CrowdSurf

Empowering Transparency in the Web

25 Aug 2016,
ACM SIGCOMM,
Florianopolis

Hassan Metwalley
Stefano Traverso
Marco Mellia
Stanislav Miskovic
Mario Baldi

POLITECNICO DI TORINO

Symantec

Introduction

Do you know what you HTTP?

Example

Web tracking

Thousands o ect our data

- ❑ Browsing h
- ❑ Religious, ferences

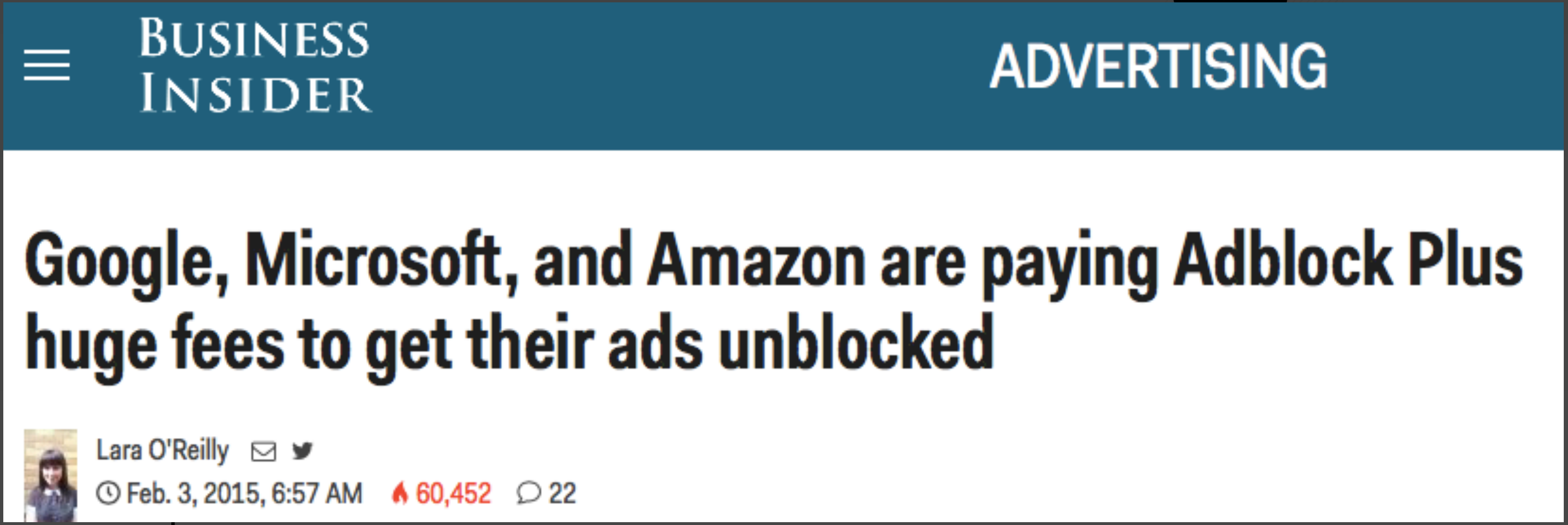
- ❑ On averag **as soon as the browser st**
- ❑ Some trac [1]
- ❑ **71% of wel** racker [1]

[1] Metwalley, H. et al. "*The On* *rements*", TMA 2015

The Open Question

How to **know** and **choose** which **services our data is exchanged** with and how?

Partial solutions



BUSINESS INSIDER

ADVERTISING

Google, Microsoft, and Amazon are paying Adblock Plus huge fees to get their ads unblocked

Lara O'Reilly

Feb. 3, 2015, 6:57 AM 60,452 22

A New System

Goal

Let **users** re-gain visibility and **control** on the **information** they exchange with **Web services**

Design Principles

- ❑ Holistic
  working in any scenario
- ❑ Client-centric
  available on any kind of device
- ❑ Practical, not revolutionary
  use existing technology
- ❑ Crowd-sourced
  knowledge built on a community of users
- ❑ Automatic
  little engagement of the user
- ❑ Privacy-safe
  never compromise users' privacy

CrowdSurf

CrowdSurf

Cloud

- A **controller** collects information about the services users visit
  - Explicit -> their opinion
  - Implicit -> traffic samples
- Users' contributions processed by **data-analyzers** and the **advising community**
- Results = **suggestions** about the reputation of services

Client

- Users download the suggestions they like
- the **CrowdSurf Layer** translates them into **rules**
- Rules = **actions** on users' traffic
  - Regexp + action

