YOU NEVER SURF ALONE. UBIQUITOUS TRACKING OF USERS' BROWSING HABITS.

Silvia Puglisi, David Rebollo-Monedero and Jordi Forné

Department of Telematics Engineering

Universitat Politecnica de Catalunya

AGENDA

- Background
- Contribution
- Modelling the user's footprint
- A metric of similarity
- Experimental results
- Conclusions

BACKGROUND

When users surf the web a network of personalisation services tracks their preferences by following their browsing habits.

These services combine information from different sources:

- profiles and accounts that users create,
- browser cookies,
- device fingerprinting and so on.

Online profiling often means collecting enough features to distinguish an individual among a group of candidates.

Knowing certain facts about an individual allows the add network (or an attacker) to verify assumptions and make predictions.

CONTRIBUTIONS

1. Introducing a model of the user online footprint.
2. Introducing a measure of similarity between the user profile and the observed advertising profile.
3. Measuring how quickly a user is uniquely identified and tracked by an advertising network.

MODELLING THE USER'S FOOTPRINT

We model the user's activity as series of events belonging to a certain identity.

Each event is a document containing different information.

Each document can be seen as a hypermedia document i.e. an object possibly containing graphics, audio, video, plain text and hyperlinks.

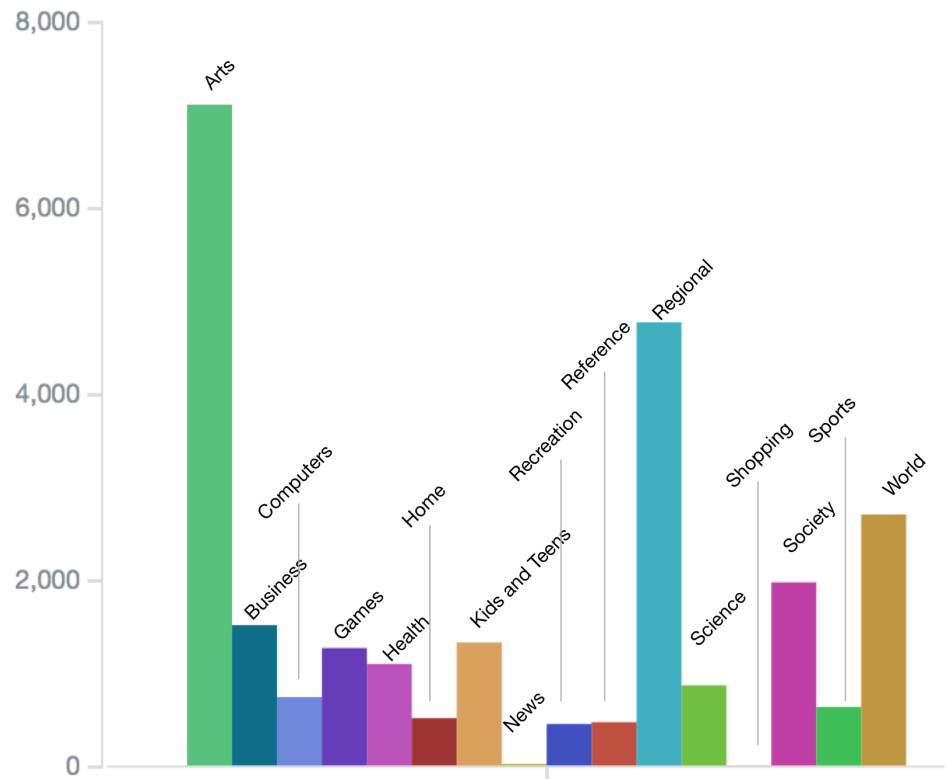
We call the hyperlinks selectors and we use these to build the connections between documents.

JSON


```
{
  "_index": "events",
  "_type": "browse",
  "_id": "AU5oiL6hZkB_U5MYIC_t",
  "_score": null,
  "_source": {
    "1-abs-norm": 0.790864961392389,
    "rd-party-requests": {
      "comment": "",
      "creator": {
        "comment": "",
        "version": "2.1.0-beta-1-legacy",
        "name": "BrowserMob Proxy"
      },
      "version": "1.2",
      "entries": [
        {
          "comment": "",
          "serverIPAddress": "192.241.188.103",
          "pageref": "https://storypirates.kindful.com/s/support-a-school/",
          "startedDateTime": "2015-07-07T14:40:58.929+02:00",
          "cache": {},
          "request": {
            "comment": "",
            "cookies": [
              {
                "comment": "",
                "name": "_session_id",
                "value": "71e2fa5736e49574ce13306614a566ef"
              }
            ],
            "url": "https://storypirates.kindful.com/s/support-a-school/",
            "queryString": [],
            "headers": [],
            "headersSize": 471,
```

We aggregate keywords each time the user creates a new event by visiting a different url.

