Re-identification of mobile devices using real-time bidding advertising networks

ACM MobiCom '20

Keen Sung, JianYi Huang, Mark D. Corner, Brian N. Levine

University of Massachusetts Amherst

JaeHyun Lee (jhlee2021@mmlab.snu.ac.kr)

Contents

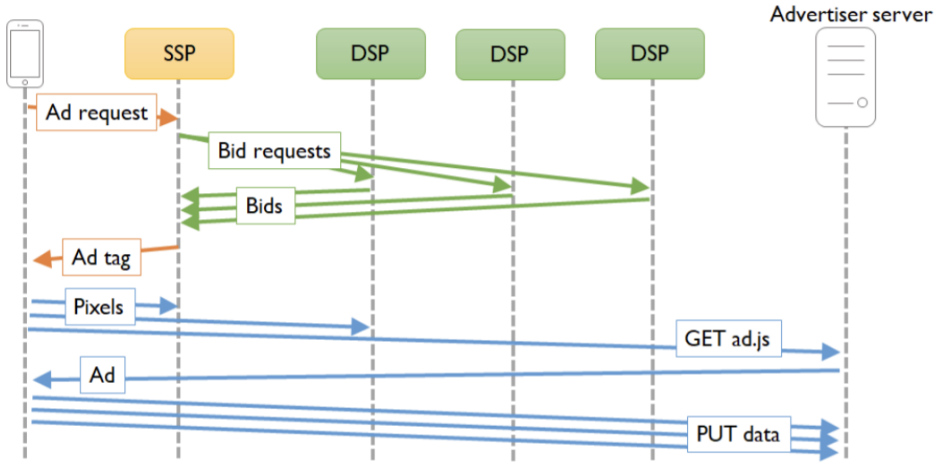
Motivation

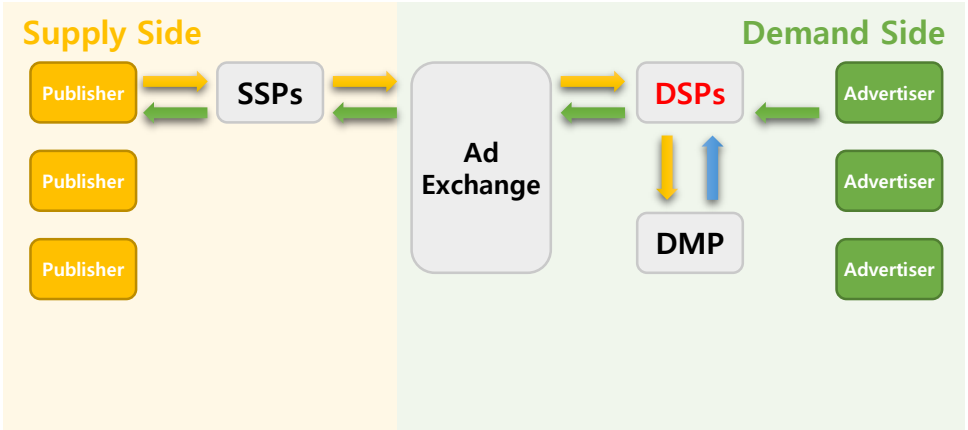
- Users allow their **information** to be shared in the online **advertising ecosystem**
  - Information: unique id (Ad ID), device info, IP, location and etc.
  - Advertising ecosystem: Real Time Bidding (RTB)

- In RTB network, ***advertisers can re-identify*** devices (users) even if users try to prevent re-identification

Real-time bidding

❖ **Real-time bidding (RTB)** is a means by which advertising inventory is bought and sold on a per-impression basis, via instantaneous programmatic auction within 100ms





- Users
  - Targets or potential consumers for the advertisers
- Supply Side Platform (SSP)
  - Gather the user profiles and forward them for the advertisers
- **Demand Side Platform (DSP)**
  - Integrate bidding information, determine bidding amount
- Ad Exchanges
  - Connect SSPs and DSPs, forward data
- Data Management Platform (DMP)
  - Possess a database comprised of multiple user profiles

*OpenRTB: https://iabtechlab.com/standards/openrtb/

Adversarial model

- Goal of adversary
  - **Maximize the recall of impressions** on a particular mobile device even when the **user tries to prevent re-identification** such as resetting or clearing the Ad ID, denying the app access to location, or using a VPN

- Vantage point
  - Self-service DSP advertising actual ads

➢ **Re-identification is quite expensive** for normal advertiser **but** it could be cheap for the motivated adversary
  - Ex. targeted fraud, espionage, cyber-stalking or Etc.

