker detection system!*

Motivation

Problem

Tracke ernet

T

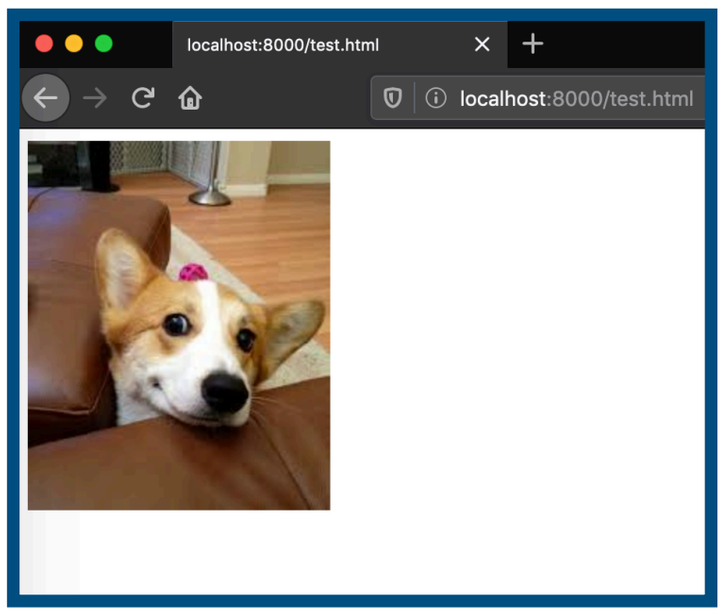
What nges

Goal: *Build a machine learning model to identify trackers for inclusion in filter lists*

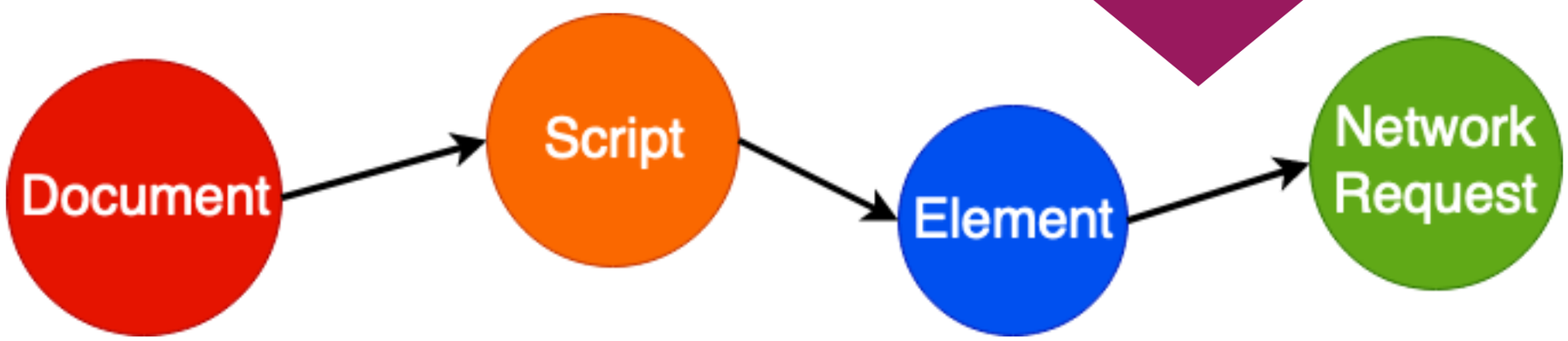
How can we solve this? *Build an automated tracker detection system!*

