

The Quest for Algorithm Diversity

<https://facebook.tracking.exposed>

Claudio Agosti @_vecna

and the fbtrex team: <https://facebook.tracking.exposed/about>

Facebook creates two big techno-social problems

1. Algorithm Hegemony

**2. Privatized Info so that FB can
assess society better than society itself**

**This project offers a practical way to study the problem and
raise critiques.**

Concept

- * A web extension

<https://github.com/tracking-exposed/web-extensions>

- * Collect the HTML snippet of the posts Facebook sends you (only if they are public)



- * They enter in a pipeline of analysis and reduction

supporters
411

timelines
158310

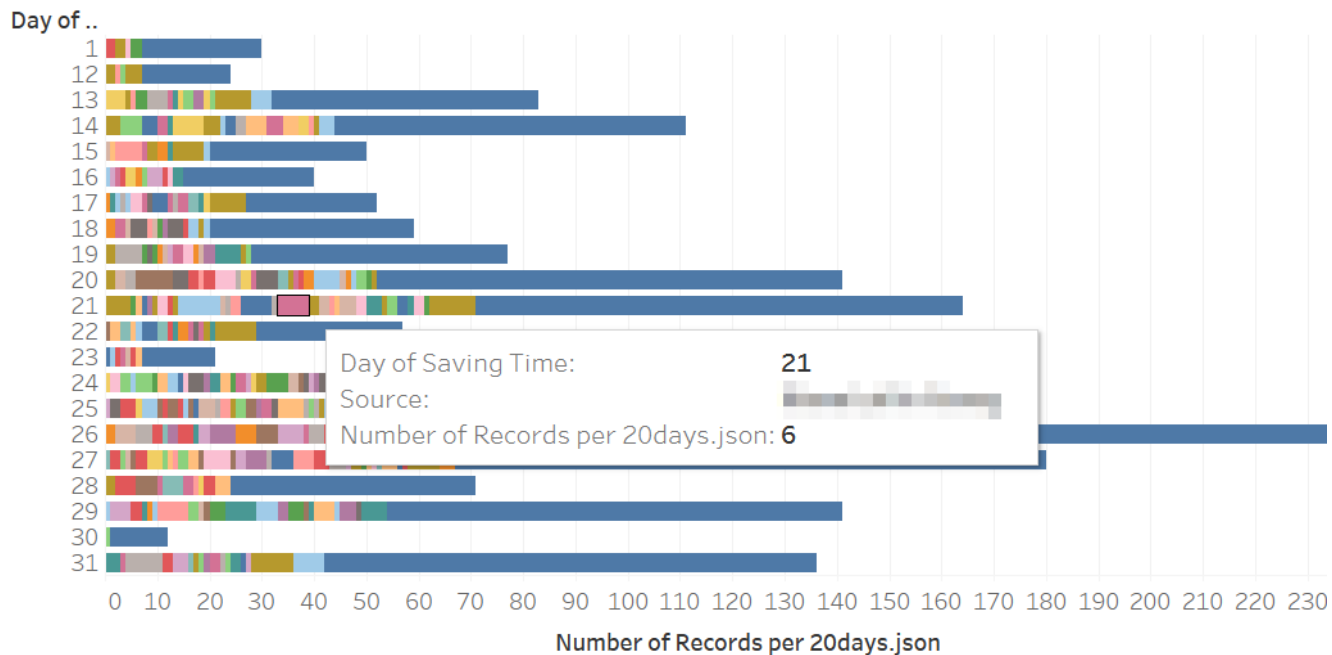
impressions
2454118

htmls
1600037

accesses
79582

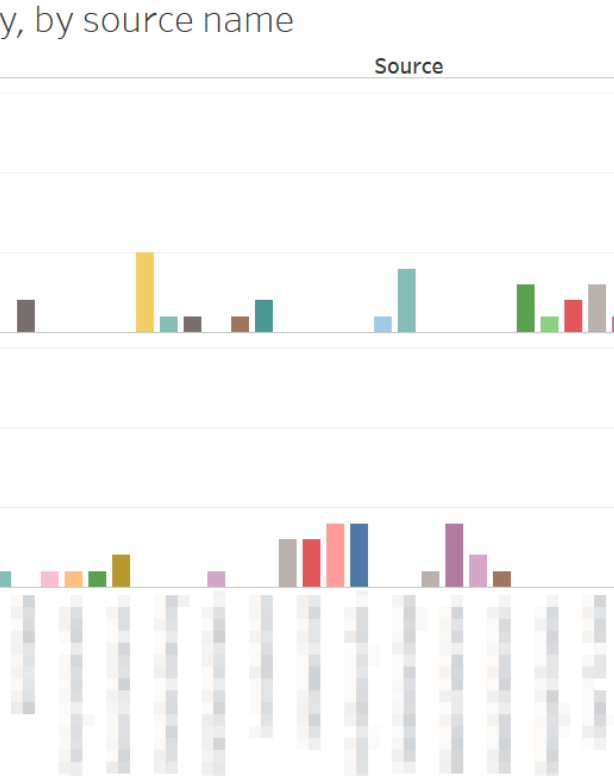
Algorithm Hegemony

Facebook controls your information feed. fbtrex maps this impact:



[https://facebook.tracking.exposed/api/v1/htmls/\\$supporterId/days/\\$numberOfDays](https://facebook.tracking.exposed/api/v1/htmls/$supporterId/days/$numberOfDays)

t type, source, keyw



Algorithm Hegemony

A post can show up at the top of a newsfeed despite chronological order.

How is it being favored by the algorithm?

A post show up only once despite multiple refreshes.

How is it being penalized by the algorithm?