Methodology

1. **Mark unique string on device** and read it to confirm the same device in subsequent ad

2. When Ad ID is not given, advertiser can reduce cost by **filtering impressions based on retargeting features**

Methodology – 1. Marking

- Mark unique identifier via 7 features
  - Confirm which device was the original one by reading the mark

	Feature	JS	Non-JS	iOS	Android
Using browser API	1. LocalStorage	✓		✓	
	2. IndexedDB	✓		✓	✓
	3. Web SQL	✓		✓*	✓
	4. Cache API	✓		✓*	✓*
JS file including mark	5. Unique, Cached JS file	✓		✓	✓
1x1 tracking pixel	6. Cookies	✓	✓	✓*	✓
	7. HTTP ETag	✓	✓	✓	✓

✓* : *partial support*

Methodology – 2. Retargeting

- Retargeting features

Feature	Prop. matched
Device OS	0.51
Bundle	0.24
OS Version	0.18
Android Model	0.012
GPS	1.1×10^{-4}
IP	6.3×10^{-5}

Unique

When the **Ad ID is not available**, must avoid infeasible costs

Combine features to filter impressions to re-identify a single user

Evaluation

❖ Goal: ***Quantifying*** *the advertiser's* ***success rate*** *and* ***economic cost*** *of re-identification*

- Key metrics
  - Impression rate
    - The number of extra impressions of untargeted devices with retargeting filter
  - Recall
    - Ability to find a device again

Evaluation (cont.)

- Data sets
- Mark persistence
- Retargeting (recall) rate & cost
- Case studies

Data sets

- Experiment environments
  - **Self-service DSP** with JS based banner advertisements
    - Ad's HTML code includes macro to return the device IP, geolocation, UA string, OS version, available Ad ID to seven marking methods
    - Hundreds of Android and iOS apps including games weather, chat and etc.
    - Maximum bid: $10 CPM / Average winning bid: $4.02 CPM*
  - **10 US towns or cities** (8 random micropolitan** + NYC + LA)

Data set	Days	Imps	Ad IDs	Filter	Section	Purpose
PERSIST	81	52,719	1,715	Ad ID	§5.2	Persistence of marking methods
LONG	33	108,084	1,384	Ad ID	§5.2	Performance of retargeting strategy
IP	3	72,508	44,055	IP	§5.3	Cost of retargeting by IP
GPS	3	44,570	28,673	GPS	§5.3	Correlation between imps rate and population density
FIRST	4	22,319	9,986	IP/GPS	§6.1	Case Study 1: Without Ad ID
VPN	28	74,026	13,778	Ad ID, App	§6.2	Case Study 2: User under VPN

Randomly sample a set of 3,099 Ad IDs for 2days

m LONG data set

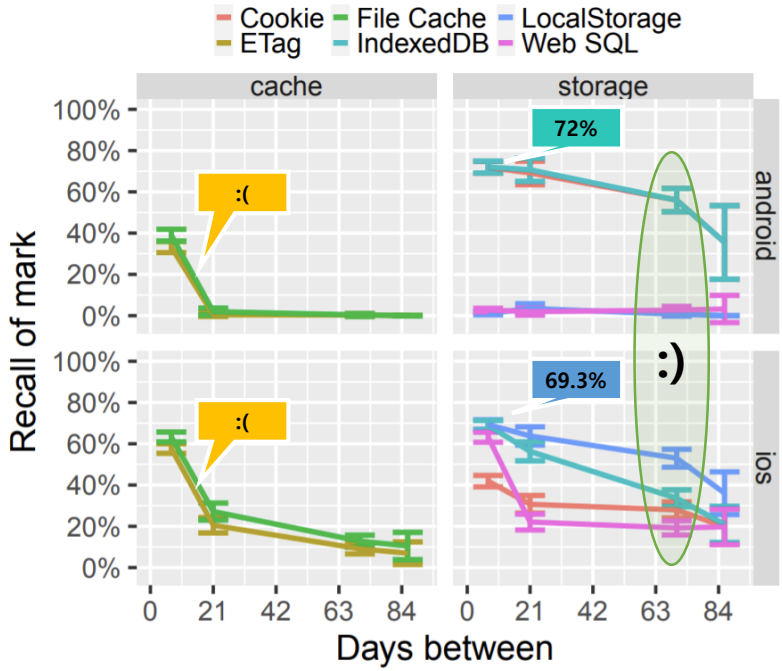
For Case Studies

*CPM (Cost per Mille): cost of 1,000 impressions, $4.02 is extremely higher than a normal price
** To avoid bias due to geography or population density

