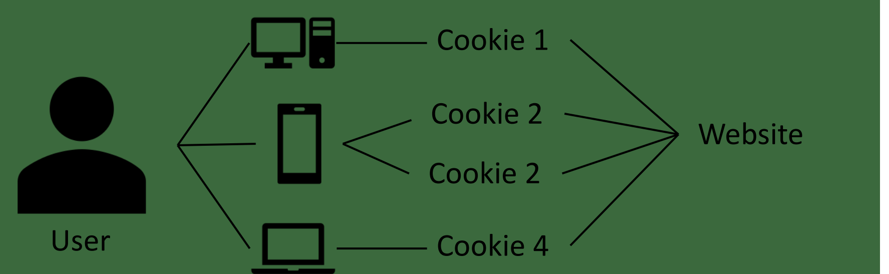
Household estimation using web traffic data

Tommaso Ricci
Msc in Big Data

UNIVERSITY of STIRLING

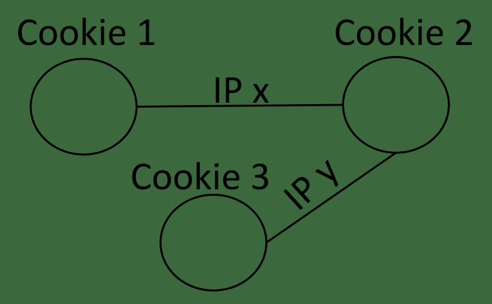
Problem and Aim

Web sites cookies are known for their unreliability and inability to detect users across different devices. When users are not registered or logged in, assigning traffic to the right entity, can be challenging.



Methodology

The problem is tackled by building a device graph: the nodes represent cookies and they are linked if they have shared the same IP at some point in time. After edge manipulation (filtering and summing parallel edges), community detection is performed on the graph. The resulting communities represent users. All the traffic coming from cookies inside the community can be now associated with the same user.



Same users on different devices can be detected using web traffic of a single internet domain using unsupervised methods.

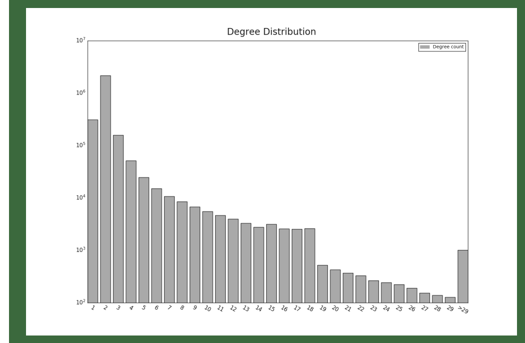
Learn more.

In collaboration with

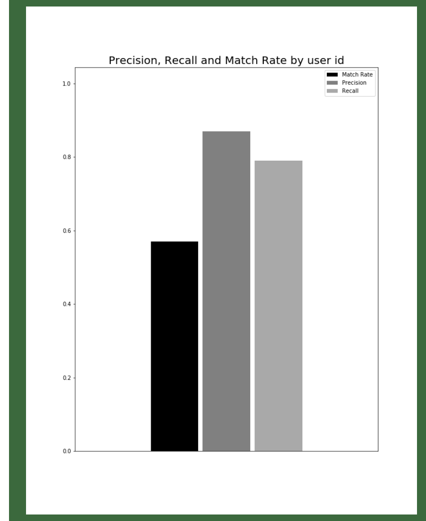
tvsquared

Initial Results

Early results shows that a good amount of users (57 %) can be detected and separated from other users. Most of the cookies can be correctly associated together (average recall=0.80) but separating them from cookies belonging to other users is still an open challenge.



Degree Distribution before edge filtering.
The graph consisted of 2.8 Million nodes and 4.2 Million edges.



W is the estimated community
V is the community of the test data

$$\text{Precision} = \frac{|V \cap W|}{|W|}, \text{Recall} = \frac{|V \cap W|}{|V|}.$$

There is a match if both Precision and Recall are above .50
The Match rate is the percentage of matches over the test data.

Next Steps

The second phase of the project will cover data coming from different internet domains, trying to reconstruct entire households instead user devices ownership.

For further information contact: *Tommaso Ricci*

tor00013@students.stir.ac.uk