Algorithm Hegemony

9 Oct 2016 19:20	9 Oct 2016 19:30	9 Oct 2016 19:40	9 Oct 2016 19:50
1 ^o 9 Oct '16, 19:18	4 ^o 9 Oct '16, 19:19	1 ^o 9 Oct '16, 19:36	1 ^o 9 Oct '16, 19:47
5 ^o 9 Oct '16, 19:11	9 ^o 9 Oct '16, 19:11	9 ^o 9 Oct '16, 19:19	6 ^o 9 Oct '16, 19:42
4 ^o 9 Oct '16, 19:11	8 ^o 9 Oct '16, 19:11	13 ^o 9 Oct '16, 19:18	7 ^o 9 Oct '16, 19:36
7 ^o 9 Oct '16, 19:03	13 ^o 9 Oct '16, 19:03	15 ^o 9 Oct '16, 19:11	15 ^o 9 Oct '16, 19:19
10 ^o 9 Oct '16, 19:00	16 ^o 9 Oct '16, 19:00	14 ^o 9 Oct '16, 19:11	18 ^o 9 Oct '16, 19:18
8 ^o 9 Oct '16, 18:57	14 ^o 9 Oct '16, 18:57	17 ^o 9 Oct '16, 19:03	20 ^o 9 Oct '16, 19:11
6 ^o 9 Oct '16, 18:49	10 ^o 9 Oct '16, 18:49	22 ^o 9 Oct '16, 19:00	19 ^o 9 Oct '16, 19:11
15 ^o 9 Oct '16, 18:30	21 ^o 9 Oct '16, 18:30	18 ^o 9 Oct '16, 18:57	23 ^o 9 Oct '16, 19:03
9 ^o 9 Oct '16, 18:28	15 ^o 9 Oct '16, 18:28	16 ^o 9 Oct '16, 18:49	24 ^o 9 Oct '16, 18:57
13 ^o 9 Oct '16, 18:25	19 ^o 9 Oct '16, 18:25	19 ^o 9 Oct '16, 18:28	21 ^o 9 Oct '16, 18:49
22 ^o 9 Oct '16, 18:06	23 ^o 9 Oct '16, 18:01	25 ^o 9 Oct '16, 18:25	25 ^o 9 Oct '16, 18:28
17 ^o 9 Oct '16, 18:01	25 ^o 9 Oct '16, 17:48	23 ^o 9 Oct '16, 16:10	9 ^o 9 Oct '16, 13:47
19 ^o 9 Oct '16, 17:48	17 ^o 9 Oct '16, 16:10	5 ^o 9 Oct '16, 13:47	4 ^o 9 Oct '16, 12:11