Mark persistence

- Assumption
  - Adversary's goal: re-mark a device before the original mark is cleared
  - Unless all stored marks are cleared at the same time, the missing values can be restored

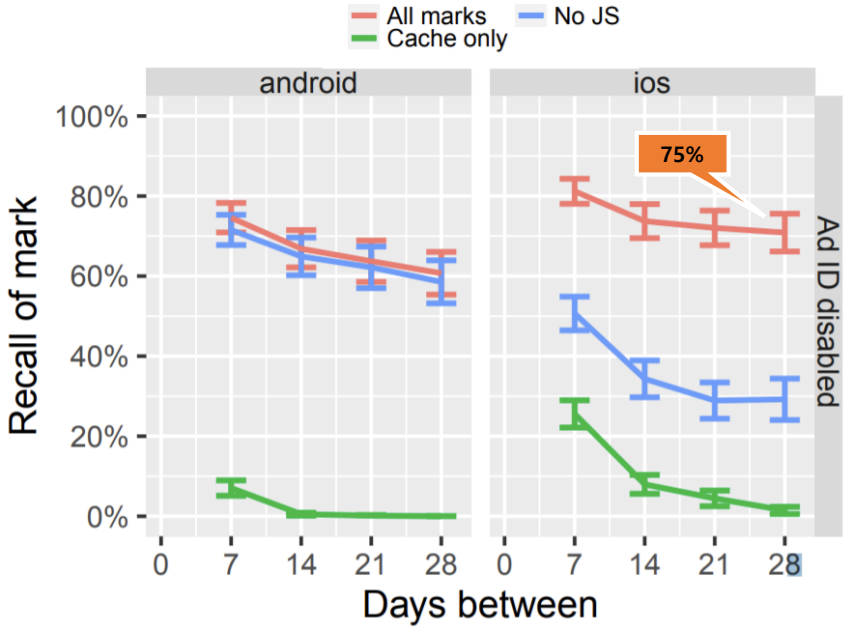
- Using **PERSIST** data set
  - Retargeting **intermittently** (7, 20, 70 and 81)
  - Cache
    - Mostly cleared after a week
  - Storage
    - Android: Indexed DB and cookies (72.0%)
    - iOS: LocalStorage (69.3%)



Mark persistence (cont.)

- Assumption
  - Adversary's goal: re-mark a device before the original mark is cleared
  - Unless all stored marks are cleared at the same time, the missing values can be restored

- Using **LONG** data set
  - Retargeting **every 2 hours** continuously



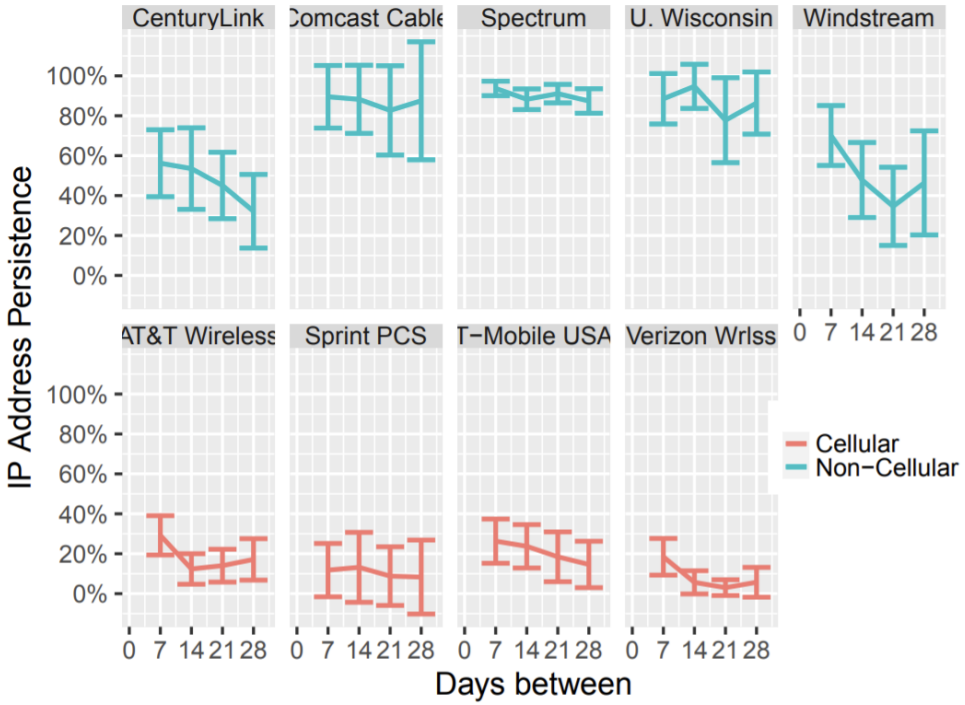
➢ **More persistent** if encountered regularly