These keywords constitute the user profile of interests.



With this model we assume that a particular category is weighted according to the number of times this has been counted in the user profile.

We define the profile of a user um as the Probability Mass Function:

$$\text{PMF } p_m = (p_{m,1}, \ldots, p_{m,L})$$

A histogram of relative frequencies of tags across the set of tag categories T.

The profile of an ad is defined as the PMF:

$$\text{PMF } q_n = (q_{n,1}, \ldots, q_{n,L})$$

Where q_{n,l} is the percentage of tags belonging to the category l which have been assigned to this specific advertising item.

A METRIC OF SIMILARITY

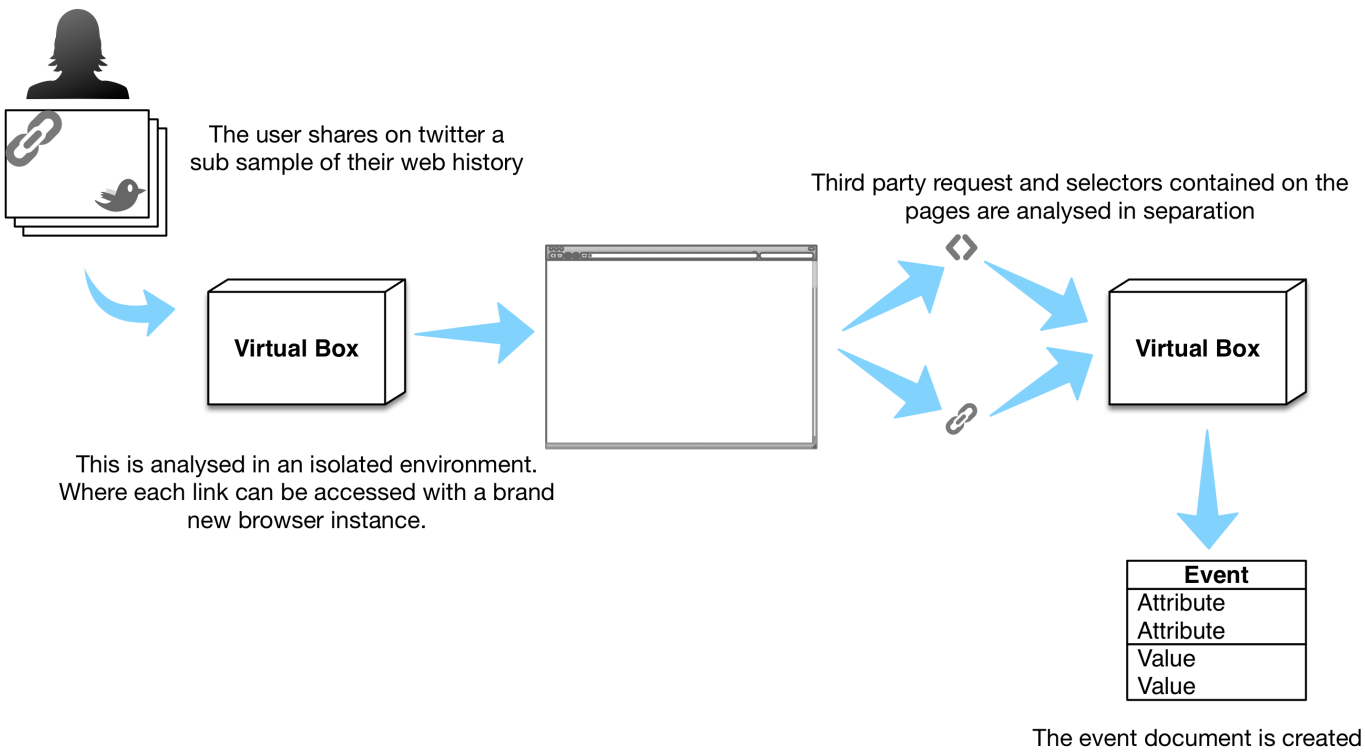
We consider the third party advertising network to operate like a recommendation system suggesting products or services based on the users' preferences.

We assume that the ad server suggest advertising based on a measure of similarity.