Generic Facebook Lawyer:

“There are many good reasons for an algorithm to do that!”

Investigation complexity

fbtrex analyzed “what the supporter get from FB”

It depends from five (high level) variables:

**How much the
supporter
access to
Facebook**

**Which sources
the supporter
follows**

**What the
sources shares**

**The Facebook
Algorithm**

**The supporter
profile**

Algorithm Hegemony

4 fresh users, all of them:

With zero friends

Following the same 12 sources

Accessing to Facebook once per hour

Scrolling automatically Facebook for three minutes each hour

The only difference? few likes to different pages, to characterize the user

we released the dataset:

<https://medium.com/@trackingexposed/facebook-algorithm-and-impact-on-media-french-election-experiment-1-d760ed5a242f>

Algorithm Hegemony

We can't escape the bubble, but we can compare our bubble with others - and understand if the algorithm hides something from us.

Post Id	User Id / Saving Time									
	100016788883580					100016926932367				
	16	17	18	19	20	16	17	18	19	20
1605177452826936	■	■	■	■	■	■	■	■	■	■
1606576672687014	■	■	■	■	■	■	■	■	■	■
1606897375988277					■					■
1606981855979829					■				■	■
1607438325934182									■	■
1607489989262349	■	■	■	■	■	■	■	■	■	■
1607751319236216	■	■	■	■	■	■	■	■	■	■
1607830219228326	■	■	■	■	■	■	■	■	■	■
1607857359225612						■	■	■	■	■
1607883549222993	■	■	■	■	■	■	■	■	■	■
1607990609212287	■	■	■	■	■	■	■	■	■	■
1608071595870855						■	■	■	■	■
1608109832533698		■		■	■					
1608182745859740			■	■	■			■		■

Algorithm Hegemony

Is everyone seeing the same comments?

(spoiler: no)



Algorithm Hegemony

Metadata extracted

- 🌸 saved a day ago.
- 🌸 **id:** bafb04a7fc20afd21f95f07f58c479f[redacted]
- 🌸 ☒ **postType**
- 🌸 **type:** feed
- 🌸 ☒ **feedUTime**
- 🌸 **publicationUTime:** 1501430167
- 🌸 ☒ **feedBasicInfo**
- 🌸 **postId:** [redacted]
- 🌸 **permaLink:** [redacted]
- 🌸 **hrefType:** post
- 🌸 ☒ **feedText**
- 🌸 **source:** [redacted] compartió la publicación de The Hill.
- 🌸 ☒ **feedHref**
- 🌸 **externalHref:** http://hill.cm/44B6SVn
- 🌸 ☒ **imageAltText**

The user saw this



[redacted] compartió la publicación de The Hill.

· 56 min ·



The Hill Me gusta esta página

· 17 h ·

Hackers at a competition in Las Vegas successfully breached the software of U.S. voting machines in less than 90 minutes, exposing massive vulnerabilities in America's voting systems.