CrowdSurf Controllers

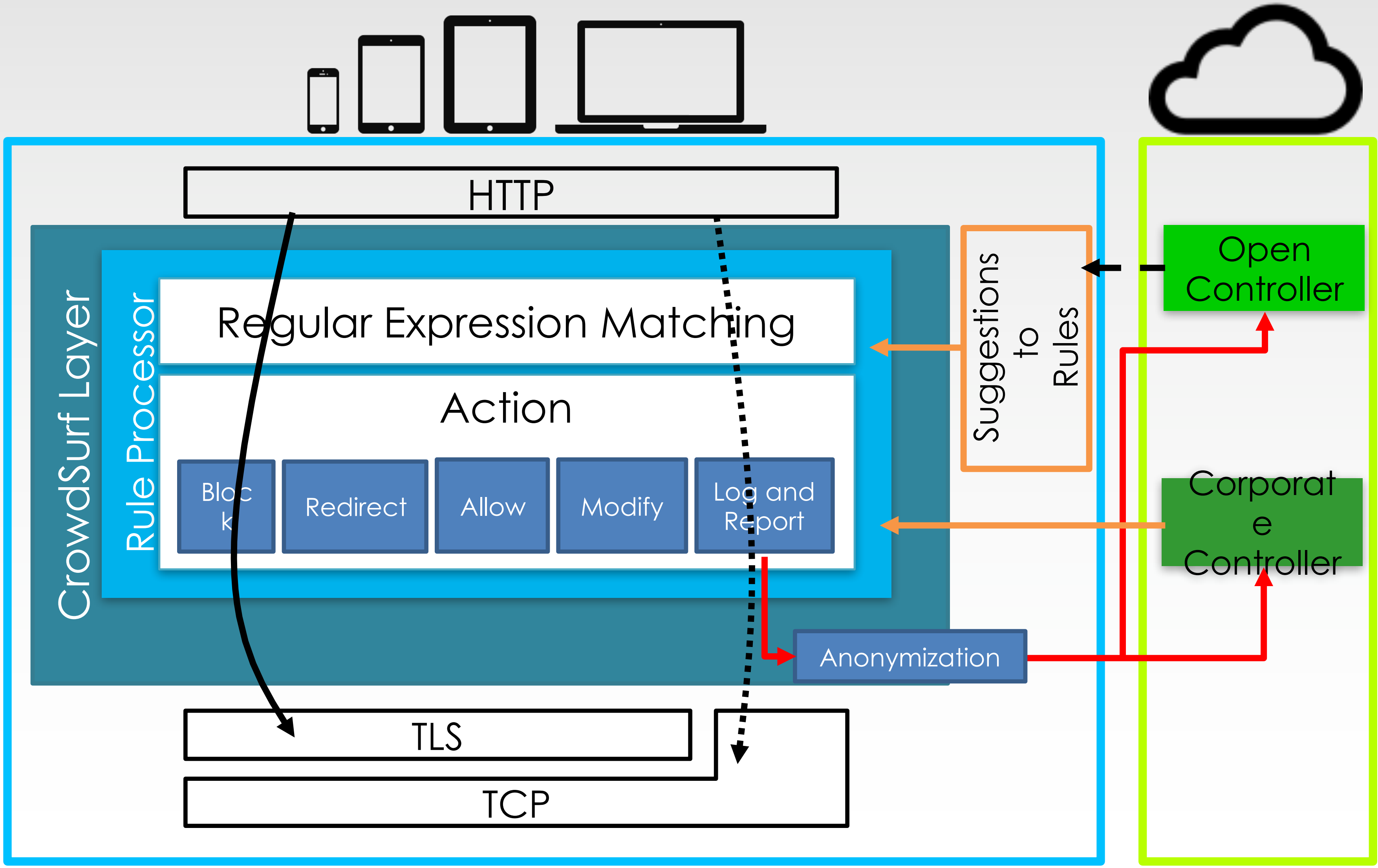
Open Controller

- **Collaborative approach**
- Users improve the wisdom of the system
  - Traffic samples and opinions
  - Build data analyzers and suggestions

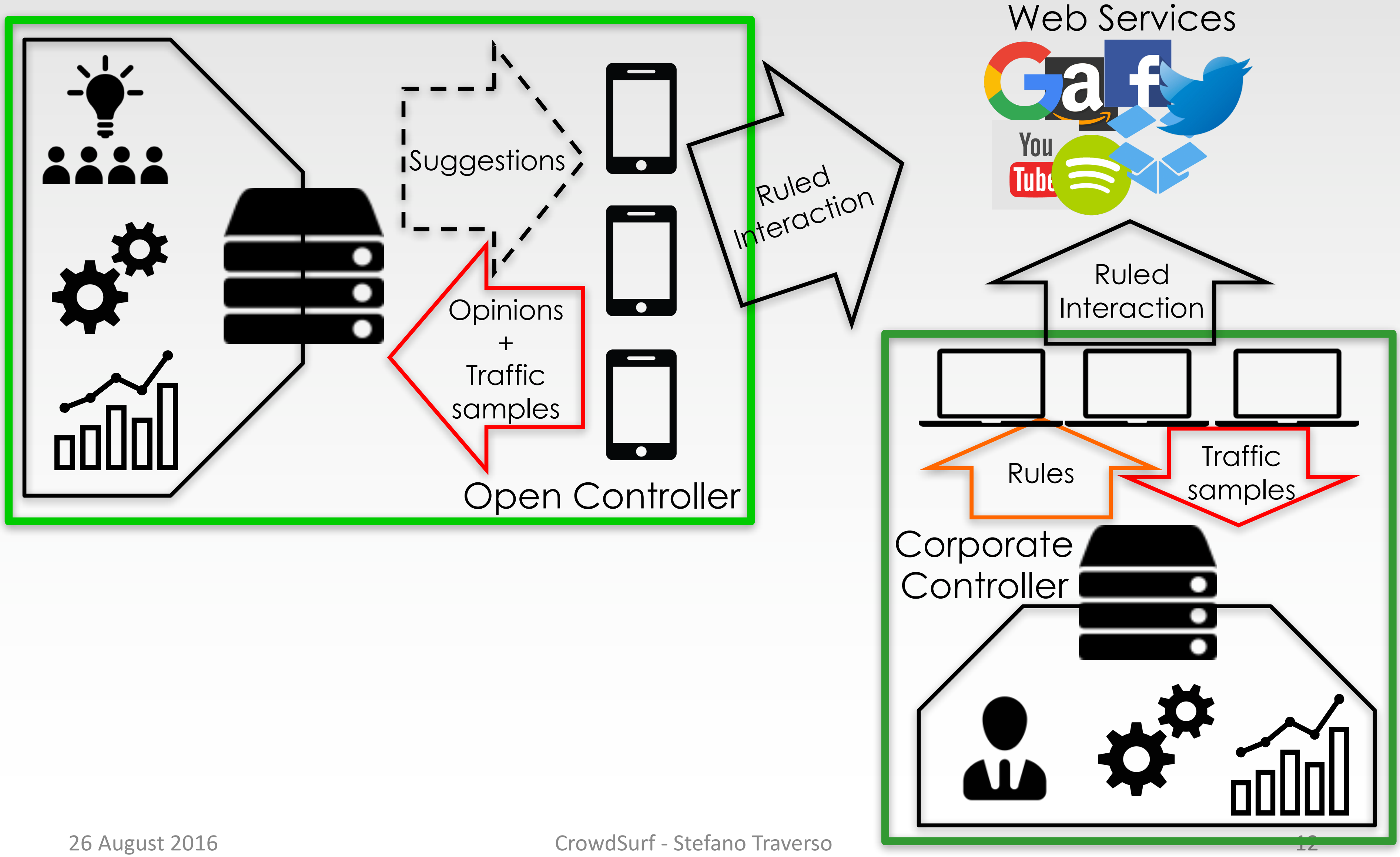
Corporate Controller

- **Builds directly rules** for employees
- Employees can not customize rules
- All devices follow the same rules

