UNIVERSITÀ DEGLI STUDI DI TRENTO

Department Of Information Engineering And Computer Science
ICT International Doctoral School

CYCLE XXX

The *Dao* of Wikipedia

Extracting Knowledge from the Structure of Wikilinks

CRISTIAN CONSONNI

Advisor:

Alberto Montresor
University of Trento, Trento

Co-advisor:

Yannis Velegrakis
University of Trento, Trento

2019

A Virginia, per essermi stata vicina.

ABSTRACT

Wikipedia is a multilingual encyclopedia written collaboratively by volunteers online, and it is now the largest, most visited encyclopedia in existence. Wikipedia has arisen through the self-organized collaboration of contributors, and since its launch in January 2001, its potential as a research resource has become apparent to scientists, its appeal lying in the fact that it strikes a middle ground between accurate, manually created, limited-coverage resources, and noisy knowledge mined from the web. For this reason, Wikipedia's content has been exploited for a variety of applications: to build knowledge bases, to study interactions between users on the Internet, and to investigate social and cultural issues such as gender bias in history, or the spreading of information.

Similarly to what happened for the Web at large, a structure has emerged from the collaborative creation of Wikipedia: its articles contain hundreds of millions of links. In Wikipedia parlance, these internal links are called *wikilinks*. These connections explain the topics being covered in articles and provide a way to navigate between different subjects, contextualizing the information, and making additional information available.

In this thesis, we argue that the information contained in the link structure of Wikipedia can be harnessed to gain useful insights by extracting it with dedicated algorithms. More prosaically, in this thesis, we explore the link structure of Wikipedia with new methods.

In the first part, we discuss in depth the characteristics of Wikipedia, and we describe the process and challenges we have faced to extract the network of links. Since Wikipedia is available in several language editions and its entire edition history is publicly available, we have extracted the wikilink network at various points in time, and we have performed data integration to improve its quality.

In the second part, we show that the wikilink network can be effectively used to find the most relevant pages related to an article provided by the user. We introduce a novel algorithm, called *CycleRank*, that takes advantage of the link structure of Wikipedia considering cycles of links, thus giving weight to both incoming and outgoing connections, to produce a ranking of articles with respect to an article chosen by the user.

In the last part, we explore applications of *CycleRank*. First, we describe the ENGINEROOM EU project, where we faced the challenge to

find which were the most relevant Wikipedia pages connected to the Wikipedia article about the *Internet*. Finally, we present another contribution using Wikipedia article accesses to estimate how the information about diseases propagates.

In conclusion, with this thesis, we wanted to show that browsing Wikipedia's wikilinks is not only fascinating and serendipitous[1], but it is an effective way to extract useful information that is latent in the user-generated encyclopedia.

1 https://xkcd.com/214/

PUBLICATIONS

This thesis is based on the following papers:

[1] Cristian Consonni, David Laniado, and Alberto Montresor. Wikilinkgraphs: A complete, longitudinal and multi-language dataset of the wikipedia link networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 598–607, 2019.

[2] Cristian Consonni, David Laniado, and Alberto Montresor. Discovering Topical Contexts from Links in Wikipedia. 2019.

[3] Cristian Consonni, David Laniado, and Alberto Montresor. CycleRank, or There and Back Again: personalized relevance scores from cyclic paths on graphs. *Submitted to VLDB 2020*, 2020.

[4] Paolo Bosetti, Piero Poletti, Cristian Consonni, Bruno Lepri, David Lazer, Stefano Merler, and Alessandro Vespignani. Disentangling social contagion and media drivers in the emergence of health threats awareness. *Science Advances*, 2019. *Under review at Science Advances.*

This Ph.D. was instrumental to study other topics, which I chose not to include in this manuscript:

[5] Cristian Consonni, Paolo Sottovia, Alberto Montresor, and Yannis Velegrakis. Discovering Order Dependencies through Order Compatibility. In *International Conference on Extending Database Technology*, 2019.

[6] Riccardo Pasi, Cristian Consonni, and Maurizio Napolitano. Open Community Data & Official Public Data in flood risk management: a comparison based on InaSAFE. In *FOSS4G-Europe 2015, the 2nd European Conference for for Free and Open Source Software for Geospatial*, 2015.

[7] Marco Cè, Cristian Consonni, Georg P. Engel, and Leonardo Giusti. Non-Gaussianities in the topological charge distribution of the SU(3) Yang-Mills theory. *Physical Review D*, 92(7):074502, 2015.

CONTENTS

1

INTRODUCTION

At a first look, the *brain*, a *knowledge base*, and the *Garden of Eden* do not seem to have anything in common. However, it can be argued that in all these metaphorical places, knowledge is encoded in the structure of a graph. A graph is a structure composed by a set of objects in which some pairs of objects possess some given property. The objects correspond to abstractions called nodes, vertices or points; and each of the related pairs of vertices is called an edge, arc, or line.

For the brain, the concept of *neural network* is well-known since the late XIXth century, and it is used as a practical tool in computer science since the 1980's [5]. In this model, individual *neurons* are the nodes of the graph, and the *synapses* are the edges. In this context, the ability of the brain of modifying the connections between neurons, called *neuroplasticity*, offers the insight that the structure of the connections in a graph are fundamental for the encoding of knowledge in the graph structure.

A knowledge base is a technology used to store information. Following the Resource Description Framework (RDF) paradigm, a knowledge base is a collection of statements of the form *subject—predicate—object*, also known as triples. Nodes are resources in the knowledge base - either subjects or objects - while edges encode the predicates.

Finally, in the Garden of Eden, the idea is literally present in the form of *Tree of the knowledge of good and evil*, besides the fascination of the fact that a *tree* is, in fact, a special and simple type of graph, more profoundly the Tree can be described as an *axis mundi*, is that is the point of connection between the divine and the mortals.

The idea that knowledge is contained or encoded in the relations among entities, or in *paths* connecting nodes, is very ancient as well. In the Chinese tradition of Taoism, the *Tao* or *Dao* - literally the "way", "path", "route", or "road" - encodes the natural order of the universe whose character one0s human intuition must discern in order to realize the potential for individual wisdom. This intuitive knowing of life cannot be grasped as a concept; it is known through actual living experience of one's everyday being. In Buddhism, the *Noble Eightfold Path* is a of Buddhist practices leading to *nirvana* and the liberation from from suffering and ignorance.

In this thesis, we start from the grand idea that paths in graphs encode some knowledge about the entities they connect and we present an algorithm that we have devised to highlight these emergent truths. In particular, we will use Wikipedia, the collaborative, web-based, free encyclopedia as a general network of concept and we will show that it is possible to extract new knowledge from this graph using dedicated algorithms.

In the following, we will briefly introduce the main subjects of our investigation namely: graphs and Wikipedia. We will also focus on the Pagerank algorithm [6] as a prime example of an algorithm that can extract knowledge, in particular in the form of scores, from the paths in a graph.

Graphs are fundamental structures that can capture many real-world phenomena. Graphs, also called *networks*, offer the foundation for modeling a variety of situations in diverse domains such as describing relations among individuals in social networks, organizational networks, semantic relations among concepts in knowledge bases, food webs and many others. The opportunity to investigate these domains is related to the availability of data.

Several trends in the last decade have contributed new sources of data in digital form: Web 2.0 and user-generated content, social media and, more recently, *Big Data* and the *Internet of Things* (IoT). Data generated by users - e.g. in Wikipedia and in online social networks - are usually augmented by the availability of metadata that are created completely automatically by sensors or without user interaction, such as the stream of the web pages visited by a user. These data present challenges related to their volume, the size of the datasets; velocity, the frequency of update; and variety, the diversity of their sources and scope. This phenomenon has been called the *data deluge* [7, 8]. To respond to this new context, computer scientists have developed new tools specifically designed to manage these new datasets.

Heterogeneous information networks are ubiquitous and form a critical component of modern information infrastructure. Despite their prevalence in our world, researchers have only recently recognized the importance of studying information networks as a whole. Hidden in these networks are the answers to important questions. For example, is there a collaborated plot behind a network intrusion, and how can a source in communication networks be identified? How can a company derive a complete view of its products at the retail level from interlinked social communities? These questions are highly relevant to a new class of analytical applications that query and mine massive information networks for pattern and knowledge discovery, data and information integration, veracity analysis and deep understanding of the principles of information networks.

From the beginning of the years 2000's graphs have been extensively employed to tackle new problems and explore new opportunities that require the ability to process massive graphs. In this context many modern applications use graphs as a data structure to provide services such as suggesting friends on social networks, answer queries on knowledge bases or modeling biological phenomena such as gene co-activations. Since they describe real-world phenomena, these systems and the graphs that model them can change over time.

Searching for information and knowledge inside networks, particularly large networks with thousands of nodes is a complex and time-consuming task. Unfortunately, the lack of a general analytical and access platform makes sensible navigation and human comprehension virtually impossible in large-scale networks. Fortunately, information networks contains massive nodes and links associated with various kinds of information. Knowledge about such networks is often hidden in massive links in heterogeneous information networks but can be uncovered by the development of sophisticated knowledge discovery mechanisms.