Hackers break into US voting machines in minutes at hacking competition

Some devices were found to have physical ports that could be used to attach devices containing malicious software. Others had insecure Wi-Fi connections, or were running outdated software with security vulnerabilities like Windows XP. thehill.com

Me gusta

ComentarCompartir

33 3

Comentarios

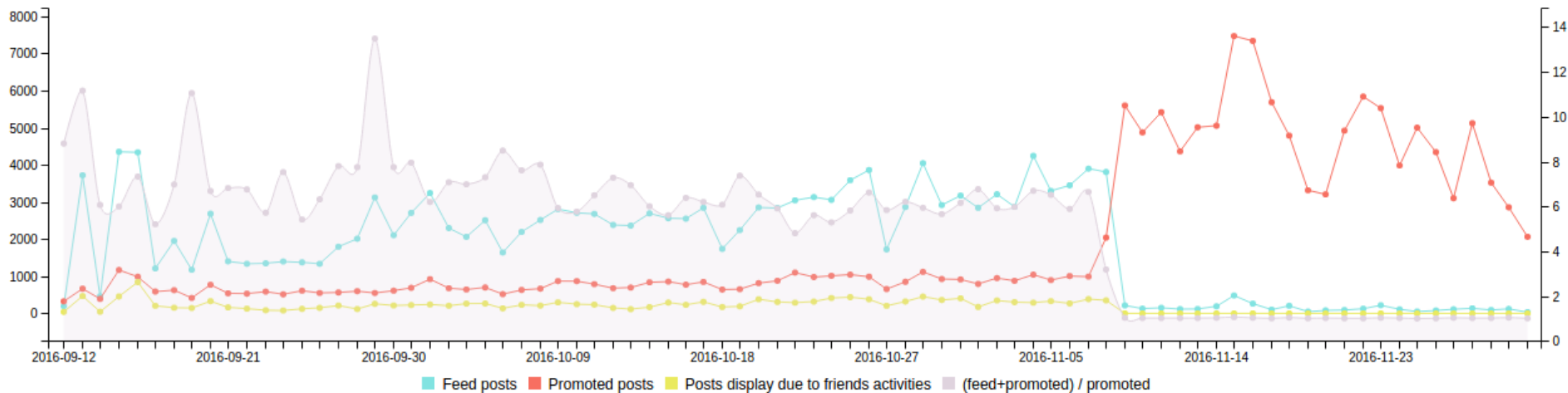


[redacted] of course

Me gusta

[how the parsing works] 1/2

Facebook use different HTML structures at the same time, and changes by surprise: this is why the parsing can't happen client side



[how the parsing works] 1/2

Small self-contained parser runs distributed and populate the metadata structure:

```
function getFeedUTime(snippet) {  
  
    var $ = cheerio.load(snippet.html);  
    var retO = { 'feedUTime': false };  
  
    try {  
        var varstr = $(".timestampContent")[0].parent.attribs['data-utime'];  
        var rv = _.parseInt(varstr);  
    } catch(error) {  
        debug("Error in extracting Publication UTime %s", snippet.id);  
        return retO;  
    }  
    if(_.isNaN(rv)) {  
        debug("Parsing failure %s ", snippet.id);  
    } else {  
        debug("Extracted %s %s ", moment(rv * 1000), snippet.id);  
        retO.feedUTime = true;  
        retO.publicationUTime = rv;  
    }  
    return retO;  
};
```

Algorithm Hegemony

**Algorithms are social policy,
therefore they should be subject to
public scrutiny**



Building Global Community



MARK ZUCKERBERG · THURSDAY, FEBRUARY 16, 2017



Kurt Gessler [Follow](#)

Deputy Editor for Digital News @ChicagoTribune. I also teach stuff at @MedillSchool. Practicing journalism ...
Apr 18 · 7 min read

Facebook's algorithm isn't surfacing one-third of our posts. And it's getting worse

Facebook is changing its News Feed to favor websites that load faster



Caroline Cakebread

🕒 18 uur 💧 11

Algorithms, *Imagine...*

Social media stores data, and return them in standard format, unaltered.

Users run their own algorithm client side,

Users exchange, customize, sell their own algorithms.