The CrowdSurf Layer



CrowdSurf in a picture



Proof of Concept

Prototype

Controller

- □ Java-based web service
- □ Communicates with CrowdSurf devices
- □ Hosts a data analyzer for identification of tracking sites
- □ Collects traffic samples
- □ Distributes suggestions

Client

- □ Implemented as a Firefox plugin
- □ Supports *block*, *redirect*, *log&report*

Example of Data Analyzer: Automatic Tracker Detector

Unsupervised methodology to identify third-party trackers [2]

❑ Observation:

  ❑ trackers usually embed UIDs as URL parameters

❑ Procedure:

1. Input: HTTP traffic samples provided by CS users
2. Take all HTTP queries to third-party services

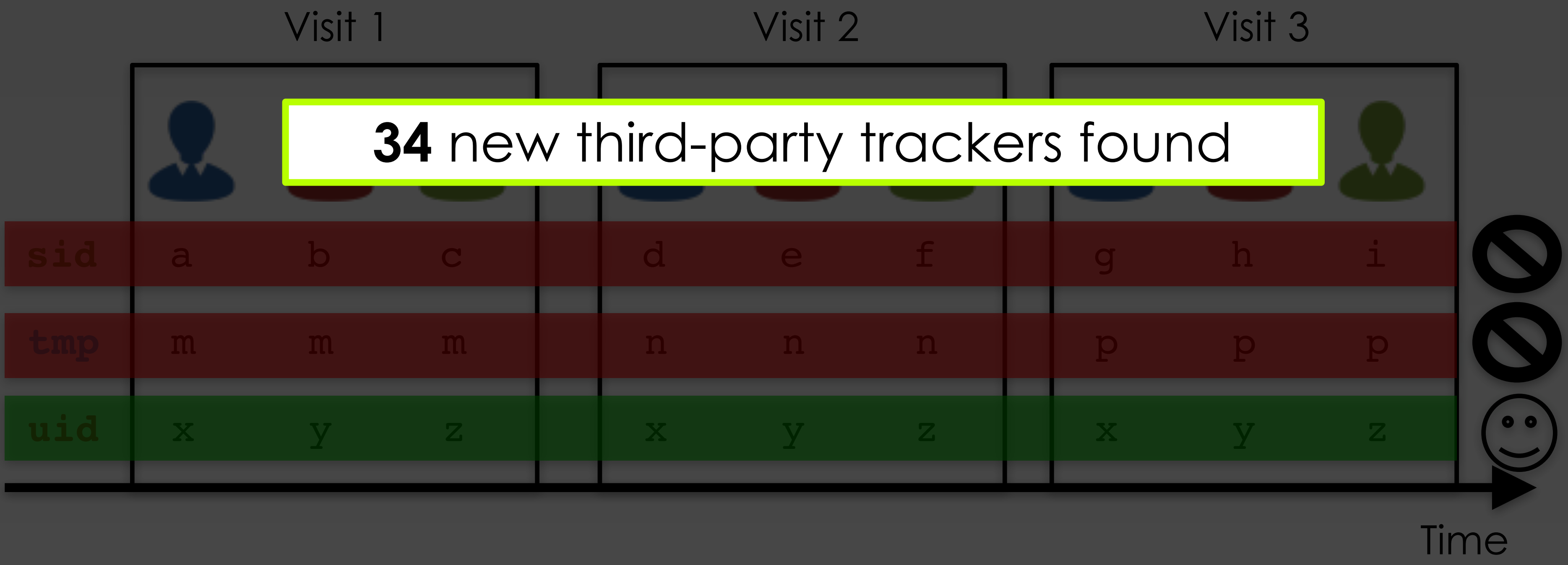
   http://acmetrack.com/query?**key1**=X&**key2**=Y

3. Extract keys (**key1**, **key2**) and their values
4. Check the presence of key values uniquely associated to the users

[2] Metwalley, H. et al "Unsupervised Detection of Web Trackers", IEEE Globecom 2015

Example of Data Analyzer: Automatic Tracker Detector

```
http://acmetrack.com/query?sid=X&tmp=Y&uid=Z
```



	Visit 1			Visit 2			Visit 3			
sid	a	b	c	d	e	f	g	h	i	🚫
tmp	m	m	m	n	n	n	p	p	p	🚫
uid	x	y	z	x	y	z	x	y	z	☺

34 new third-party trackers found

Performance Implications of running CrowdSurf

Different user profiles

Paranoid Profile

- □ **Blocks**
  - □ adv/tracking
  - □ JS code
- □ **Does not report** traffic samples

