Web communities network mapping with Hyphe
Benjamin Ooghe-Tabanou

To cite this version:
Benjamin Ooghe-Tabanou. Web communities network mapping with Hyphe. NodeXL Academy: Social Media Research Winter School 2022, Social Media Research Foundation, Jan 2022, online, France. hal-03904193

HAL Id: hal-03904193
https://sciencespo.hal.science/hal-03904193
Submitted on 16 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License
Web communities network mapping with Hyphe

NodeXL Academy '22
January 19th 2022

Benjamin Ooghe-Tabanou (@boogheta)
Sciences Po médialab (@medialab_ScPo)
médialab @ Sciences Po

- Pluridisciplinary Research Lab created by Bruno Latour in May 2009, led by Dominique Cardon since 2017
- Social Sciences, Engineering & Design
- Articulate qualitative & quantitative methods through a digital approach
- Work with digital traces
- METAT: a monthly Open Support Workshop https://www.sciencespo.fr/recherche/fr/content/metat-latelier-de-methodes

https://medialab.sciencespo.fr
Bruno Latour, médialab founder

« Google is nice, but we need something better »

The Indian Express, 2011
Webcrawlers? What is that?

« *Web crawling is the process of building a collection of webpages by starting with an initial set of URLs and recursively traversing the corresponding pages to find additional links.* »

Gabe Ignatow & Rada Mihalcea, in: Text Mining: A Guidebook for the Social Sciences

But why would anyone want to do that?

- create an index for a search engine (like... you know...)
- extract contents or structured data (like scraping)
- perform some text mining
- study network communities through hyperlinks
Crawling the web for social sciences?

« Hyperlink » is at the core of the Web's architecture → charged with meaning and structure

« The texts are linked together in a way that one can go from one concept to another to find the information one wants. The network of links is called a web. [...] The texts are known as nodes. The process of proceeding from node to node is called navigation. »


« A hyperlink is a manifestation of intention. By linking one page to another, one piece of text to another, people intend to do particular things. »

Ryfe, Mensing, & Kelley, 2016, What is the meaning of a news link?
The Web: a place of controversies and debates

ARE YOU COMING TO BED?

I CAN'T. THIS IS IMPORTANT.

WHAT?

SOMEONE IS WRONG ON THE INTERNET.

CC-BY-NC - Randall Munroe - XKCD
A bottom-up hierarchy emerged from hyperlinks

« Matthew effect » : preferential attachment

→ new web pages tend to cite the already most cited ones
Hyphe: a research directed crawler

Build your own web documents corpus to study social phenomena online

→ gather « web actors »

→ explore hyperlinks between them

http://hyphe.medialab.sciences-po.fr/demo/

COP 21
Vie privée
Extrême droite
Tissu associatif
Produits laitiers
Cellules souches
Administrations culturelles

http://medialab.github.io/double-dating-data/

http://utopies-concretes.org/
There is no such thing as a « website »!

→ « WebEntities »: bundles of webpages aggregating coherent actors to answer a specific research question

= set of URL prefixes
Finely delimit the web territories of actors

Manually setup prefix patterns to adjust the cursor of « WebEntities »
Which data structure to manage hypertexts?

- **Tree** of urls → Hyphe's Traph
- **Graph** of hyperlinks
- Dynamic branches aggregates

https://medialab.github.io/hyphe-traph/fosdem2018/#/

Carla studies how immigration is discussed in the public debate.
**Hyphe's crawling strategy: leverage hyperlinks**

- Classical crawlers (**DMI's IssueCrawler**): Snowball
  → Top Layer attraction (Google, YouTube, Wikipedia...)
  → Topic drifts

- Hyphe:
  - crawl exclusively pages within the chosen WebEntities
  - sort discovered entities by degree of citation
  - humanly select new entities to include and crawl
**Web prospection loop: curating a corpus iteratively**

- Step by step iterative expansion & curation of entities

  - **Selection of Web Entities**
  - **Crawling**
  - **Url Candidates**

- **Human/Time cost**
- **How to know when to stop?**
  - → hyperlink citations threshold
Qualify the corpus entities (tagging)