Photographer: Iain Macmillan / © Yoko Ono

Facebook as reality mirror

**Every user just see a personalized version of the reality;
Nobody, can get a true idea of what's going on.
Exception made for Facebook**

A supporter is an observation point among 1.8 billions*

**** Never forget: they are not neutral***

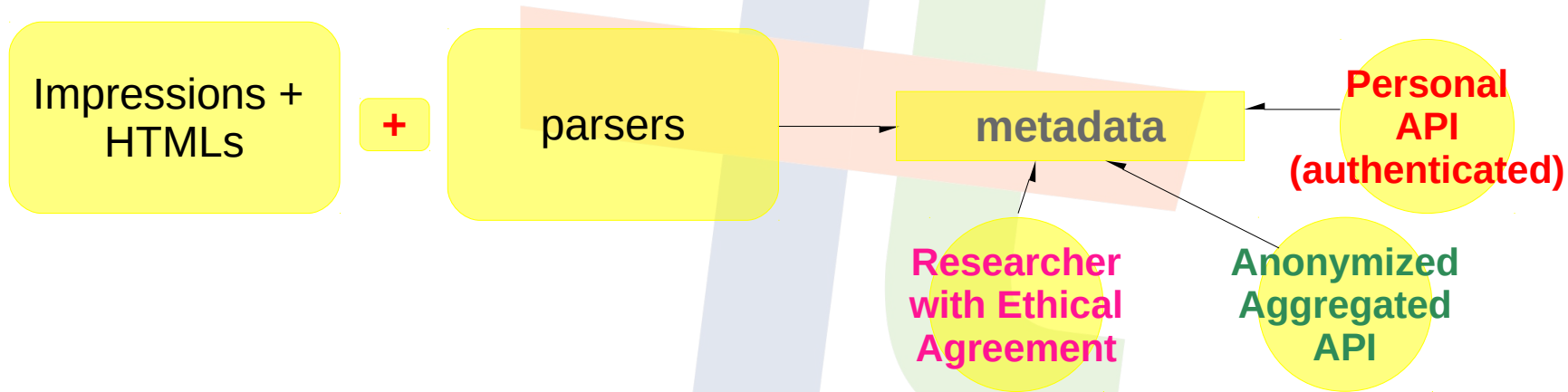
Facebook as reality mirror, the data collected

Let's aim to the exactly opposite of Facebook on:

**Protect individuals personal data,
Listening users needs,
Data transparency policy,
Give back agenda in user's hands.**

Facebook as reality mirror, Data reuse

“they are public” it is not enough to make them acritically open: data collected could be privacy sensitive



DuckDuckGo [Follow](#)

DuckDuckGo is the search engine that doesn't track you.

Jul 5 · 2 min read

Privacy Mythbusting #3: Anonymized data is safe, right? (Er, no.)

**WE ARE STILL IN BETA
FEW SAFEGUARDS ARE IN PLACE**

Facebook as reality mirror, Data reuse



Article

On the Anonymity Risk of Time-Varying User Profiles

Silvia Puglisi *, David Rebollo-Monedero and Jordi Forné

"We use actual Facebook data to show how our model can be applied to a real-world scenario." April '17

Next steps

Recruitment of: users, writers, analyst, developers

Explore the diversity of needs: Facebook graph is an unexplored resource, Facebook keep it close as it is the main asset. We look for experiment and data reuse

*but we are not Facebook,
experiments should be interesting as
part of the information ecology, not ROI based*

Merge the improvements in the main software

Practical idea?

Helps researcher in figure out way to use it

Run an experiment with some friends, and write for <https://medium.com/@trackingexposed> in your native language


Hint: run your bots with an auto-scroller

Write new parsers, (we didn't yet extract the number of like and the number of comments!)

Use the APIs (ask for new one)

Configure the extension with a control group

Imagine and experiment how to visualize the data



Tomorrow day5 - 10:50 AM
Caffè @ItalianHackerEmbassy
a dive into #fbtrex dataset