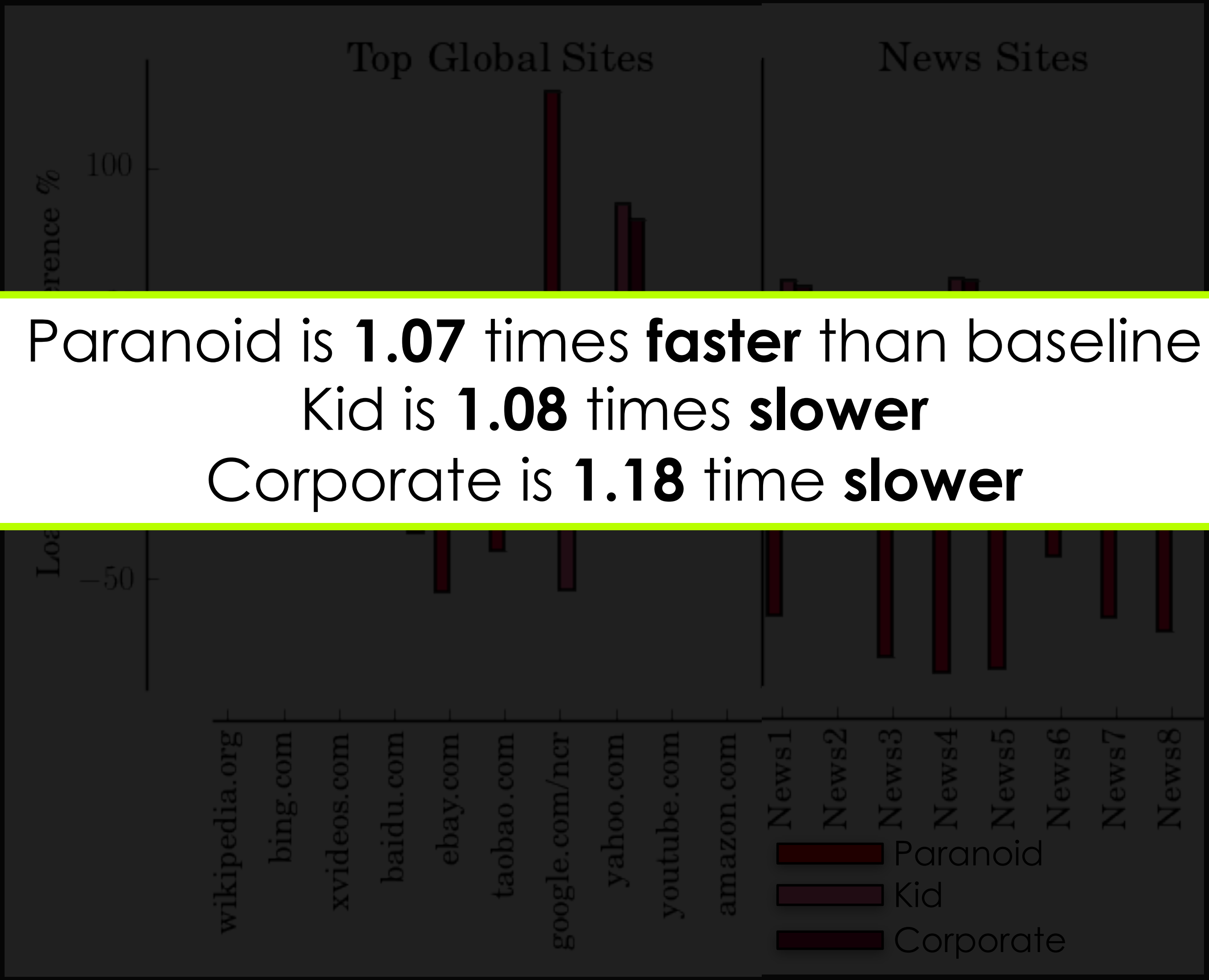
Kid Profile

- □ Activates **child protection rules**
- □ **Reports** traffic to trackers

Corporate Profile

- □ **Redirects** *search.google.com* to *search.bing.com*
- □ **Blocks** social networks, e-commerce sites, trackers
- □ **Reports** acitivity on DropBox

Impact on Web site loading time



Paranoid is **1.07** times **faster** than baseline
Kid is **1.08** times **slower**
Corporate is **1.18** time **slower**

Conclusion

Open Problems

- Lot of details to consider
- Design/develop/stardardize a new network layer
- Protecting users’ privacy
  - Anonymizing HTTP/S traffic
- Usability
- Involve users to join
- Protection from malicious biases