- Free notes
- Categories
HyBro: a web browser designed for corpus curation

https://github.com/medialab/hyphe-browser/releases/

« NaviCrawler » heritage: build a Hyphe web corpus while browsing

→ « in-situ » prospection & tagging (digital field work)
→ teach the web to students
Many ways to use Hyphe

• Complete methodology includes:
  
sourcing, automatized collection, iterative corpus building
  qualitative categorization, exploratory analysis,
  network visualization, quantitative statistical analysis

• Diverse audiences:
  – Research: help social scientists work on digital fields
  – Pedagogy: teach students what the web is beyond Google & Facebook

• Possible small & large scale analyses:
  – a website's internal structure
  – a theme's ensemble of actors and their ties
  – a controversy's alliances & oppositions
  – etc.
From above: clusters, opposition & affinity

Web corpus on climate change: type of supported policy

The network represents how actors hyperlink on the web. The most cited actors have the biggest representation. Links are colored according to their source.
From above: clusters, opposition & affinity

Double Dating Data: Climate change on the web

The network represents how actors hyperlink on the web. The most cited actors have the biggest representation. Links are colored according to their source.

Web corpus on climate change: anthropogenic climate change

Le réseau représente les acteurs tels qu’ils sont reliés sur le web par des hyperliens. Les acteurs les plus cités dans le corpus ont une plus grande taille. La couleur des liens est celle de l’acteur qui en est à l’origine.

Corpus web sur le changement climatique: le changement climatique est-il anthropique?
From within: explore webpages contents

Privacy Web Corpus Datascape

NodeXL Academy '22 - Jan 19th 2022  Web communities network mapping with Hyphe
From the sides: a hierarchy of directed hyperlinks

Conspiracy theorists
- Not reliable
- Not cited
- Cite many reliable sources

Right and extreme right
- Not reliable
- Cited only by conspirationists
- Cite many reliable sources

French national media
- Reliable *(in this qualification)*
- Cited by everyone
- Cite only platforms and regional media

Visual Network Exploration for Data Journalists
Hyperlinks directionality: a bottom-up hierarchy

Most hyperlinks stem from the unreliable and aim at the reliable resources.

Each bloc’s surface is proportional to the count of websites. The color code is the same as the “Décodex”.

<table>
<thead>
<tr>
<th>Category</th>
<th>Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliable</td>
<td>46% 305 sites</td>
</tr>
<tr>
<td>Satirical</td>
<td>11% 76 sites</td>
</tr>
<tr>
<td>Imprecise</td>
<td>13% 89 sites</td>
</tr>
<tr>
<td>Unreliable</td>
<td>30% 197 sites</td>
</tr>
<tr>
<td>Right + extreme-right</td>
<td>9% 59 sites</td>
</tr>
<tr>
<td>Conspiracy theorist</td>
<td>11% 73 sites</td>
</tr>
<tr>
<td>Other</td>
<td>10% 65 sites</td>
</tr>
</tbody>
</table>

The thickness is proportional to the normalized link density.*

*Accounts for the different group sizes.

Densities below 2% have been omitted.

Visual Network Exploration for Data Journalists

NodeXL Academy '22 - Jan 19th 2022

Web communities network mapping with Hyphe
Explore polarization dynamics

Unfolding the multi-layered structure of the French Mediascape
Bibliography

Reference publications:

  
  https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679

  
  https://spire.sciencespo.fr/hdl:/2441/6obemb2hsj9pboj9bbvc7sftne

– Plique G., Jacomy M., Ooghe-Tabanou B., Girard P. (2018), **It's a Tree... It's a Graph... It's a Traph! Designing an on-file multi-level graph index for the Hyphe web crawler**, FOSDEM 2018, Bruxelles
  
  https://medialab.github.io/hyphe-traph/fosdem2018/#/

  
Now let’s play!

https://hyphe.medialab.sciences-po.fr/demo/
https://github.com/medialab/hyphe

Any question first?

benjamin.ooghe@sciencespo.fr
@boogheta  @medialab_ScPo