Retargeting

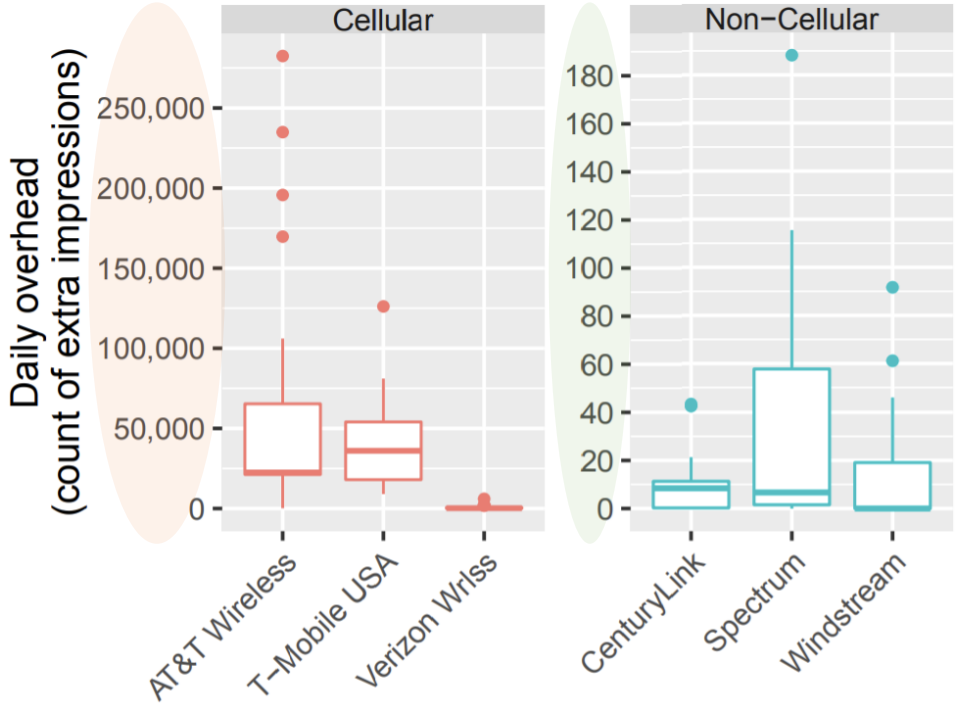
❖ Retargeting strategy

1. Targeting the **IP address** seen during training
2. Additionally targeting some radius of **GPS coordinates** seen
3. Filtering out irrelevant bids based on **static features** (model, OS, app)