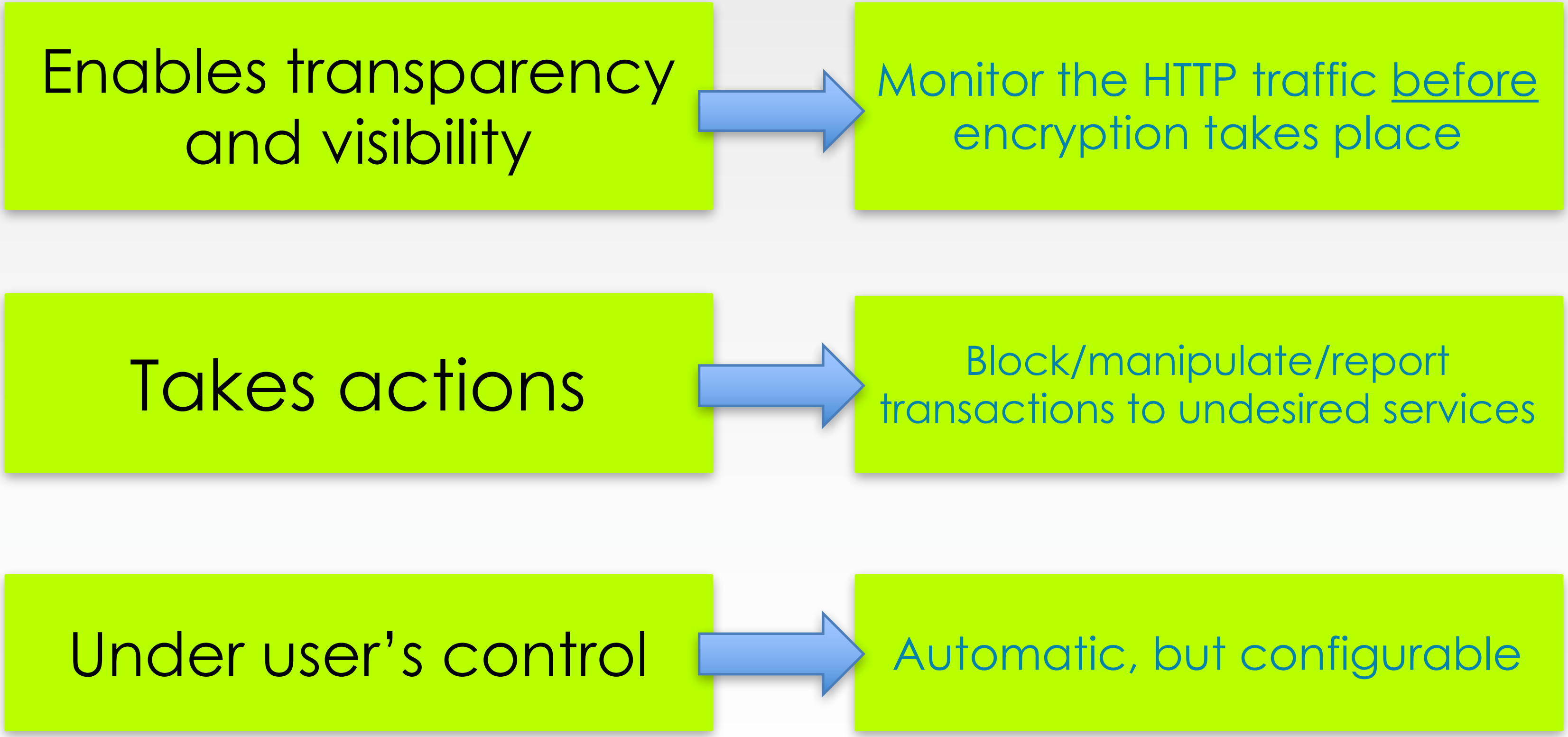
CrowdSurf

Holistic, crowd-sourced system for the auditing of the information we expose in the Web

https://www.myermes.com

Thank you!

Need a new model that…



Enables transparency and visibility → Monitor the HTTP traffic before encryption takes place

Takes actions → Block/manipulate/report transactions to undesired services

Under user's control → Automatic, but configurable

Example of Data Analyzer: Automatic Tracker Detector

Automatic Tracker Detector

vs

Dataset

HTTP trace from ISP running Tstat

- ❑ 10 days of October 2014
- ❑ ~19k monitored users
- ❑ ~240k HTTP transactions per day