Part I

GRAPHS FROM WIKIPEDIA

Wikipedia articles contain multiple links connecting a subject to other pages of the encyclopedia. In Wikipedia parlance, these links are called internal links or *wikilinks*. We present a complete dataset of the network of internal Wikipedia links for the 9 largest language editions. The dataset contains yearly snapshots of the network and spans 17 years, from the creation of Wikipedia in 2001 to March 1st, 2018. While previous work has mostly focused on the complete hyperlink graph which includes also links automatically generated by templates, we parsed each revision of each article to track links appearing in the main text. In this way we obtained a cleaner network, discarding more than half of the links and representing all and only the links intentionally added by editors. We describe in detail how the Wikipedia dumps have been processed and the challenges we have encountered, including the need to handle special pages such as *redirects*, i.e., alternative article titles. We present descriptive statistics of several snapshots of this network. Finally, we propose several research opportunities that can be explored using this new dataset.

2

WIKILINKGRAPHS: A COMPLETE, LONGITUDINAL AND MULTI-LANGUAGE DATASET OF THE WIKIPEDIA LINK NETWORKS

Wikipedia[1] is probably the largest existing information repository, built by thousands of volunteers who edit its articles from all around the globe. As of March 2019, it is the fifth most visited website in the world [9]. Almost 300k active users per month contribute to the project [10], and more than 2.5 billion edits have been made. The English version alone has more than 5.7 million articles and 46 million pages and is edited on average by more than 128k active users every month [11]. Wikipedia is usually a top search-result from search engines [12] and research has shown that it is a first-stop source for information of all kinds, including information about science [13, 14], and medicine [15].

The value of Wikipedia does not only reside in its articles as separated pieces of knowledge, but also in the links between them, which represent connections between concepts and result in a huge conceptual network. According to Wikipedia policies[2] [16], when a concept is relevant within an article, the article should include a link to the page corresponding to such concept [17]. Therefore, the network between articles may be seen as a giant mind map, emerging from the links established by the community. Such graph is not static but is continuously growing and evolving, reflecting the endless collaborative process behind it.

The English Wikipedia includes over 163 million connections between its articles. This huge graph has been exploited for many purposes, from natural language processing [18] to artificial intelligence [19], from Semantic Web technologies and knowledge bases [20] to complex networks [21], from controversy mapping [22] to human way-finding in information networks [23].

1 https://www.wikipedia.org

2 In what follows, we will refer to the policies in force on the English-language edition of Wikipedia; we will point out differences with local policies whenever they are relevant.

This paper presents a new dataset, WikiLinkGraphs, that makes the networks of internal links in the nine largest editions of Wikipedia available to researchers and editors, opening new opportunities for research.

Most previous work on the Wikipedia link graph relies on wikilink data made accessible through the Wikipedia API[3] and through database dumps[4] These data include also all transcluded links, i.e. links automatically generated by templates defined in another page; templates typically add all possible links within a given group of articles, producing big cliques and inflating the density of connections.

Inserting a template in Wikipedia merely amounts to writing a small snippet of code, which in the final article is rendered as a collection of links. Figure 1 shows a rendering of the navigation template `{{Computer science}}`[5] from English Wikipedia, which produces a table with 146 links to other articles within the encyclopedia. Navigation templates are very general by design serve to group links to multiple related articles. They are not specific to a given page: in fact, the content of a template can be changed independently from editing the pages where it is included.

We argue that considering only the links explicitly added by editors in the text of the articles may provide a more trustful representation of semantic relations between concepts, and result in a cleaner graph by avoiding the cliques and other potential anomalous patterns generated by transcluded links.

The aim of this work is to build a dataset of the graph of the specific link between Wikipedia articles added by the editors. The WikiLinkGraphs dataset was created by parsing each article to extract its links, leaving only the links intentionally added by editors; in this way, we discarded over half of the overall links appearing in the rendered version of the Wikipedia page.

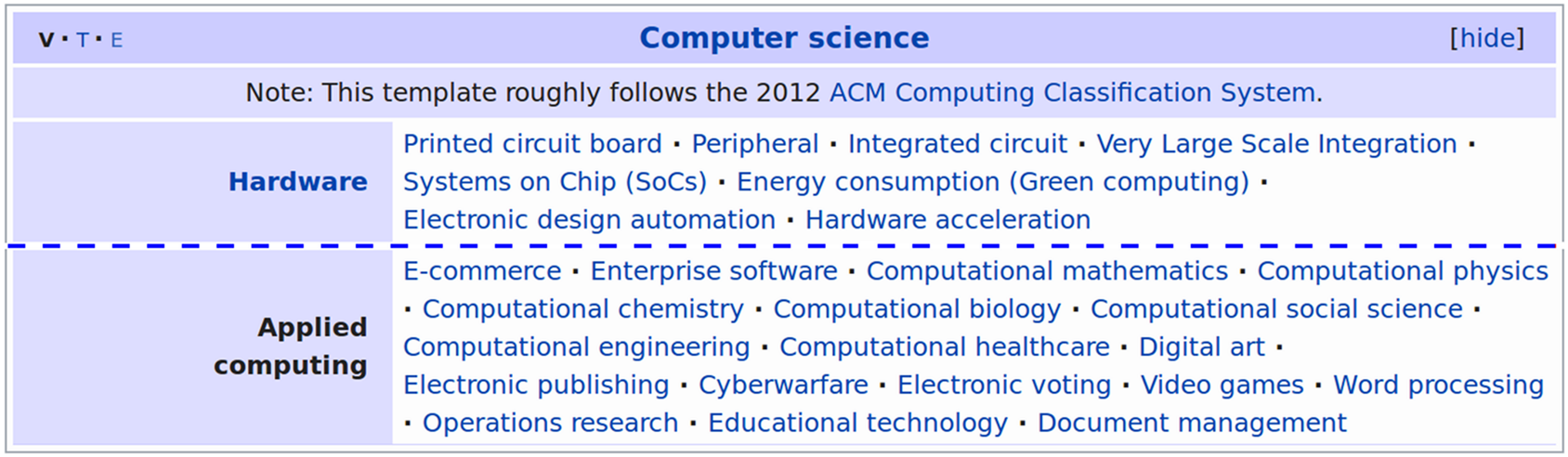
Furthermore, we tracked the complete history of each article and of each link within it, and generated a dynamic graph representing the evolution of the network. Whilst the dataset we are presenting in this paper consists of yearly snapshots, we have generated several supporting dataset as well, such as a large collection tracking the timestamp in which each occurrence of a link was created or removed.

Redirects, i.e. special pages representing an alternative title for an article, are a known issue that was shown to affect previous research [24]. In our dataset, we tracked the redirects over time, and resolved all of them according to the corresponding timestamp. The complete history of all redirects is made available as well.

3 Hyperlinks in the current version of Wikipedia are available through the "Link" property in the Wikipedia API: https://www.mediawiki.org/wiki/API

4 https://meta.wikimedia.org/wiki/Data_dumps

5 https://en.wikipedia.org/wiki/Template:Computer_science



v · T · E	**Computer science**	[hide]
Note: This template roughly follows the 2012 ACM Computing Classification System.		
Hardware	Printed circuit board · Peripheral · Integrated circuit · Very Large Scale Integration · Systems on Chip (SoCs) · Energy consumption (Green computing) · Electronic design automation · Hardware acceleration	
Applied computing	E-commerce · Enterprise software · Computational mathematics · Computational physics · Computational chemistry · Computational biology · Computational social science · Computational engineering · Computational healthcare · Digital art · Electronic publishing · Cyberwarfare · Electronic voting · Video games · Word processing · Operations research · Educational technology · Document management	

Figure 1: A portion of the navigational template `{{Computer science}}` from English Wikipedia as of revision n^{o} 878025472 of 12 January 2019, 14:12. The dashed line indicates that a portion of template has been stripped for reasons of space.

The code used to generate the dataset is also entirely made available on GitHub, so that anybody can replicate the process and compute the wikilink graphs for other language editions and for future versions of Wikipedia.

2.1 THE WIKILINKGRAPHS DATASET

This section describes how we processed the Wikipedia dumps of the complete edit history to obtain the dataset.

2.1.1 DATA PROCESSING

The WikiLinkGraphs dataset was created from the full Wikipedia revision history data dumps of March 1, 2018[6], as published by the Wikimedia Foundation, and hence includes all entire months from January 2001 to February 2018.

These XML dumps contain the full content of each Wikipedia page for a given language edition, including encyclopedia articles, talk pages and help pages. Pages are divided in different *namespaces*, that can be recognized by the prefix appearing in the title of the page. The encyclopedia articles are in the *main namespace*, also called `namespace 0` or `ns0`. The content of the pages in Wikipedia is formatted with *Wikitext* [25], a simplified syntax that is then rendered as HTML by the MediaWiki software[7]. For each edit a new *revision* is created: the dump contains all revisions for all pages that were not deleted.