Retargeting – IP Addresses



<Persistence of IP address from various ISP>

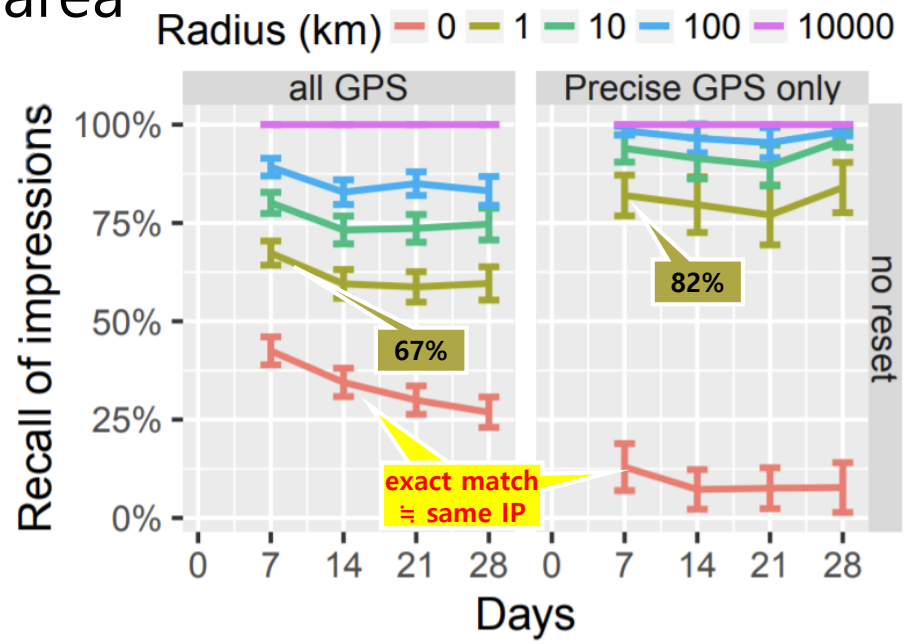


<Impression rate for an IP per day>

➢ Devices on **non-cellular** ISPs can be targeted **cheaply and effectively**

Retargeting – Geocoordinates

- Geofencing can **boost recall** if the device uses another IP but remains in the same general area



➢ **Geocoordinates** reduce impression rate so **reduce cost** of retargeting

Retargeting – Device attributes

- Further minimize the cost by filtering out devices that do not match the OS, model, and app used by the target

- Overall, the impression rates is reduced to 0.01% (android) and 0.39% (iOS) using a static filter, compared to the rate without a filter

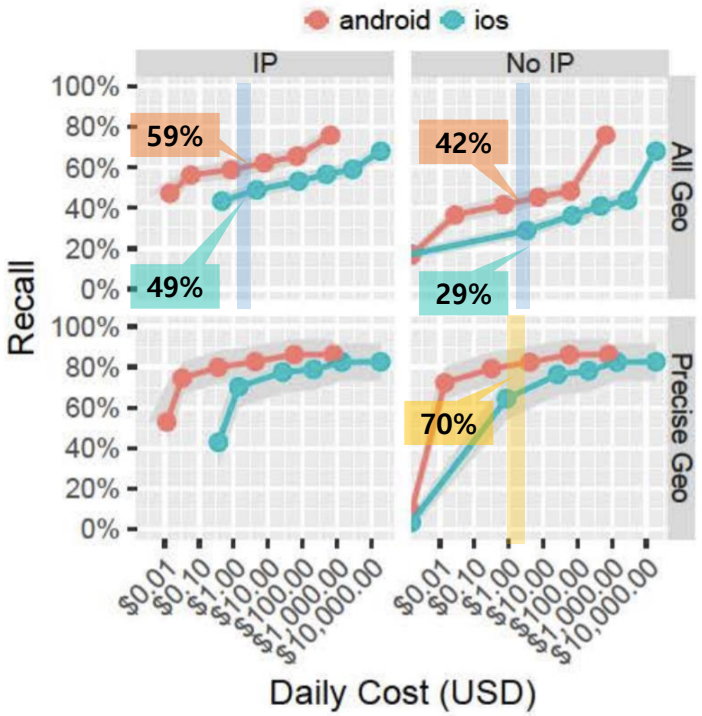
➢ **Filtering out irrelevant devices** based on these static attributes can **reduce cost** by several orders of magnitude

Cost

- Average CPM **$4.02**
  - Very high for RTB banner advertisements,
    but it was intentional to maximize recall
  - Daily cost = Impression rate x Average CPM

✓ IP address + Geo + static filter
- 49% (iOS), 59% (android) recall for less than $5/day

✓ Precise Geo alone (geo app)
- 70% recall for $2/day



➢ Shared geolocation or IP can significantly reduce the cost of tracking;
while recall is **still possible when both are hidden, more expensive**

Case Studies

❖ *Extra case studies for further insight into the (1) cost of rediscovering a user without Ad ID, and (2) the effectiveness of user's privacy decisions*