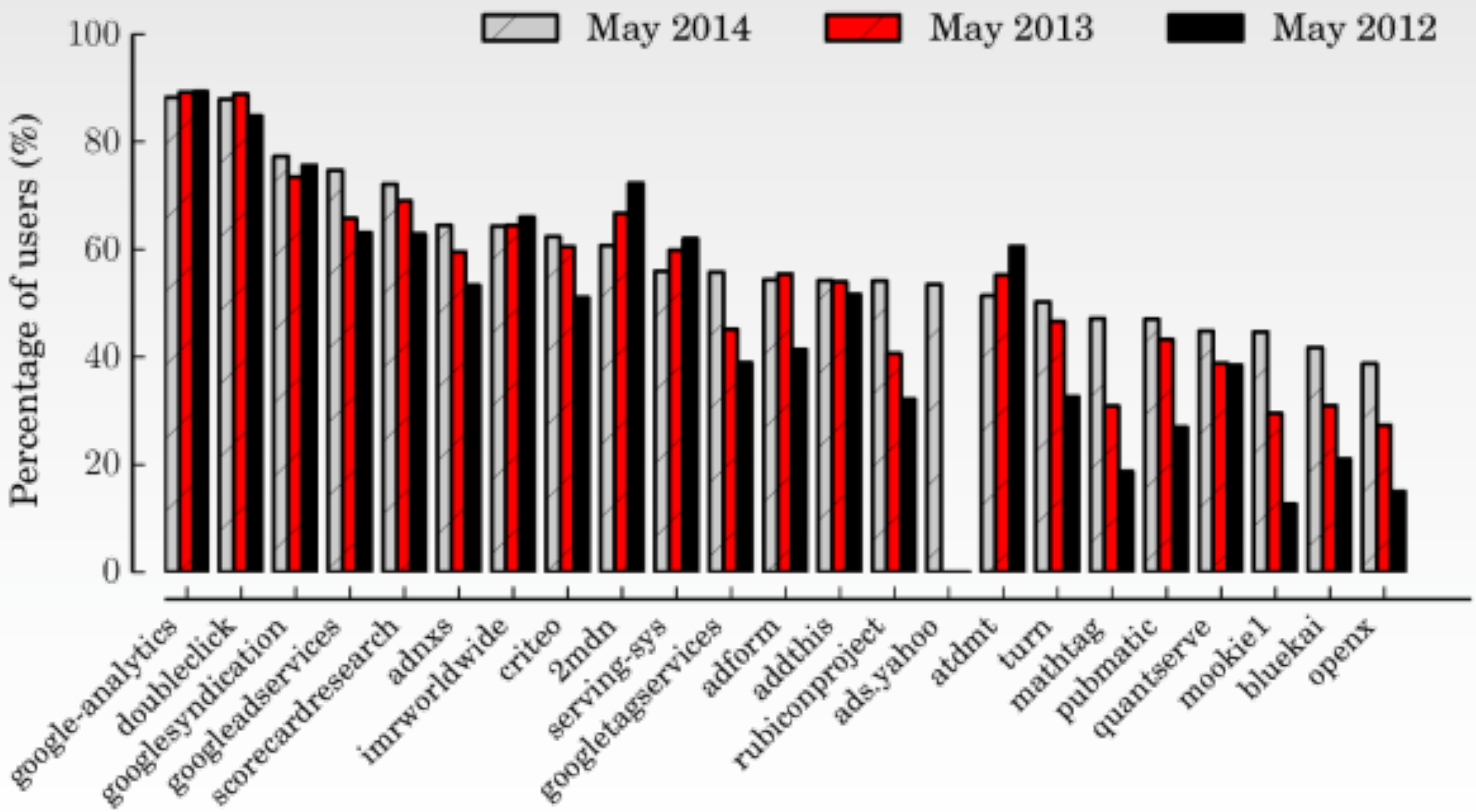
34 new third-party trackers found

	Thir	
News1	cl.adform.net	xid
	atemda.com	bidderuid
	x.bidswitch.net	user_id
	www.77tracking.com	rand
	rack.movad.net	us
	ovo01.webtrekk.net	cs2
	dis.criteo.com	uid
	p.rfihub.com	bk-uuid
	ib.adnxs.com	xid

	bedded Third-arty Trackers
Portal1	26
News1	13
E-commerce1	12
E-commerce2	9
E-commerce3	4
Portal2	4
Porn	3
Sportnews	1
SearchEngine	1

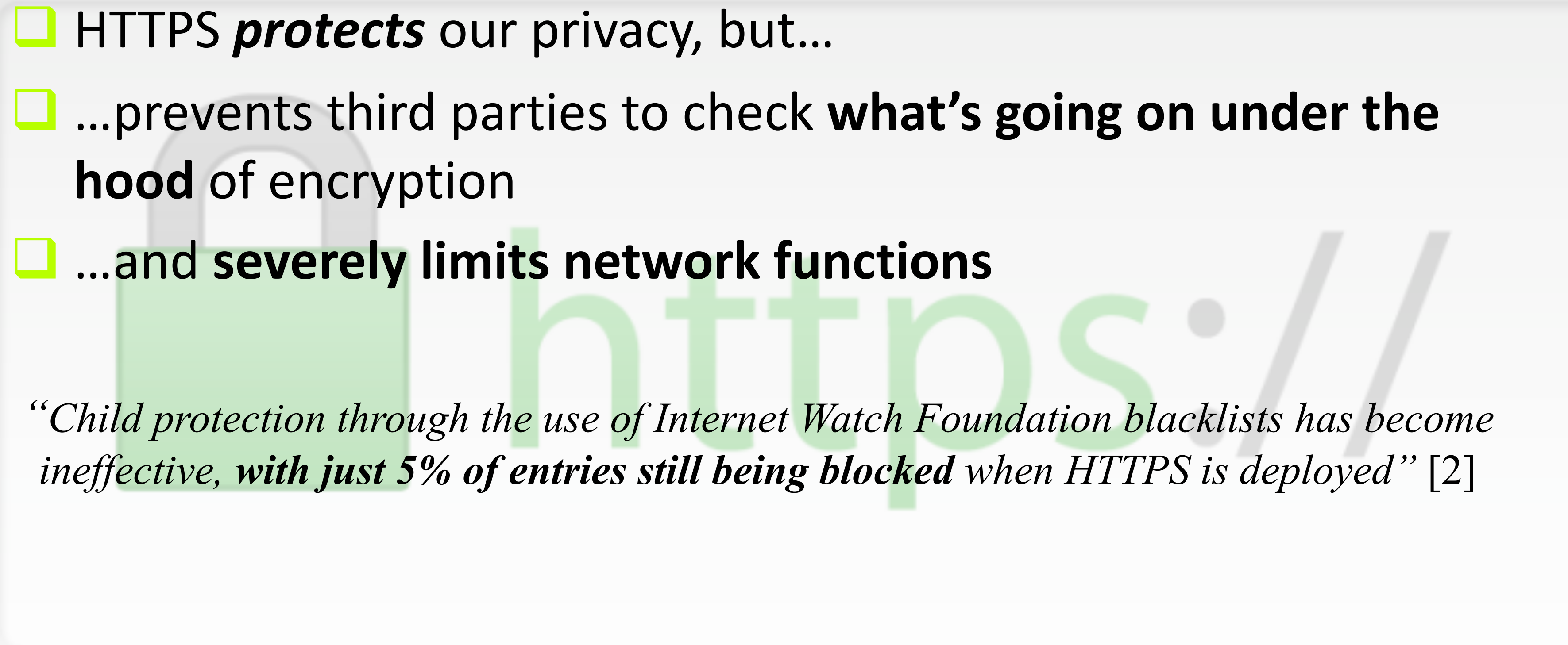
Example

A growing business around our data



[3] Metwalley, H. et al. "*The Online Tracking Horde: A View from Passive Measurements*", TMA 2015

