The Morphology of the Web is

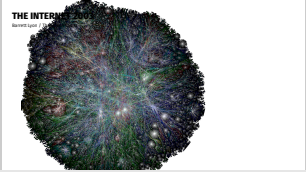
Martin Lopatka
mlopatka@mozilla.com
Data Scientist/Applied Statistician

- Research Engineering and Data Science at Mozilla!
- The world's largest shared public resource... The Web.
- you may have heard of it.. it's called Firefox.
- "the study of form and structure without consideration of function"
- carrying out research on the nature, structure, and technology on the Modern web!

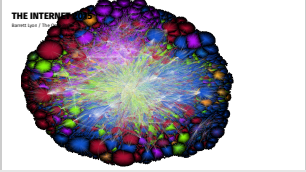
Disclaimer

Martin Lopatka provides this contribution to the Data Natives 2018 conference in a personal capacity. The views expressed are his own and do not necessarily represent the views of Mozilla Corporation or the Mozilla Foundation.

Disclaimer! GO fast!



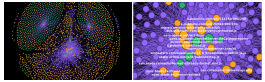
- Paint a picture of the web's morphology in 2003
- Wikipedia young < 100K articles vs 6 million
- Reddit did not exist yet, Google had just released Page-rank (household name).
- Facebook won't launch until next year.
- Op-tee project visualisation made by a massive crawl fo the web,
- 30 million TLDs registered.
- links -> content creators. outbound links to other content. less log in. No SEO.



- Emphasise differences in Morphology
- Web includes over 314 Million (geo diversity)
- more links generated algorithmically (2016 G will punish this)
- online advertising market 125 billion dollars.
- emergence of more hyperlinked nodes

The centrality of advertisers

Are trackers the new backbone of the Web?
https://medium.com/firefox-context-graph/are-trackers-the-new-backbone-of-the-web-f1b80435da15



- investigation of these hyperlinked nodes to explore their nature
- we begin exploring the role of highly connected, hub nodes
- resources accessed in a third-party context. Community detection was performed using Louvain Modularity
- fewer hyperlinks between pages belonging to separate tracking/advertising communities
- This is the first indication of a advertisement

How did we get to an ecosystem of

- We’ve already made some scary discoveries about the diversity
- ”Comme ci comme ca” using the May 2018 batch of common crawl
- Easylist ruleset ->f Adblock adblocker
- Morphology is changing, important to discuss traffic and therefore content access

Social platforms were designed to facilitate; they became attention brokers, they are designed to help companies reach users as these sort of middlemen where they know a ton about you and thats the service they are providing to the advertisers.

Renee DiResta "The Internet's Original Sin"
27-Oct-2018 12:45

- Mozfest, Renee DiResta: abuses of social media platforms
- platform purpose built for advertisement.
- *infrastructure for advertisement* content and social platforms have evolved to ”maximise sustained engagement on site”
- Example: you consume your news, *on* facebook (60% US)

The highest trafficked pages on the Web are search engines, Social media platforms, and Commerce platforms

- highest trafficked pages on the web -> business model where engaged time on site is directly beneficial in a a business sense.
- Web commerce platforms, advertisers, search engines, and Social media platforms.

Dynamic pages feature changing content, showing different text, images and videos depending on who's visiting and when. This dynamic content includes different hyperlinks.

The new ecology of outbound links

- Retargeting, advertiser to tag you (with a cookie) and redirect as a specific media clicker, this requires a link through an advertiser's intermediate page
- Advertisement content, usually dynamically generated for the individual or profile of the page visitor
- Real Time bidding, the links surfaced may vary from one page visitor to another, based on estimates of *your* profile and value to advertisers, meaning crawlers will be served a fundamentally different

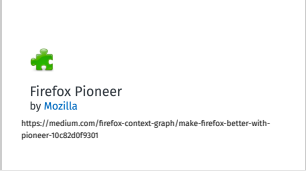
The problem is that the web is no longer built upon the simple premise of a collection of small static HTML and image files served up with a simple tag structure and readily parsed with a few lines of code. Today's web is richly dynamic, multimedia and increasingly broken into walled gardens and device-specific parallel webs.

Kalev Leetaru "Are Web Archives Failing The Modern Web: Video, Social Media, Dynamic Pages and The Mobile We"
24-Feb-2017 22:41

- from Kalev Letaru in a 2017 Forbes article discussing the role of the internet archive and common crawl initiatives in helping us make sense of the Web!

- leverage crawler technology for it’s robust technical measurement features.
- also want to user traffic patterns to ensure our feature development is relevant
- Crawlers, have a limited View - the open web makes up less and less of the Web.
- how do we study the nature of the web as people actually experience it in a Privacy respectful way that also takes advantage of scalable technology like Web crawlers?

Make Firefox better with pioneer



- We ask permission,
- compliment the shortcomings of various data collection strategies
- The Firefox Pioneer program is an opt-in data collection
- real human readable consent policy

https://github.com/mozilla/overscripted/

- selective Javascript execution stack for tracking related activity
- 70Gb and touches over 2 million web pages seeded from the Alexa top 10K
- aspiring or established data scientist, come get your hands dirty
- data science can also work in a collaborative open source manner

Acknowledgements

- my team
- interns
- ops, secEng
- community

- These slides are available on my GitHub