Build a *graph representation* of page load events

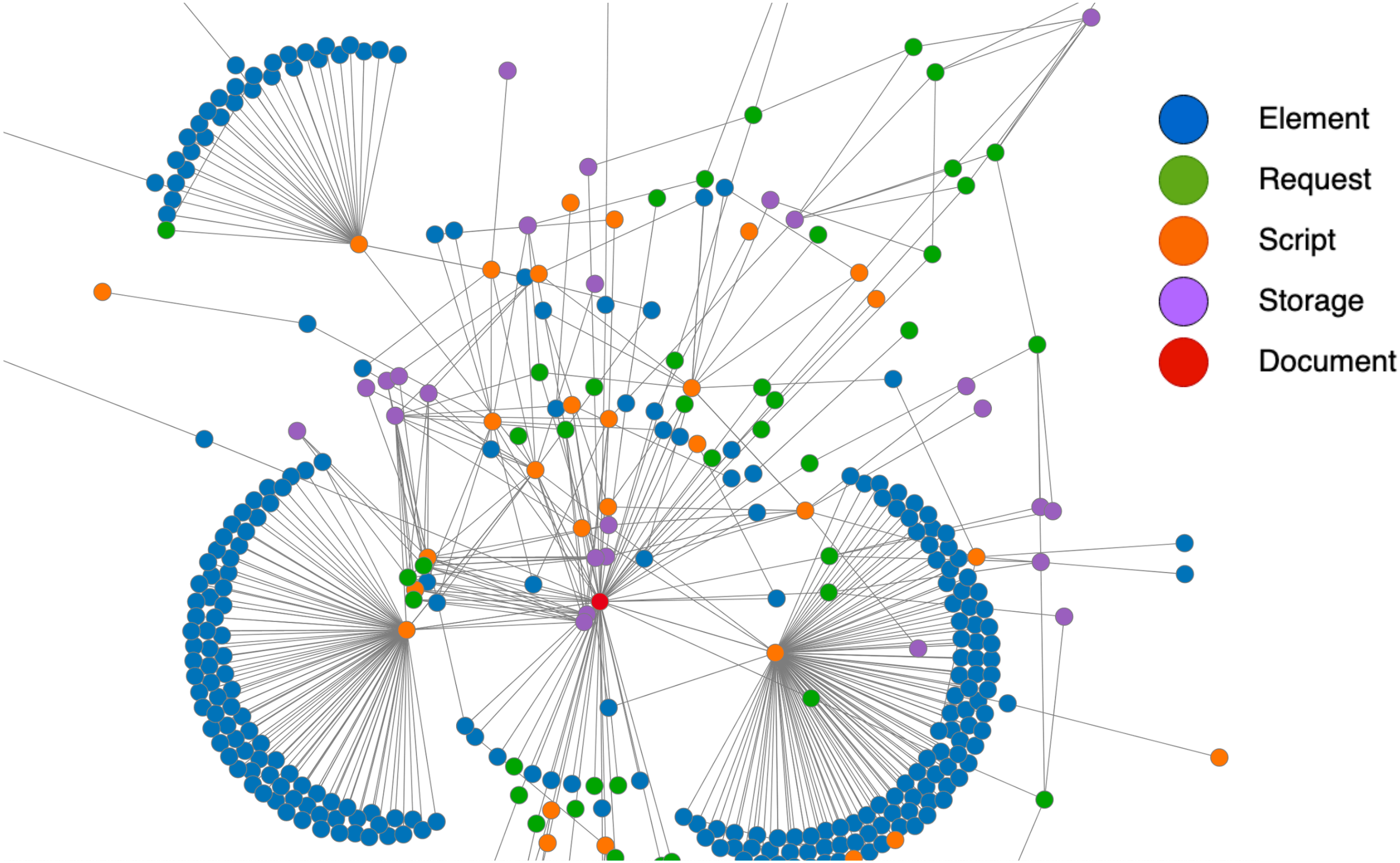
Simple Example



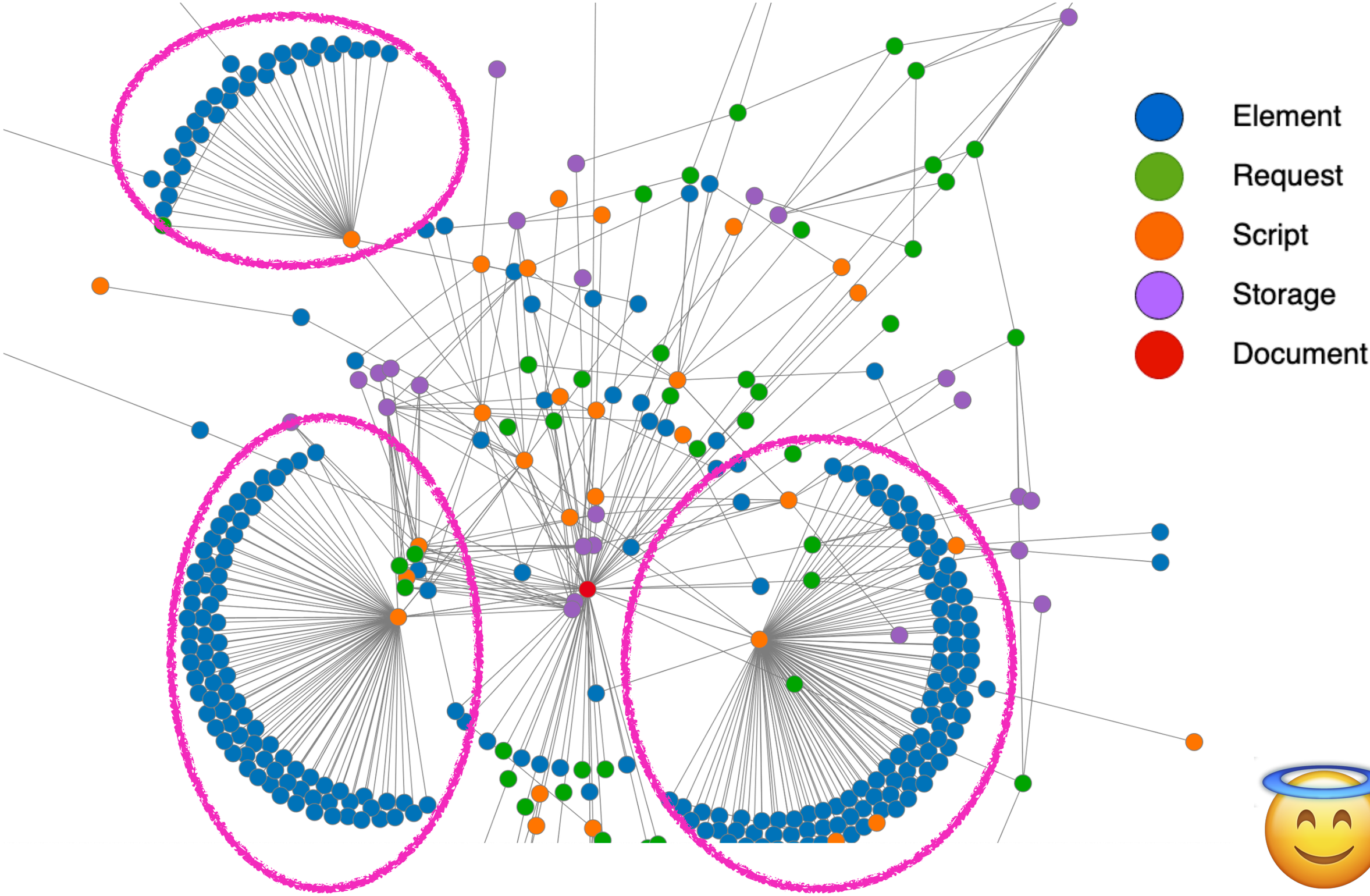

```
<html>
  <body>
     <script>
       var newimg = document.createElement("img");
       newimg.src = "https://encrypted-
tbn0.gstatic.com/images?
q=tbn:ANd9GcQgrjE6kteaw0qJ2jeeeeOUyR_-qb-
DQ9ke3j33pS_dE0vvQXDR";
       document.body.appendChild(newimg);
     </script>
  </body>
</html>
```