Table 1 presents the compressed sizes for the XML dumps that have been downloaded and the number of pages and revisions that have been processed. We extracted all the article pages. This resulted in 40M articles being analyzed. In total, more than 1B revisions have been processed to produce the WikiLinkGraphs dataset.

2.1.1.1 *Link Extraction*

Wikipedia articles have *revisions*, which represent versions of the Wikitext of the article at a specific time. Each modification of the page

6 All files under "All pages with complete edit history (.7z)" at https://dumps.wikimedia.org/enwiki/20180301/. Wikipedia dumps are available up to 3 months prior to the current date, so those specific dumps are not available anymore. However, any dump contains the whole Wikipedia history dating from 2001 onwards. So our results can be replicated with any dump taken later than March 1st, 2018.

7 https://www.mediawiki.org

lang	size (GB)	files	pages	revisions
de	33.0	109	3,601,030	113,836,228
en	138.0	520	13,750,758	543,746,894
es	27.0	68	3,064,393	77,498,219
fr	26.0	95	3,445,121	99,434,840
it†	91.0	61	2,141,524	68,567,721
nl	7.4	34	2,627,328	38,226,053
pl	15.0	34	1,685,796	38,906,341
ru	24.0	56	3,362,946	63,974,775
sv	9.0	1	6,139,194	35,035,976

Table 1: Statistics about the processed Wikipedia dumps: size of the dowloaded files and number of processed pages and revisions for each dump. (†) the Italian Wikipedia dumps were downloaded in `.bz2` format.

(an *edit* in Wikipedia parlance) generates a new revision. Edits can be made by *anonymous* or *registered* users.

A revision contains the wikitext of the article, which can have sections, i.e. header titles. Sections are internally numbered by the MediaWiki software from 0, the *incipit* section, onwards. As for HTML headers, several section levels are available (sections, subsections, etc.); section numbering does not distinguish between the different levels.

While a new visual, WYSIWYG editor has been made available in most Wikipedia editions starting since June 2013 [26], the text of Wikipedia pages is saved as *Wikitext*. In this simplified markup language, internal Wikipedia links have the following format `[[title|anchor]]`; for example,

```
[[New York City|The Big Apple]]
```

This wikitext is visualized as the words `The Big Apple` that gets translated into HTML as:

```
<a href="/wiki/New_York_City"
   title="New York City">The Big Apple</a>
```

pointing to the Wikipedia article *New York City*. If the page exists, as in this example, the link will be blue-colored, otherwise it will be colored in red, indicating that the linked-to page does not exist [27]. The anchor is optional and, if it was omitted, then the page title, in this case `New York City`, would have been visualized.

For each revision of each page in the Wikipedia dump, we used the following regular expression in Python[8] to extract *wilinks*:

```
1  \[\[
2  (?P<link>
3     [^\n\|\]\[\<\>\{\}]{0,256}
4  )
5  (?:
6    \|
7    (?P<anchor>
8        [^\[]*?
9    )
10 )?
11 \]\]
```

Line 1 matches two open brackets; then, Lines 2–4 capture the following characters in a named group called `link`. Lines 5–10 match the optional anchor: Line 5 matches a pipe character, then Lines 6–8 match non-greedily any valid character for an anchor saving them in a named group called `anchor`. Finally, Line 10 matches two closed brackets. The case of links pointing to a section of the article is handled *a posteriori*, after the regular expression has captured its contents. When linking to a section, the `link` text will contain a pound sign (`#`); given that this symbol is not allowed in page titles, we can separate the title of the linked page from the section.

The RawWikilinks Dataset.

The link extraction process produces a dataset with the following information:

- `page_id`: an integer, the page identifier used by MediaWiki. This identifier is not necessarily progressive, there may be gaps in the enumeration;
- `page_title`: a string, the title of the Wikipedia article;
- `revision_id`: an integer, the identifier of a revision of the article, also called a *permanent id*, because it can be used to link to that specific revision of a Wikipedia article;
- `revision_parent_id`: an integer, the identifier of the parent revision. In general, each revision as a unique parent; going back in time before 2002, however, we can see that the oldest articles present non-linear edit histories. This is a consequence of the import process from the software previously used to power Wikipedia, MoinMoin, to MediaWiki;

8 https://github.com/WikiLinkGraphs/wikidump/blob/70b0c7f929fa9d66a220caf11c9e31691543d73f/wikidump/extractors/misc.py#L203

- `revision_timestamp`: date and time of the edit that generated the revision under consideration;
- `user_type`: a string (`"registered"` or `"anonymous"`), specifying whether the user making the revision was logged-in or not;
- `user_username`: a string, the username of the user that made the edit that generated the revision under consideration;
- `user_id`: an integer, the identifier of the user that made the edit that generated the revision under consideration;
- `revision_minor`: a boolean flag, with value 1 if the edit that generated the current revision was marked as *minor* by the user, 0 otherwise;
- `wikilink.link`: a string, the page linked by the wikilink;
- `wikilink.tosection`: a string, the name of the section if the link points to a section;
- `wikilink.anchor`: a string, the anchor text of the *wikilink*;
- `wikilink.section_name`: the name of the section wherein the *wikilink* appears;
- `wikilink.section_level`: the level of the section wherein the *wikilink* appears;
- `wikilink.section_number`: the number of the section wherein the *wikilink* appears.

2.1.1.2 *Redirects and Link Resolution*

A redirect in MediaWiki is a page that automatically sends users to another page. For example, when clicking on a *wikilink* `[[NYC]]`, the user is taken to the article *New York City* with a note at the top of the page saying: "(Redirected from NYC)". The page *NYC*[9] contains special Wikitext: `#REDIRECT [[New York City]]`which defines it as a redirect page and indicates the target article. It is also possible to redirect to a specific section of the target page. Different language editions of Wikipedia use different words[10], which are listed in Table 2.

In general, a redirect page can point to another redirect page creating a chain of multiple redirects[11]. These pages should only be temporary

9 https://en.wikipedia.org/w/index.php?title=NYC&redirect=no

10 https://github.com/WikiLinkGraphs/wikidump/blob/70b0c7f929fa9d66a220caf11c9e31691543d73f/wikidump/extractors/redirect.py#L14

11 For example, a live list of pages creating chains of redirect on English Wikipedia is available at https://en.wikipedia.org/wiki/Special:DoubleRedirects.

lang	words
de	`#WEITERLEITUNG`
en	`#REDIRECT`
es	`#REDIRECCIÓN`, `#REDIRECCION`
fr	`#REDIRECTION`
it	`#RINVIA`, `#RINVIO`, `#RIMANDO`
nl	`#DOORVERWIJZING`
pl	`#PATRZ`, `#PRZEKIERUJ`, `#TAM`
ru[‡]	`#Perenapravlenie`, `#perenapr`
sv	`#OMDIRIGERING`

Table 2: Words creating a redirect in MediaWiki for different languages. `#REDIRECT` is valid on all languages. ([‡]) For Russian Wikipedia, we present the transliterated words.

and they are actively eliminated by Wikipedia volunteers manually and using automatic scripts.

Despite the name, redirects are served as regular pages by the MediaWiki software so requesting a redirect page, for example by visiting the link https://en.wikipedia.org/wiki/NYC, returns an HTTP status code of 200.

2.1.1.3 *Resolving Redirects*

We have extracted one snapshot per year on March, 1st from the RawWikilinks dataset. The creation of a snapshot for a given year entails the following process:

1. we list all *revisions* with their timestamps from the dumps;
2. we filter the list of revisions keeping only those that existed on March 1st, i.e. the last revision for each page created before March 1st;
3. we resolve the redirects by comparing each page with the list of redirects obtained as described above;

At the end of this process, we obtain a list of the pages that existed in Wikipedia on March, 1st of each year, together with their target, if they are redirects. We call this dataset ResolvedRedirects.

It should be noted that even if we resolve redirects, we do not eliminate the corresponding pages: in fact, redirects are still valid pages belonging to the `namespace 0` and thus they still appear in our snapshots as nodes with one outgoing link, and no incoming links.

2.1.1.4 *Link Snapshots*

We then process the RawWikilinks dataset and we are able, for each revision of each page, to establish whether a wikilink in a page was pointing to an existing page or not. We add this characteristics to the RawWikilinks dataset in the field `wikilink.is_active`: a boolean representing whether the page pointed to by the link was existing in that moment or not. Revisions are then filtered so to obtain the lists of links existing in each page at the moment of interest; we call this new dataset WikiLinkSnapshots.