Loss of visibility and control

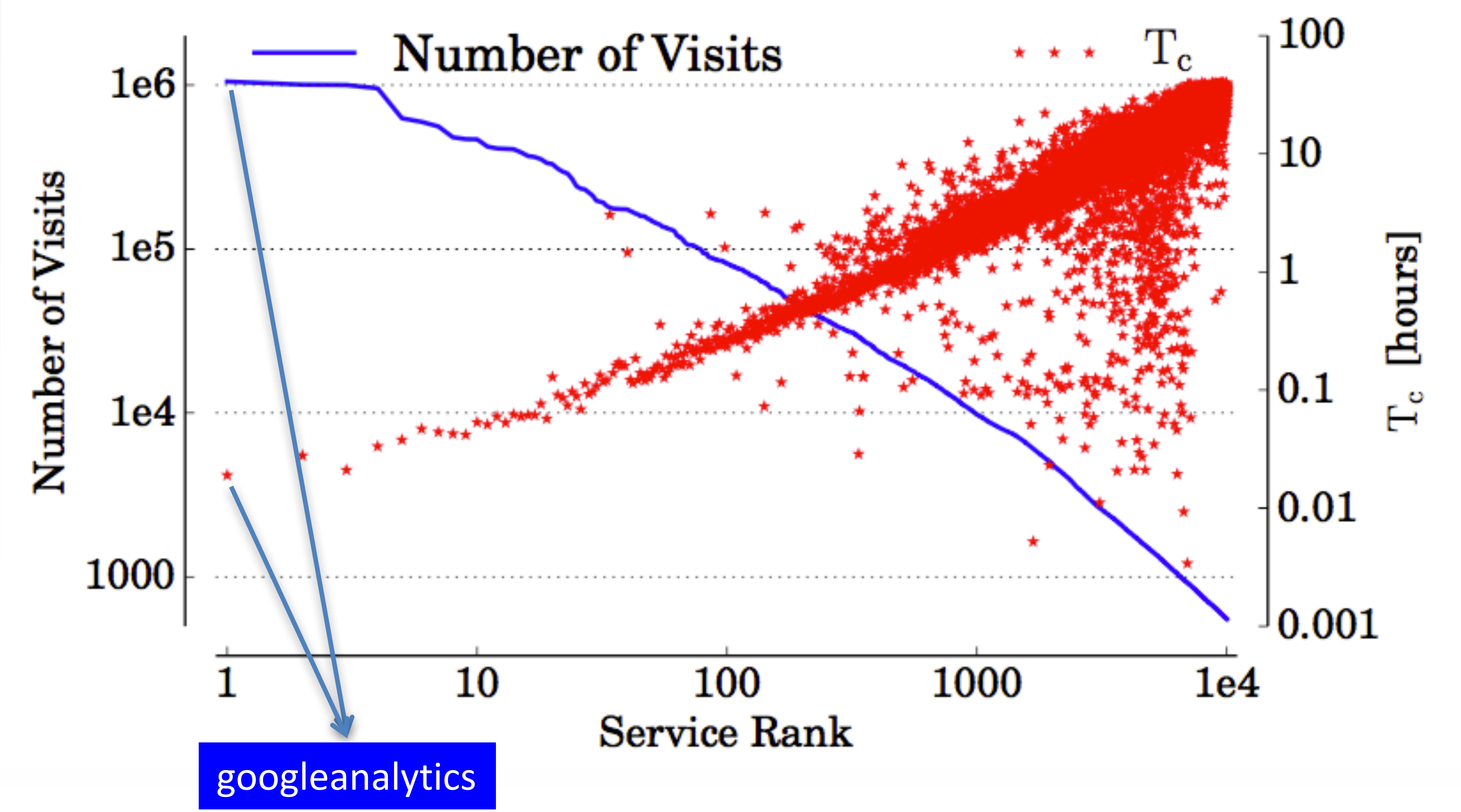


- HTTPS ***protects*** our privacy, but…
- …prevents third parties to check **what's going on under the hood** of encryption
- …and **severely limits network functions**

*"Child protection through the use of Internet Watch Foundation blacklists has become ineffective, **with just 5% of entries still being blocked** when HTTPS is deployed"* [2]

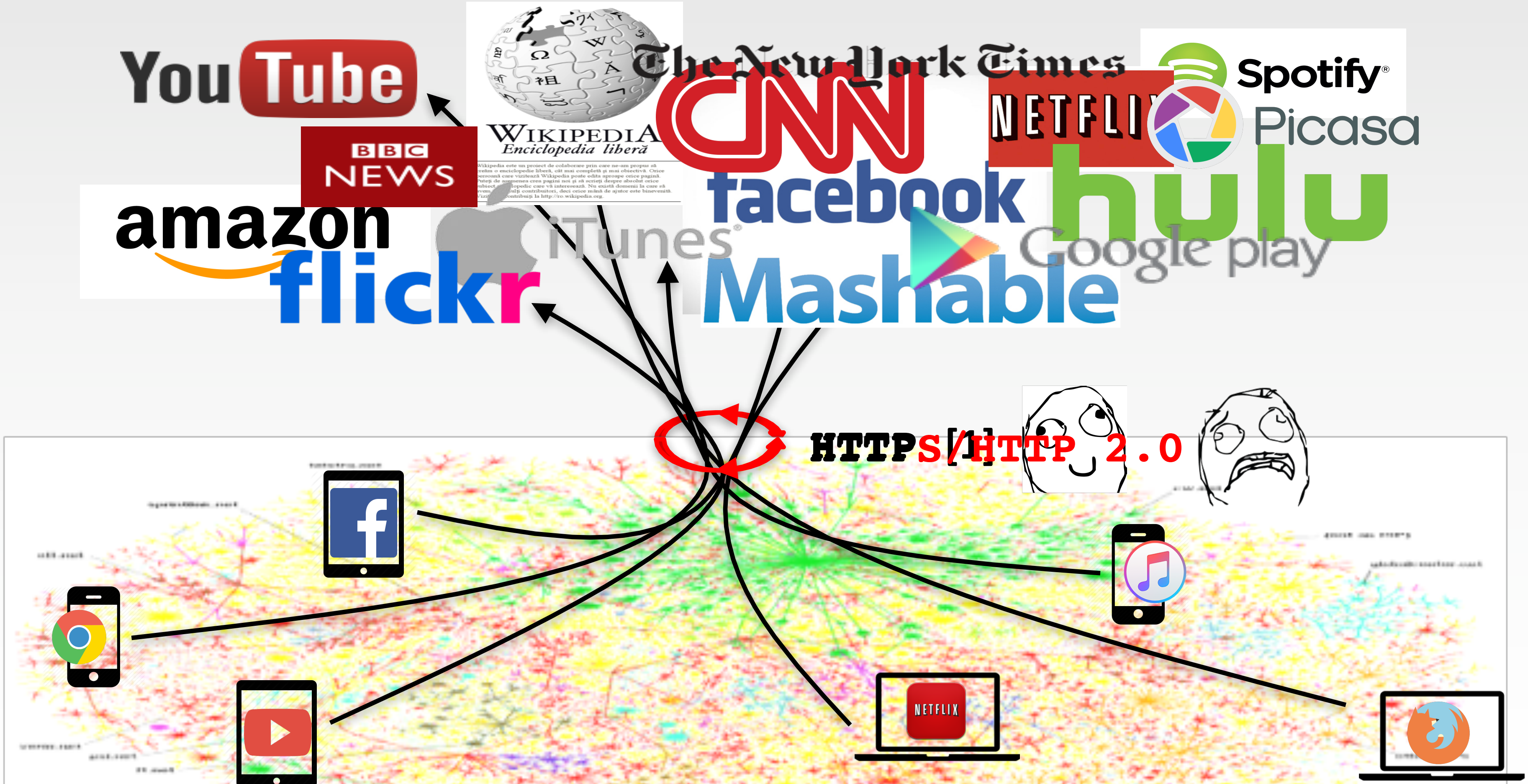
[2] Naylor, D. et al. *"The Cost of the "S" in HTTPS"*, CoNEXT 2014