We use the 1 − norm between the user and the advertisting profile as a measurement of how the advertising network is tracking the user profile:

$$\|p_m, q_n\|_1 = \sum_i p_{mi} - q_{ni}$$

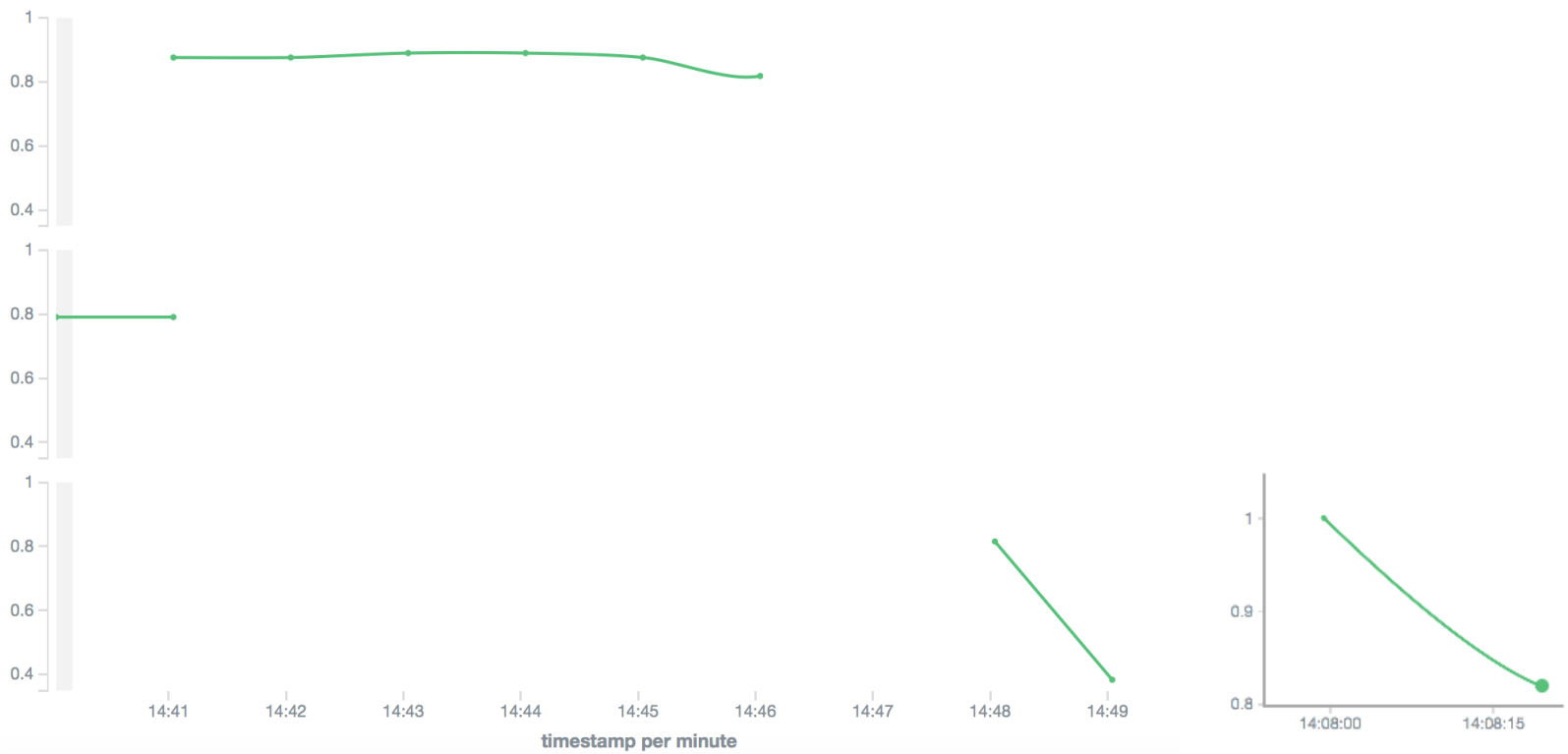
EXPERIMENTAL METHODOLOGY



WHY WE CRAWLED TWITTER FEEDS

"https://blog.twitter.com/2014/introducing-the-website-tag-for-remarketing"

EXPERIMENTAL RESULTS



CONCLUSIONS

Web tracking happens very quickly in a few subsequent visits to websites in a large advertising network.

Hypermedia models are able to capture both connections between entities as weel as individual events features and metrics.

FUTURE DEVELOPMENT

We would like to introduce a full set of metrics to measure the distance between the advertising and the user's profile:

- 2-norm,
- KL-Divergence between the advertising profile and the observed user profile,
- Fisher information.

We are interested in measuring how social networks sharing buttons and/or commenting services, included on websites, are able to track users even when these have not signed in with their account.

We are interested to measure how Privacy Enhancing Techniques affect advertising networks.

"I do not want to live in a world where everything I do and say is recorded. That is not something I am willing to support or live under."

Edward Snowden