2.1.1.5 *Graph Snapshots (*WikiLinkGraphs*)*

Armed with the WikiLinkSnapshots and the ResolvedRedirects dataset we can extract the WikiLinkGraphs as a list of records with the following fields:

- `page_id_from`: an integer, the identifier of the source article.
- `page_title_from`: a string, the title of the source article;
- `page_id_to`: an integer, the identifier of the target article;
- `page_title_to`: a string, the title of the target article;

If a page contains a link to the same page multiple times, this would appear as multiple rows in the WikiLinkSnapshots dataset. When transforming this data to graph format we eliminate these multiple occurrences, because we are only interested in the fact that the two pages are linked. Wikipedia policies about linking [16] state that in general a link should appear only once in an article and discourage contributors to put multiple links to the same destination. One clear example is the page *New York City* where, for example, the expression *"United States"* is used to link to the corresponding article only once, at the first occurrence. For these reasons, we do not think it is justified to assign any special meaning to the fact that two page have multiple direct connections between them.

Figure 2 summarizes the steps followed to produce the WikiLinkGraphs from the Wikipedia dumps with the intermediate datasets produced.

2.1.2 DATASET DESCRIPTION

The WikiLinkGraphs dataset comprises data from 9 Wikipedia language editions: German (`de`), English (`en`), Spanish (`es`), French (`fr`), Italian (`it`), Dutch (`nl`), Polish (`pl`), Russian (`ru`), and Swedish (`sv`).

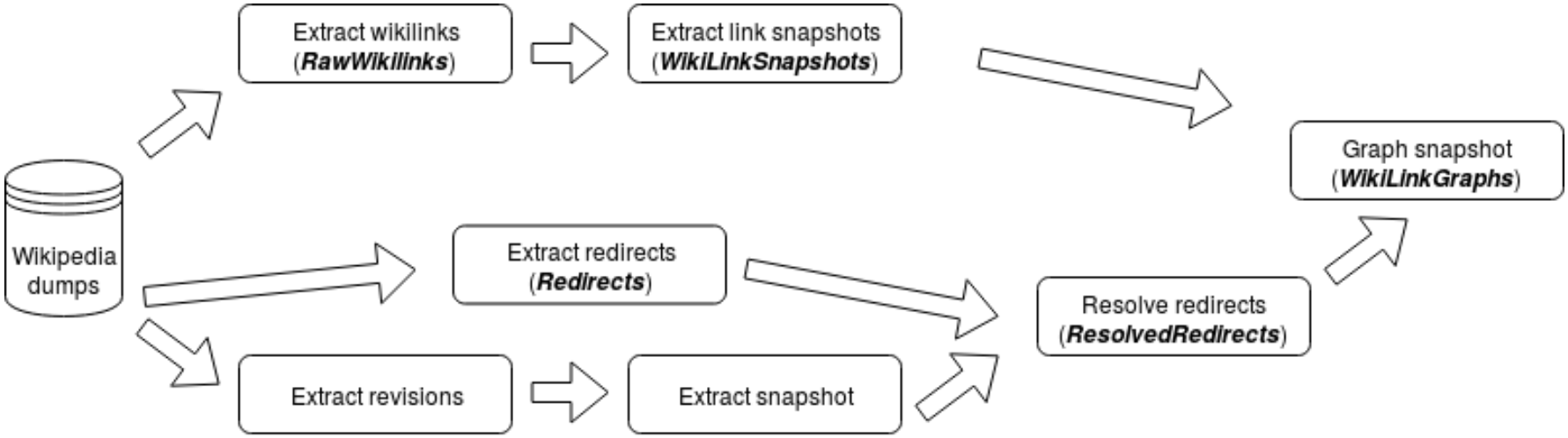


Figure 2: The process to produce the WikiLinkGraphs dataset from the Wikipedia dumps. In bold and italics the name of the intermediate datasets produced.

date	de		en		es		fr		it	
	N	**E**	**N**	**E**	**N**	**E**	**N**	**E**	**N**	**E**
2001-03-01	0	0	37	31	0	0	0	0	0	0
2002-03-01	900	1,913	27,654	223,705	1,230	2,664	55	53	0	0
2003-03-01	14,545	126,711	118,946	1,318,655	2,786	13,988	6,694	58,027	1,036	9,695
2004-03-01	63,739	794,561	248,193	3,170,614	17,075	162,219	28,798	338,108	6,466	97,721
2005-03-01	244,110	3,659,389	624,287	8,505,195	43,114	457,032	96,676	1,238,756	32,834	355,197
2006-03-01	474,553	7,785,292	1,342,642	18,847,709	112,388	1,351,111	283,831	3,926,485	149,935	1,434,869
2007-03-01	775,104	11,946,193	2,425,283	34,219,970	253,569	3,327,609	555,471	7,900,561	302,276	3,960,767
2008-03-01	1,063,222	15,598,850	3,676,126	50,270,571	452,333	6,292,452	1,113,622	12,546,302	507,465	7,239,521
2009-03-01	1,335,157	19,607,930	4,848,297	61,318,980	762,234	9,504,039	1,369,619	16,546,043	693,445	10,713,417
2010-03-01	1,603,256	23,834,140	5,937,618	71,024,045	1,159,567	12,844,652	1,632,118	21,064,666	877,089	14,120,469
2011-03-01	1,879,381	28,457,497	7,027,853	82,944,163	1,693,815	17,454,997	1,890,614	25,704,865	1,043,648	17,496,901
2012-03-01	2,163,719	33,036,436	7,922,426	93,924,479	1,944,529	21,167,388	2,137,209	30,422,158	1,213,961	21,069,750
2013-03-01	2,461,158	37,861,651	8,837,308	105,052,706	2,198,429	24,314,571	2,369,365	34,791,331	1,377,144	24,694,404
2014-03-01	2,712,984	42,153,240	9,719,211	116,317,952	2,409,026	27,090,659	2,594,282	39,257,288	1,511,827	26,821,204
2015-03-01	2,933,459	46,574,886	10,568,011	127,653,091	2,561,516	29,529,035	2,809,572	43,831,574	1,643,387	29,867,490
2016-03-01	3,155,927	50,904,750	11,453,255	139,194,105	2,728,713	32,633,513	3,037,908	48,659,900	1,802,952	32,521,188
2017-03-01	3,372,406	55,184,610	12,420,400	150,743,638	2,881,220	35,546,330	3,239,160	53,126,118	1,917,410	35,158,350
2018-03-01	3,588,883	59,535,864	13,685,337	163,380,007	3,034,113	38,348,163	3,443,206	57,823,305	2,117,022	37,814,105

Table 3: Number of nodes N and edges E for each graph snapshot of WikiLinkGraphs dataset obtained for the English (en), German (de), Spanish (es), French (fr), and Italian (it) Wikipedia editions.

date	nl		pl		ru		sv	
	N	E	**N**	**E**	**N**	**E**	**N**	**E**
2001-03-01	0	0	0	0	0	0	0	0
2002-03-01	368	728	698	1,478	0	0	122	184
2003-03-01	5,182	41,875	8,799	68,720	108	239	6,708	33,473
2004-03-01	23,059	225,429	24,356	299,583	1,600	3,927	22,218	171,486
2005-03-01	62,601	669,173	61,378	779,843	11,158	63,440	66,673	651,671
2006-03-01	169,193	1,850,260	234,506	2,218,720	64,359	422,903	163,988	1,605,526
2007-03-01	338,354	3,746,141	395,723	4,575,510	246,494	1,849,540	269,599	2,627,901
2008-03-01	523,985	6,037,117	546,236	7,151,435	459,863	3,762,487	370,569	3,746,860
2009-03-01	667,311	7,900,852	690,887	9,663,964	703,316	6,395,215	452,132	4,841,861
2010-03-01	764,277	9,467,588	822,868	11,776,724	962,680	9,881,672	542,900	5,856,848
2011-03-01	879,062	11,120,219	953,620	13,959,431	1,295,284	13,955,827	712,129	6,922,100
2012-03-01	1,358,162	14,255,313	1,091,816	15,813,952	1,562,821	17,882,908	800,776	7,945,812
2013-03-01	1,550,027	16,241,260	1,208,355	17,405,307	1,862,035	21,724,380	1,424,006	16,812,447
2014-03-01	2,332,477	19,940,218	1,322,701	19,244,972	2,098,071	25,100,193	2,422,972	26,497,619
2015-03-01	2,424,624	21,638,960	1,414,645	20,838,508	2,350,262	28,242,878	3,218,352	33,025,219
2016-03-01	2,500,880	23,252,874	1,513,239	22,445,122	2,782,155	31,467,831	4,470,345	38,864,469
2017-03-01	2,569,547	24,691,572	1,597,694	24,238,529	3,094,419	34,441,603	6,062,996	51,975,115
2018-03-01	2,626,527	25,834,057	1,684,606	25,901,789	3,360,531	37,394,229	6,131,736	52,426,633

Table 4: Number of nodes N and edges E for each graph snapshot of WikiLinkGraphs dataset obtained for the Dutch (nl), Polish (pl), Russian (ru), and Swedish (sv) Wikipedia editions.