Time to collect a dataset



googleanalytics

Monitoring the Web



HTTP[1]

HTTPS/HTTP 2.0

[1] Popa, L. et al., “HTTP As the Narrow Waist of the Future Internet,” *ACM HotNets*, 2010

CrowdSurf Controllers

Open Controller

- **Collaborative approach**
- Users improve the wisdom of the system
  - Traffic samples and opinions
  - Build data analyzers and suggestions

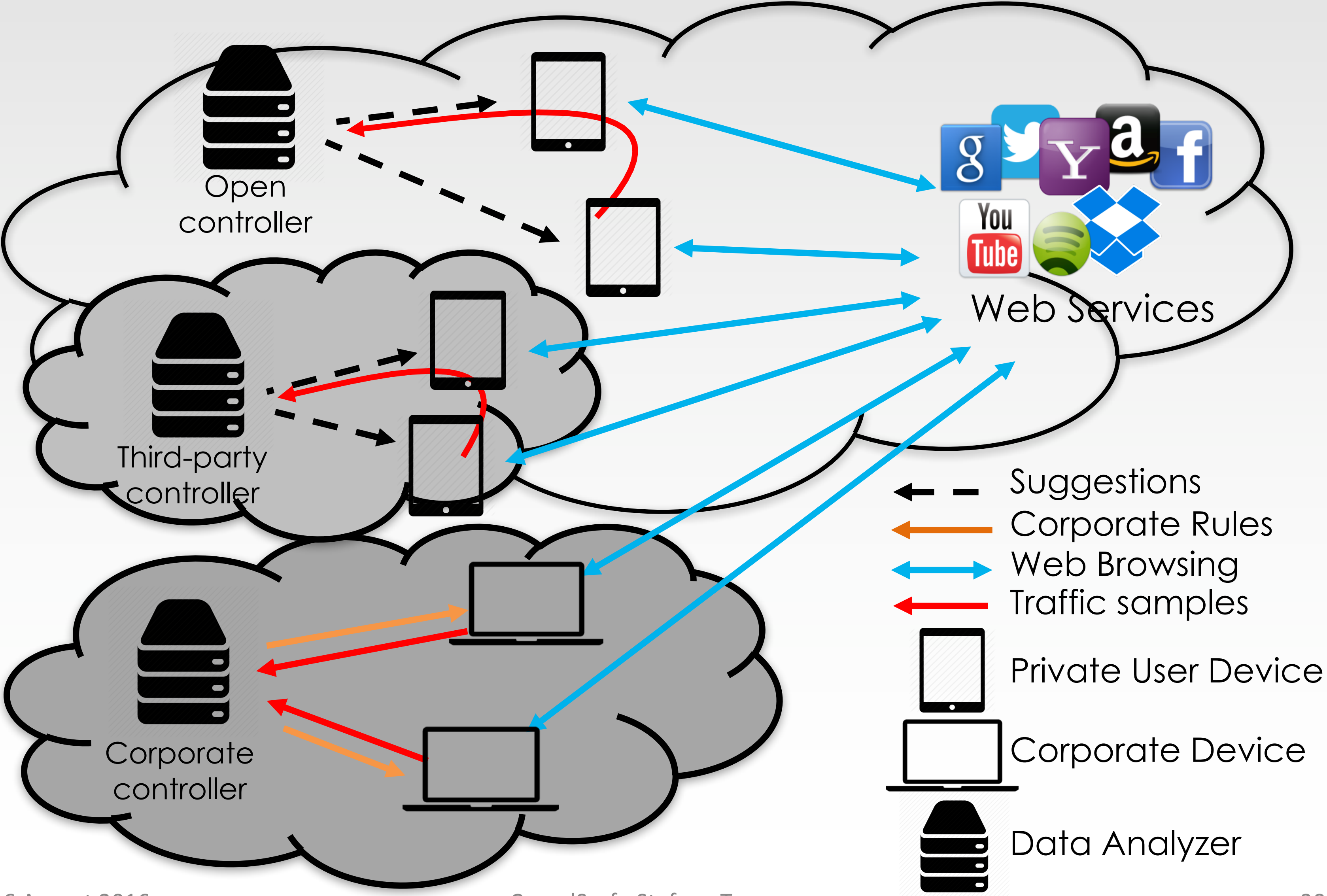
Third party Controller

- Suggestions for **commercial purposes**
- Opens to a market of suggestions

Corporate Controller

- **Builds directly rules** for employees
- Employees can not customize rules
- All devices follow the same rules

CrowdSurf in a picture


Open controller
Web Services
Third-party controller
Corporate controller
Suggestions
Corporate Rules
Web Browsing
Traffic samples
Private User Device
Corporate Device
Data Analyzer