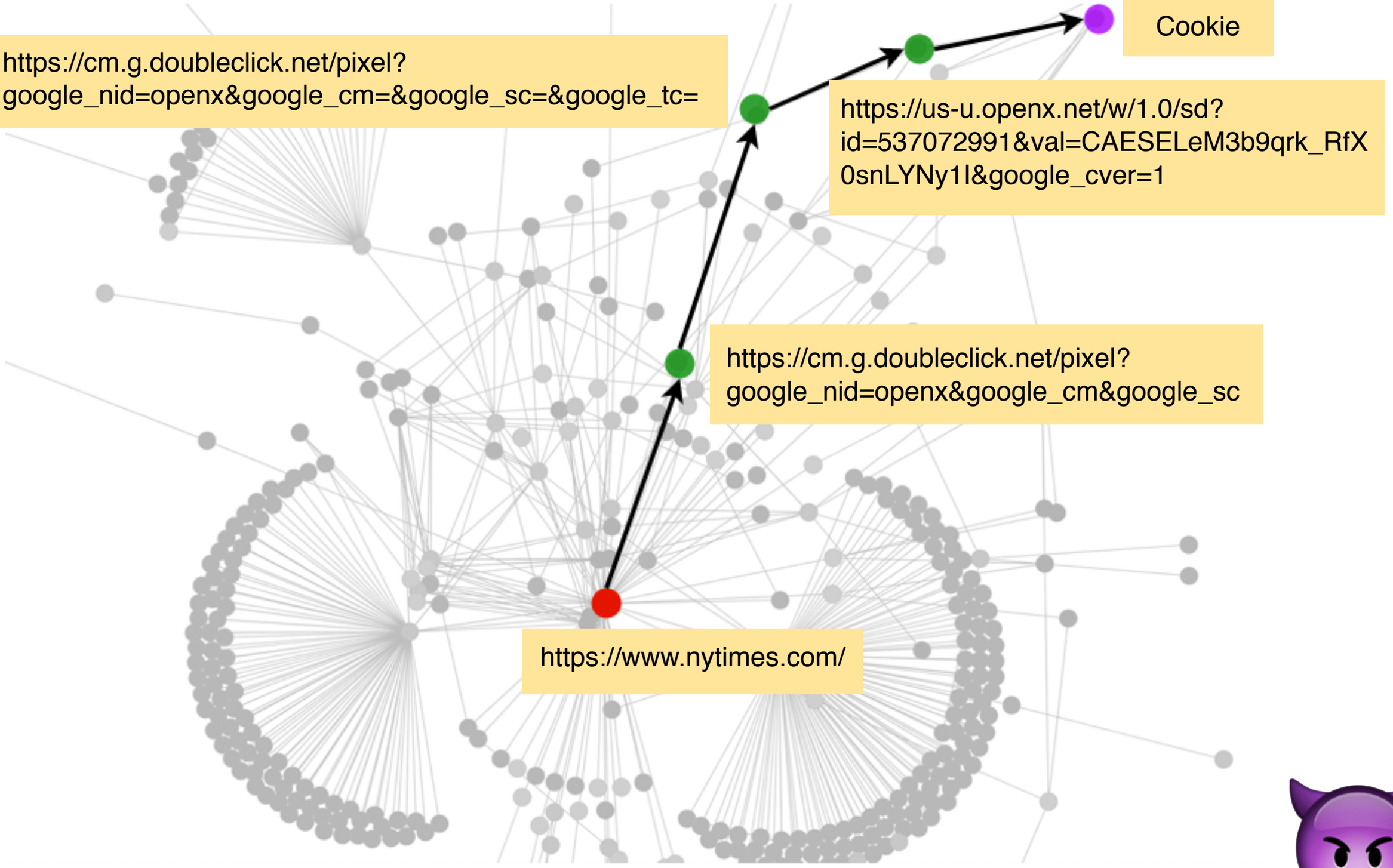
Graph Representation — Example



Graph Representation — Normal events



Graph Representation — Tracker events



Extracting features from the graph

Node Features

- Node type
- Node sub-type
- URL length
- Presence of ad keywords

.
.

Connectivity Features

- Neighbors
- Centrality measures

.
.
.

Present in existing systems

Extracting features from the graph

Node Features

- Node type
- Node sub-type
- URL length
- Presence of ad keywords

.
.

Connectivity Features

- Neighbors
- Centrality measures

.
.
.

Data Flow Features

- Content transmissions
- Cookie related actions

.
.
.

Present in existing systems

Questions we want to answer

- Can data flow features from our graph representation help us create a reliable tracker classification system?
- Do the data flow features enable us to discover fundamental tracker behavior that is hard to hide?
- Will the classifier be robust to temporal changes in websites?