These editions are the top-9 largest editions per number of articles, which also had more than 1,000 active users [10]. We excluded Cebuano Wikipedia, because notwithstanding being at the moment the second-largest Wikipedia, its disproportionate growth with respect to the number of its active users has recently been fueled by massive automatic imports of articles. For fairness, we note that also the growth of Swedish Wikipedia has been led in part by automatic imports of data [10], but we have decided to keep it in given it has a reasonably large active user-base.

The WikiLinkGraphs dataset comprises 172 files for a total of 142 GB; the average size is 244 MB and the largest file is 2.4 GB. For each of the 9 languages, 18 files are available with the snapshots of the *wikilink* graph taken on March, 1st from 2001 to 2018. As specified in Section 2.1.1.5, these are CSV files that are later compressed in the standard gzip format. The remaining 10 files contain the hash-sums to verify the integrity of files and a `README`.

2.1.2.1 *Where to Find the* WikiLinkGraphs *Dataset and Its Supporting Material*

The WikiLinkGraphs dataset is published on Zenodo at https://zenodo.org/record/2539424 and can be referenced with the DOI number `10.5281/zenodo.2539424`. All other supporing datasets are available at https://cricca.disi.unitn.it/datasets/. The code used for data processing has been written in Python 3 and it is available on GitHub under the *WikiLinkGraph* organization https://github.com/WikiLinkGraphs.

All the datasets presented in this paper are released under the *Creative Commons - Attribution - 4.0 International* (CC-NY 4.0) license[12]; the code is released under the GNU General Public License version 3 or later[13].

2.1.2.2 *Basic statistics*

Tables 3 and 4 present the number of nodes (N) and edges (E) for each snapshot included in the WikiLinkGraphs dataset. The number of nodes is much larger that the number of "content articles" presented in the main pages of each Wikipedia version. For reference, in March, 2018 English Wikipedia had 5.6M articles [28], however in our snapshot there are more than 13.6M nodes. This is due to the fact that we have left in the graph redirected nodes, as described above, whilst we have resolved the links pointing to them; redirects remain as orphan nodes

12 https://creativecommons.org/licenses/by/4.0/
13 https://www.gnu.org/licenses/gpl-3.0.en.html

in the network, receiving no links from other nodes, and having one outgoing link.

Figure 3 shows a plot of the growth over time of the number of links in the WikiLinkGraphs of each language we have processed. The plot is drawn in linear scale to give a better sense of the relative absolute proportions among the different languages. After the first years all language editions exhibit a mostly stable growth pattern with the exception of Swedish, that experienced anomalous growth peaks probably due to massive bot activity.

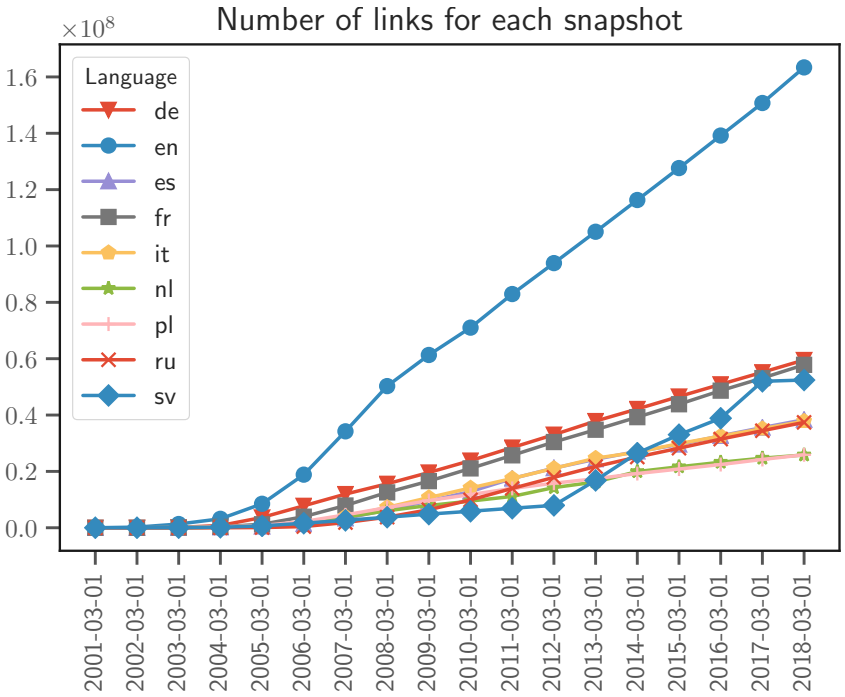


Figure 3: Overview of the growth over time of the number of links in each snapshot in the WikiLinkGraphs dataset.

2.2 ANALYSIS AND USE CASES

In this Section we analyse the WikiLinkGraphs dataset to provide some useful insights in the data that will help to demonstrate the opportunities opened by this new dataset.

2.2.1 COMPARISON WITH WIKIMEDIA'S PAGELINKS DATABASE DUMP.

To start, we compare our dataset with an existing one provided by the Wikimedia Foundation: the PAGELINKS table dump.[14] This table tracks all the internal links in a wiki [29], whether they are links in non-articles pages, link pages across different namespaces, or if they are *transcluded* in a page with a template[15]. The table presents information about the source page identifier and namespace, and the linked-to article title and namespace. There are no duplicates of the same combination of source page id, source page namespace and target title. For this reason, only distinct links in a page are recorded in the table. When updating this table, MediaWiki does not check if the target page exists or not.

lang	pagelinks all	pagelinks ns0	WLG
de	156,770,699	106,488,110	59,535,864
en	1,117,233,757	476,959,671	163,380,007
es	88,895,487	51,579,346	38,348,163
fr	270,129,151	144,469,298	57,823,305
it	187,013,995	118,435,117	37,814,105
nl	88,996,775	66,606,188	25,834,057
pl	131,890,972	79,809,667	25,901,789
ru	152,819,755	108,919,722	37,394,229
sv	133,447,975	111,129,467	52,426,633

Table 5: Comparison of the number of links between articles in the `ns0` as they result from Wikimedia's PAGELINKS database table dump (PAGELINKS ns0) and from the WIKILINKGRAPHS dataset (WLG). The total number of rows, counting links between other namespaces is given in (PAGELINKS all).

Table 5 present a comparison of the number of links extracted from the PAGELINKS table and the WIKILINKGRAPHS.

Links in WIKILINKGRAPHS are much less because links transcluded from templates are not considered. Given the specific research question

14 For the latest versions of the database dumps, all Wikipedia hyperlinks are available in the "pagelinks" files at https://dumps.wikimedia.org/.

15 We take the occasion to point out that throughout this paper we refer to "internal links" or *wikilinks* only as links between articles of the encyclopedia, however Wikipedia guidelines use the term more interchangeably to refer both to "links between articles" and "all the links that stay within the project", i.e. including links in other namespaces or that go across namespaces. Whilst it seems that the same confusion exists among the contributors of the encyclopedia, we have decided here to adopt the view for which the proper *wikilinks* are only the links between articles of the encyclopedia.

or application under consideration, it may be more suitable to include or exclude the links that were added to the page by templates; for example, to reconstruct navigational patterns it may be useful not only to consider links from templates, but also links in the navigational interface of MediaWiki.

In this sense, WikiLinkGraphs provides a new facet of the links in Wikipedia that was not readily available before. These two dataset can be used in conjunction, also taking advantage of the vast amount of metadata available accompanying the WikiLinkGraphs dataset, such as the RawWikilinks and ResolvedRedirects datasets.

2.2.2 Cross-language comparison of Pagerank scores

A simple, yet powerful application that can exploit the WikiLinkGraphs dataset is computing the general Pagerank score over the latest snapshot available [6]. Pagerank borrows from bibliometrics the fundamental idea that being linked-to is a sign of relevance [30]. This idea is also valid on Wikipedia, whose guidelines on linking state that:

> "Appropriate links provide instant pathways to locations within and outside the project that are likely to increase readers' understanding of the topic at hand." [16]

In particular, articles should link to articles with relevant information, for example to explain technical terms.

Tables 6 and 7 presents the Pagerank scores obtained by running the implementation of the Pagerank algorithm from the `igraph` library[16].

Across 7 out of the 9 languages analysed, the Wikipedia article about the *United States* occupies a prominent position being either the highest or the second-highest ranked article in direct competition with articles about countries were the language is spoken. In general, we see across the board that high scores are gained by articles about countries and cities that are culturally relevant for the language of the Wikipedia edition under consideration.

Remarkably, Dutch and Swedish Wikipedia present very different types of articles in the top-10 positions: they are mainly about the field of biology. A detailed investigation of the results and the causes for these differences is beyond the scope of this paper, but we can hypothesize differences in the guidelines about linking that produce such different outcomes.