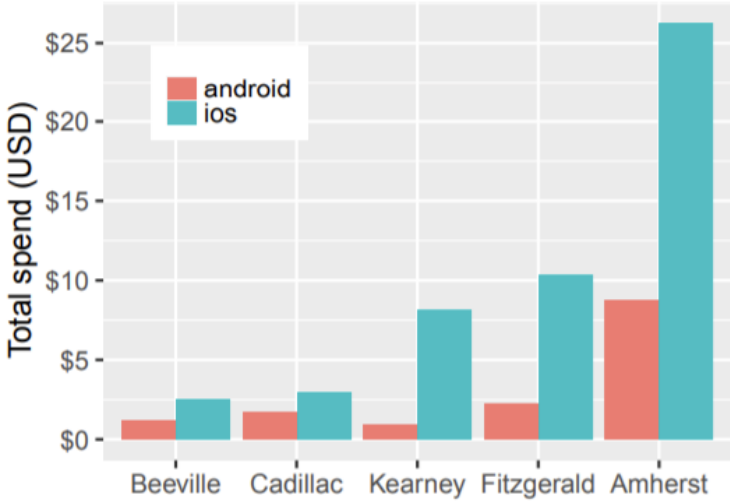
- Case Study 1: Rediscovering devices with one impression

- Case Study 2: Re-identification of VPN users

CS1: Rediscovering with one impression

❖ Goal: *Rediscover as many as possible while minimizing costs w/o Ad ID*

- Based on app, device and marking method,
  - 1,727 unique devices* were found on 5 towns
  - Within 48 hours, rediscover 660 (**38%**) devices
  - Retargeting cost: **$86.73**
    - $21.65(IP) + $14.85(Android) + $50.23(iOS)



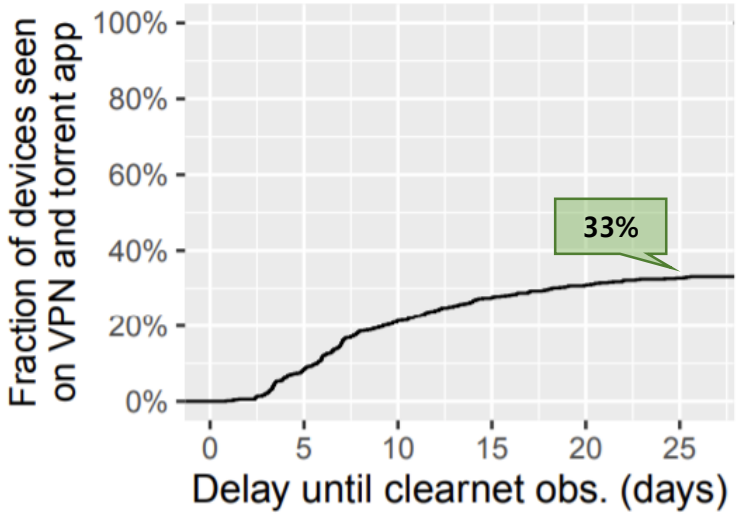
*Unique device: unique combination of app and device name Or IP address

CS2: Re-identification of VPN users

❖ Goal: *Determine the clearnet IP* (and geographic location) of devices masked by a VPN*

- Targeted ads to µTorrent, BitTorrent, tTorrent and atorrent app
  - Well known for trading copyrighted and illicit content
  - Many torrent users mask their activities **behind VPN**
  - (Android only)

- Clearnet identification
  - Using Ad ID
  - Using Ad ID + geographic location



*Clearnet IP: an address assigned by an ISP that maintains DHCP records and accurate billing information

Discussion

- Privacy-preserving solution
  - User side
    - Reset web storage and caching regularly Or just disable them
    - Use cellular network or a VPN over WiFi
  - OS platform side
    - Re-consider cache, storage policy for webview
  - RTB platform side
    - Obfuscate the IP address and precise GPS location

- Conclusion
  - Advertisers can re-identify devices relatively successfully and inexpensively on RTB networks **even when users have taken steps to prevent it**
  - Surprising that marking features are allowed and how **difficult it is for users to protect themselves in ad networks**

Thank you

AD Tag HTML

```
<!-- This ad is part of a research project at
        Location A. ... //-->
<a href="https://www.example.tld">
<img id="splash_pixel" style="display:none"
        src="https://api.example.tld/pixel.gif?impression=
        ${impression_id}&device_identifier=${
        device_identifier}&${imp_id}&gps=${gps}&
        campaign=${campaign_id}..." /></a>
<script src="https://api.example.tld/mark.js">
</script>
```