16 https://igraph.org/c/doc/igraph-Structural.html#igraph_pagerank

#	de		en		es		fr		it	
	article	**score** ($\times 10^{-3}$)	**article**	**score** ($\times 10^{-3}$)	**article**	**score** ($\times 10^{-3}$)	**article**	**score** ($\times 10^{-3}$)	**article**	**score** ($\times 10^{-3}$)
1	Vereinigte Staaten	1.646	United States	1.414	Estados Unidos	2.301	France	2.370	Stati Uniti d'America	3.076
2	Deutschland	1.391	World War II	0.654	España	2.095	États-Unis	2.217	Italia	1.688
3	Frankreich	1.020	United Kingdom	0.618	Francia	1.281	Paris	1.228	Comuni della Francia	1.303
4	Zweiter Weltkrieg	0.969	Germany	0.557	Idioma inglés	1.073	Allemagne	0.977	Francia	1.292
5	Berlin	0.699	The New York Times	0.527	Argentina	0.955	Italie	0.812	Germania	1.257
6	Österreich	0.697	Association football	0.525	Alemania	0.909	Royaume-Uni	0.773	Lingua inglese	1.228
7	Schweiz	0.691	List of sovereign states	0.523	Latín	0.867	Anglais	0.764	Roma	0.961
8	Englische Sprache	0.620	Race and ethnicity in the United States Census	0.500	Animalia	0.866	Français	0.748	Centrocampista	0.861
9	Italien	0.614	India	0.491	México	0.853	Espèce	0.731	Europa	0.805
10	Latein	0.599	Canada	0.468	Reino Unido	0.820	Canada	0.710	2004	0.778

Table 6: Top-10 articles with the highest Pagerank score computed over the most recent snaphost of the WikiLinkGraphs dataset (2018-03-01).

#	nl		pl		ru		sv	
	article	**score** ($\times 10^{-3}$)	**article**	**score** ($\times 10^{-3}$)	**article**[‡]	**score** ($\times 10^{-3}$)	**article**	**score** ($\times 10^{-3}$)
1	Kevers	3.787	Stany Zjednoczone	2.763	Soedinjonnye Shtaty Ameriki	3.290	Familj (biologi)	5.489
2	Vlinders	3.668	Polska	2.686	Sojuz Sovetskih Socialisticheskih Respublik	2.889	Släkte	5.184
3	Dierenrijk	3.294	Francja	2.360	Rossija	2.233	Nederbörd	4.696
4	Vliesvleugeligen	3.084	Język angielski	2.110	Francija	1.190	Grad Celsius	4.144
5	Insecten	2.164	Łacina	1.914	Moskva	1.135	Djur	4.114
6	Geslacht (biologie)	2.101	Niemcy	1.698	Germanija	1.080	Catalogue of Life	3.952
7	Soort	1.954	Włochy	1.229	Sankt-Peterburg	0.881	Årsmedeltemperatur	3.878
8	Frankrijk	1.932	Wielka Brytania	1.124	Ukraina	0.873	Årsnederbörd	3.366
9	Verenigde Staten	1.868	Wieś	1.095	Velikobritanija	0.811	Växt	2.810
10	Familie (biologie)	1.838	Warszawa	1.083	Italija	0.763	Leddjur	2.641

Table 7: Top-10 articles with the highest Pagerank score computed over the most recent snaphost of the WikiLinkGraphs dataset (2018-03-01). (‡) Russian Wikipedia article titles are transliterated.

2.3 RESEARCH OPPORTUNITIES USING THE WIKILINKGRAPHS DATASET

The WikiLinkGraphs dataset and its supporting dataset can be useful for research in a variety of contexts. Without pretending to be exhaustive, we present here a few examples.

2.3.1 GRAPH STREAMING.

Stream data processing has gained particular consideration in recent years since it is well-suited for a wide range of applications, and streaming sources of data are commonplace in the big data era [31]. The WikiLinkGraphs dataset, together with the RawWikilinks dataset, can be represented as a graph stream, i.e. a collection of events such as node and link additions and removals. Whilst other datasets are already available for these kind of problems, such as data from social networks, WikiLinkGraphs, being open, can facilitate the reproducibility of any research in this area and can be used as a benchmark.

2.3.2 LINK RECOMMENDATION.

West, Paranjape, Ashwin and Leskovec [32] have studied the problem of identifying missing links in Wikipedia using web logs. More recently, Wulczyn, West, Zia, and Leskovec [33] have demonstrated that it is possible to produce personalized article recommendations to translate Wikipedia articles across language editions. The WikiLinkGraphs dataset could be used in place of the web logs for a similar study on recommending the addition of links in a Wikipedia language edition based on the fact that some links exist between the same articles in other Wikipedia language editions.

2.3.3 LINK ADDITION AND LINK REMOVAL.

The problem of predicting the appearance of links in time-evolving networks has received significant attention [34]; the problem of predicting their disappearance, on the other hand, is less studied. Preusse and collaborators [35] investigated the structural patterns of the evolution of links in dynamic knowledge networks. To do so, they adapt some indicators from sociology and identify four classes to indicate growth, decay, stability and instability of links. Starting from these indicators, they

identify the underlying reasons for individual additions and removals of knowledge links. Armada et al. [36] investigated the link-removal prediction problem, which they call the *unlink prediction*. Representing the ever-evolving nature of Wikipedia links, the WikiLinkGraphs dataset and the RawWikilinks datasets are a natural venue for studying the dynamics of link addition and link removal in graphs.

2.3.4 ANOMALY DETECTION.

A related problem is the identification of spurious links, i.e., links that have been erroneously observed [37, 38]. An example of the application of this approach is the detection of links to spam pages on the Web [39]. Similarly, the disconnection of nodes has been predicted in mobile ad-hoc networks [40].

2.3.5 CONTROVERSY MAPPING.

Given the encyclopedic nature of Wikipedia, the network of articles represents an emerging map of the connections between the corresponding concepts. Previous work by Markusson and collaborators [22] has shown how a subportion of this network can be leveraged to investigate public debate around a given topic, observing its framing and boundaries as emerging from the grouping of concepts in the graph. The availability of the WikiLinkGraphs dataset can foster controversy mapping approaches to study any topical subpart of the network, with the advantage of adding a temporal and a cross-cultural dimension.

2.3.6 CROSS-CULTURAL STUDIES.

More than 300 language editions of Wikipedia have being created since its inception in 2001 [41], of which 291 are actively maintained. Despite the strict neutral point of view policy which is a pillar of the project [42, 43], different linguistic communities will unavoidably have a different coverage and different representations for the same topic, putting stronger focus on certain entities, and or certain connections between entities. As an example, the articles about bullfighting in different languages may have a stronger connection to concepts from art, literature, and historical figures, or to concepts such as cruelty and animal rights [44]. Likewise, the networks from different language versions give prominence to different influential historical characters [45, 46]. The WikiLinkGraphs dataset allows to compare the networks of 9 editions of Wikipedia, which are not only big editions, but have a fairly

large base of contributors. In this paper, we have presented a simple comparison across the 9 languages represented, and we have found an indicator of the prominence of the United States and the local culture almost across the board. Many more research questions could be addressed with the WikiLinkGraphs dataset.

2.4 CONCLUSIONS

The dataset we have presented, WikiLinkGraphs, makes available the complete graph of links between Wikipedia articles in the nine largest language editions.

An important aspect is that the dataset contains only links appearing in the text of an article, i.e. links intentionally added by the article editors. While the Wikimedia APIs and dumps provide access to the currently existing wikilinks, such data represent instead all hyperlinks between pages, including links automatically generated by templates. Such links tend to create cliques, introducing noise and altering the structural properties of the network. We demonstrated that this is not an anecdotal issue and may have strongly affected previous research, as with our method we obtain less than the half of the links contained in the corresponding Wikimedia pagelinks dump.

Another limitation of the Wikimedia dumps is that data are available only for the current version of Wikipedia or for a recent snapshot; the WikiLinkGraphs dataset instead provides complete longitudinal data, allowing for the study of the evolution of the graph over time. We provided both yearly snapshots and the raw dataset containing the complete history of every single link within the encyclopedia.

The WikiLinkGraphs dataset is currently made available for the nine largest Wikipedia language editions, however we plan to extend it to other language editions. As the code of all steps is made available, other researchers can also extend the dataset by including more languages or a finer temporal granularity.

Beyond the opportunities for future research presented above, we believe that also research in other contexts can benefit from this dataset, such as Semantic Web technologies and knowledge bases, artificial intelligence and natural language processing.

Part II

RELEVANCE ON A GRAPH

Surfing the links between Wikipedia articles constitutes a valuable way to acquire new knowledge related to a topic. The density of connections in Wikipedia makes that, starting from a single page, it is possible to reach virtually any other topic on the encyclopedia. This abundance highlights the need for dedicated algorithms to identify the topics that are more relevant to a given concept. In this sense, a well-known option is *Personalized PageRank*; its performance, however, is hindered by pages with high indegree that function as hubs and obtain high scores regardless of the starting point. In this work, we present *CycleRank*, a novel algorithm based on cyclic paths aimed at finding the most relevant nodes related to a topic. We compare the results of *CycleRank* with those of *Personalized PageRank* and other algorithms derived from it, both with qualitative examples and with an extensive quantitative evaluation. We perform different experiments based on ground truths such as the number of clicks that links receive from visitors and the set of related articles highlighted by editors in the "See also" section of each article. We find that *CycleRank* tends to identify pages that are more relevant to the selected topic. Finally, we show that computing *CycleRank* is two orders of magnitude faster than computing the other baselines.

3

CYCLERANK, OR THERE AND BACK AGAIN: PERSONALIZED RELEVANCE SCORES FROM CYCLIC PATHS ON DIRECTED GRAPHS

Wikipedia is one of the biggest and most used sources of knowledge on the Web. As of this writing, it is the fifth most visited website in the world [9]. Wikipedia exists in more than 290 active different language editions [10], and its pages have been edited over 2.5 billion times.

Wikipedia is not only a huge repository and collaborative effort; it is also a giant hypertext in which each article has links to the concepts that are deemed relevant to it by the editors [17].

Such vast network emerging from the collaborative process provides a rich representation of the connections between concepts, entities and pieces of content, aimed at encompassing "the sum of all human knowledge" [47]. This huge graph has been leveraged for different purposes in a variety of fields including natural language processing [18], semantic networks [19], cross-cultural studies [45, 46], complex networks modelling [21], automatic and human navigation of information networks [23, 48].

While one cannot assume that a single article completely encapsulates a concept [49], the link network can be useful in defining the context of an article. Previous research in controversy mapping has shown how this network can be leveraged to analyze the dominating definition of a topic, such as *"Geoengineering"* [22], shedding light on its boundary, context and internal structure. Furthermore, each linguistic community in Wikipedia produces a different network, which allows for comparing the emerging definition of a topic across different language editions [44].

The connections between Wikipedia articles are valuable, but they are also very abundant. The English version has more than 160 million links between its 5.7 million articles [1]. How can one find guidance within this wealth of data? In particular, how can we analyze the network around a specific topic, to characterize its definition as emerging from the collaborative process?

The contribution of this work is a novel approach to make sense of the Wikipedia link network, capable to answer *queries* like "Which are the concepts that are more relevant with respect to a given topic?"

Such inquiries can be translated into a graph problem. The topic we are interested in can be represented by one article, i.e. a node in the graph called *reference node*. Given a reference node r, we want to assign a score to every other node in the graph that captures its relevance to r, based on the link structure. The final output is a ranking of nodes, such that the more relevant nodes are ranked higher.

One established algorithm to answer this question is *Personalized PageRank*: a variant of *PageRank* where the user can specify one or more nodes as queries (seeds) and obtain a score for all the other nodes in the graph that measures the relatedness with respect to the seeds. However, we have found that, when applied in the context of Wikipedia, this algorithm does not produce satisfactory results since it usually includes very general articles in top positions.

To overcome these limitations, we have developed a novel algorithm to find the most relevant nodes in the Wikipedia link network related to a topic. The technique, called *CycleRank*, takes advantage of the cycles that exist between the links and produces a ranking of the different articles related to one chosen by a user. In this way, this technique accounts for links in both directions, and it can provide results that are more accurate than those produced by the well-known *Personalized PageRank* algorithm.

The Chapter is organized as follows. We first formalize the problem we want to solve in Section 3.1. We provide insights on why *Personalized PageRank* is not a good choice in Section 3.2 and we discuss related work in Section 3.3. We describe *CycleRank* in Section 3.4 and we evaluate its performance in Section 3.5. Conclusions are drawn in Section 3.6.

3.1 PROBLEM STATEMENT

Given a graph $G = (V, E)$, where V is a finite set containing n nodes (articles) and $E \subseteq V \times V$ is a set containing m directed edges (links between articles), we are seeking to build a *ranking*, i.e. an order relationship between nodes based on their relevance with respect to a reference node r.

In order to achieve this goal, we build a *ranking function* rf_r that assigns a non-negative score to every node $v \in V$:

$$rf_r : V \mapsto [0, +\infty)$$

The ranking $\nu_r = [v_1, v_2, \ldots, v_n]$ is thus given by the total order of scores: if $rf_r(v_i) > rf_r(v_j)$, then node v_i should appear before node v_j (i.e., $i < j$). Note that we assume that there are no ex-aequo in any given ranking; this can by achieved by breaking ties randomly.

3.2 BACKGROUND

The *PageRank* algorithm represents an established relevance measure for directed networks [50]; its variant *Personalized PageRank* may be used to measure relevance within a certain context. *PageRank* is a measure based on incoming connections, where connections from relevant nodes are given a higher weight. Intuitively, the *PageRank* score of a node represents the probability that, following a random path in the network, one will reach that node. It is computed in an iterative process, as the *PageRank* score of a node depends on the *PageRank* scores of the nodes that link to it. There are however efficient algorithms to compute it. The idea behind *PageRank* is that of simulating a stochastic process in which a user follows random paths in a hyperlink graph. At each round, the user either keeps surfing the graph following the link network with probability α, or is teleported to a random page in the graph with probability $1 - \alpha$. The parameter α is called *damping factor* and is generally assumed to be 0.85 [51, 52]. During the surfing process, the algorithm assumes equal probability of following any hyperlink included in a page; similarly, when teleported, every other node in graph can be selected with equal probability.

Personalized PageRank is a variant of the original *PageRank* algorithm, where the user provides a set of *seed* nodes. In *Personalized PageRank*, teleporting is not directed to some random node taken from the entire graph, but to one taken from the seed set. In this way, the algorithm models the relevance of nodes around the selected nodes, as the probability of reaching each of them, when following random walks starting from a node in the seed set.

Limitations of PageRank. At first look, *Personalized PageRank* seems to be suitable for our use case, as it can be used to represent a measure of relevance of Wikipedia articles strongly linked (directly or indirectly) to the seed.

However, we found unsatisfactory results when applying this algorithm. Very often, pages that are found to be very central in the overall network, such as *"United States"* or *"The New York Times,"* are included in the top results of completely unrelated queries.

Such central articles act as hubs in the graph; they have such a strong relevance overall that, even starting from a seed article which is not

specially related to them, one is very likely to end up reaching them while exploring the graph.

We argue that this is due to different factors. First, paths of any length can be followed; therefore, in a densely connected graph many paths will tend to converge towards the most relevant nodes. This aspect can be limited only partially by lowering the value of the damping factor.

Second, *PageRank* only accounts for inlinks, not for outlinks. This is reasonable for web search and other contexts where inlinks are a good proxy for relevance, as they represent somehow the value attributed to a node by the other nodes of the graph. In such cases, outlinks have basically no value: it is very easy to add into one's web page many outlinks to other pages. In the context of Wikipedia, instead, links from an article to other articles may be subject to being inserted and accepted by the editors' community as much as incoming links from other articles. So, both outgoing and incoming links can be considered as indicators of relevance.

In particular, outlinks to other pages from an article can be a very valuable indicator that these pages are actually related to the topic. For example, if an article contains links to *"Computer Science,"* then we can assume that its content is related to *"Computer Science;"* on the other hand, we can expect the article *"United States"* to have only a few links to articles related to *"Computer Science,"* as it is not the main subject of the article.

3.3 RELATED WORK

We discuss here relevant related studies used to establish the foundation of our work.

Identifying related content in Wikipedia. Schwarzer et al. [53] have studied the problem of recommending relevant Wikipedia articles, starting from a given article: they used citation-based document similarity measures, such as Co-Citation (CoCit), and Co-Citation Proximity Analysis (CPA). They compared the performance of these two measures against a more general test-based measure implemented in the MoreLikeThis function provided by Apache Lucene. They evaluate the effectiveness of these measures using two datasets as ground truth: the *See-Also* dataset, consisting of a list of links added as related resources to a Wikipedia article; and the *ClickStream* dataset, consisting of a list of links in an article ordered by the number of clicks that they have received from Wikipedia readers. The authors show that MLT finds articles with similar structure that use similar words, while citation-based measures are better able to find topically-related infor-

mation, with CPA consistently outperforming CoCit. With respect to their work, the main difference of our approach is that we focus on the problem of finding relevant related nodes on a graph, and we do not use the text of Wikipedia articles. Our approach has not only the advantage of being completely language-independent, but it is applicable to a much broader set of problems.

Link Structure in Wikipedia. The foundation of this work is based on the idea that inlink and outlinks in Wikipedia have a similar role to establish relevance. Kamps and Koolen [54] performed a comparative analysis of the link structure of Wikipedia and a selection of the Web - built from `.gov` websites - and found that traditional information retrieval algorithms such as HITS do not work well on Wikipedia. The root case of this problem, as they observe, is that in Wikipedia inlinks and outlinks are good indicators of relevance, contrasting the general behavior of the web where only the former provide this indication.

PageRank and variations. Boldi et al. [52] studied the behavior of *Personalized PageRank* as a function of the damping factor α. While they acknowledge that a popular choice of α is 0.85 – following the suggestion of the authors of *PageRank* itself [50] – they discuss both the possibility of choosing smaller value of α as well as values close to 1, finding the latter to be a choice with several theoretical and computational shortcomings.

Gleich et al. [51] studied the problem of determining the empirical value for α from the visitor logs of a collection of websites, including Wikipedia. They found Wikipedia visitors do not tend to teleport, and estimated the distribution of the values of α for Wikipedia to a β distribution with maximum at $\alpha = 0.30$. In our experiments, we have considered $\alpha = 0.30$, and $\alpha = 0.85$ as values for the damping parameter when executing *PageRank*.

We focus on the variations of *PageRank* that use reverse links or take into account both the existence of inlinks and outlinks. In 2010, Chepelianskii [55] introduced the idea of calculating the pagerank score of nodes on the transposed graph – called *CheiRank* – as well as on the original graph and performed a study of the correlation between the two scores on a collaboration network. Later, Zhirov [56] combined *CheiRank* and *PageRank* to produce a single two-dimensional ranking of Wikipedia articles, *2DRank*. This method does not assign a score to each node, but just produces a ranking. It was used together with *PageRank* to rank biographies across different language editions [46].

Cycles in Non-Directed Graphs. Finally, we present related work about cycles in undirected graphs. This area of work is interesting because it provides a broader context in which to insert our algorithm and it could be used as a guide to extend our algorithm to undirected graphs.

However, we consider this line of work to be very different in scope and purpose from our current work. It has been shown recently that graphs with different structure can be distinguished from one another using a measure defined with non-backtracking cycles, i. e. a closed walk that does not retrace any edge immediately after traversing them [57]. This method is tied to the idea of using the length spectrum of a graph from its Laplacian matrix. Graph spectra are extensively covered in literature [58].

3.4 THE CYCLERANK ALGORITHM

We propose a more general approach to the problem, defining a new measure of the relevance with respect to a given node in a directed network, that accounts for both incoming and outgoing links. We call this measure *CycleRank*, as it is based on the idea of circular random walks.

Starting from the observation that *PageRank* is not suitable for our context because random walks may easily lead to paths that are not related to the topic under consideration, we thought of the idea of only considering random walks coming back to the starting point within a maximum of K steps. In this way, we guarantee that we only touch pages that are, at least indirectly, both linked from and linking to the reference article. Furthermore, we do not need a damping factor, as we can assume that all walks just start from the reference node and come back.

Intuitively, a node that is linked from the reference article but does not link to it is likely to be a concept that is not related to that subject, even if it is important to its definition. Specularly, a node that links to the reference article but is not linked from it is likely to be related to it, but not relevant. Nodes that are linked both from and to a reference node are the ones that we expect to be relevant.

Extending this principle, we want then to be able to quantify the relevance of a node with respect to a given reference node, accounting also for the indirect links, i.e. for the amount of paths that can be found linking it from and to the reference node.

We do this by counting the cycles involving the reference node that pass through a given other node. As short distances represent a stronger relationship, shorter cycles should get higher weights.

We define the *CycleRank* score $CR_r(i)$ of a node i with respect to a reference node r as follows:

$$CR_r(i) = \sum_{k=2}^{K} \ell_r^k(i) \cdot \sigma(k) \tag{1}$$

Algorithm 1 *CycleRank*

Input: G: a directed graph $G = (V, E)$
Input: r: the refence node
Input: K: threshold parameter, $K \in \mathbb{N}^+$
Output: *score*: a vector of CycleRank scores for each $v \in V$

```
1: function CycleRank(G, r, K)
2:     r ← FilterGraph(G, r, K)
3:     score ← ComputeScore(G, r, k)
4:     return score
5: end function
```

where $\ell_r^k(i)$ is the number of simple cycles of length k that include both node i and r, K is a parameter representing the maximum length considered for cycles, and $\sigma(\cdot)$ is a scoring function giving different weights to cycles of different length.

In this way, given a reference node r, the *CycleRank* of a node i represents the number of cycles including both r and node i, weighted by the scoring function, which depends on the length of the cycle.

The reference node is also considered in this computation, and it gets the maximum *CycleRank* score as by definition it is included in all the cycles considered.

The threshold K is a parameter whose value can be specified according to the context. It can be set to infinite, but it will never exceed the number of nodes n. It can be typically set to a much lower value for two main reasons: to reduce the computational load and to avoid potential noise deriving from long cycles that include popular nodes far from the reference node. For both reasons, and after manually inspecting the results for different values of K, we chose to apply thresholds $K = 3$ and $K = 4$, which produce good results with a limited computational effort.

The main *CycleRank* algorithm is shown in Algorithm 1. In order to optimize the score computation, we first filter the graph G through function FilterGraph(G, r, K), removing those nodes that could never appear in cycles including the reference node r with length limited by K. We then compute the score on this network using function ComputeScore().

3.4.1 PRELIMINARY FILTERING

To reduce the size of the network, Algorithm 2 employs two breadth-fist searches to compute the distance from and to the reference node r,

and discards all the nodes whose cumulative distance (back and forth) is larger than K:

1. We compute the distance $df[v]$ of each node v from the reference node by performing a breadth-first visit of the graph, early-terminating the visit when we reach distance K (Algorithm ETBFS – Early-Terminated Breadth-First Search, Line 6);
2. We discard all nodes for which $df[i] > K - 1$ (Line 7), including those unreachable from r whose distance is $+\infty$. The function removeNodes($G, key{=}cond$) eliminates all nodes in G that do not satisfy the condition expressed by the boolean expression $cond$;
3. We compute the distance $dt[v]$ on the transposed network, i.e. the distance from each node v to the reference node (Line 9);
4. We compute the length $df[v] + dt[v]$ of the minimum cycle including v and the reference node and we discard all nodes for which $df[v] + dt[v] > K$ (Line 10);

Algorithm 2 FilterGraph

Input: G: a directed graph $G = (V, E)$
Input: r: the refence node
Input: K: threshold parameter, $K \in \mathbb{N}^+$
Output: r: the reference node in the filtered graph

```
1:  function FilterGraph(G, r, K)
2:      for v ∈ V do
3:          df[v] ← +∞          ▷ Distance from the reference node r
4:          dt[v] ← +∞          ▷ Distance from the reference node r
5:      end for
6:      ETBFS(G, r, K, df)                                  ▷ Step 1
7:      removeNodes(G, key=(df[v] > K − 1))
8:      r ← remapNodes(G, r)
9:      ETBFS(G^T, r, K, dt)                                ▷ Step 2
10:     removeNodes(G, key=(df[v] + dt[v] > K))
11:     r ← remapNodes(G, r)
12:     return r
13: end function
```

In this way we discard all the nodes that are not reached by any cycle of length lower than K. We remap node indexes at each step (Lines 8 and 11) so we effectively work with smaller networks. It should be noted that, in case of $K \geq n$, only nodes unreachable from r will be removed, as the length of simple cycles is bounded by the number of nodes n. The removed nodes will all receive a score of zero.

3.4.2 CYCLE ENUMERATION

Algorithm 3 ComputeScore

Input: G: a directed graph $G = (V, E)$
Input: r: the refence node
Input: K: threshold parameter, $K \in \mathbb{N}^+$
Output: *score*: a vector of CycleRank scores for each $v \in V$

```
1:  function ComputeScore(G, r, K)
2:      for v ∈ V do
3:          score[v] ← 0                          ▷ CycleRank score
4:          blocked[v] ← false
5:          B[v] ← List()
6:      end for
7:      S ← Stack()
8:      circuit(G, r, r, K, S)
9:      return score
10: end function
```

We then proceed to enumerate all simple cycles in the reduced graph. Our algorithm is based on Johnson's algorithm [59], limited to the query node r and early-terminated. Algorithm 3 presents the details. Each node in our algorithm is associated with the following values:

- $score[v]$, the *CycleRank* score of node v
- $blocked[v]$, a boolean indicating whether v cannot be further visited when searching for a cycle because we already went through it. The purpose of this vector is to avoid going through the same node more than once, since we are only interested in simple cycles.
- $B[v]$, a list of nodes that can be unblocked when node v is unblocked.

These variables are then considered global in the rest of the algorithms, to avoid long signatures.

Cycle discovery is performed through a recursive backtrack visit (Algorithm 4). In function circuit(G, r, v, K, S), G is the graph, r is the reference node, v is the current visited node, K is the threshold and S is a stack of nodes that have been visited so far.

We use K to early-terminate the search for a cycle when we arrive at the maximum length: in Line 3, we check that current whether the current size of the stack is smaller than K, in which case we can proceed in exploring the graph; otherwise, the function returns immediately.

The Circuit() function works by recursively visiting the nodes on the graph; when we visit a node v we add it to the stack S and mark it as blocked, then we visit its neighbors by looping over the adjacent nodes

$v.adj$. If the neighbor w we are visiting is the target node, r in our case, then we have found a cyclic path: the score is updated by calling function UpdateScore() and the unblocking flag *flag* is set to true. Otherwise, we check if w is unblocked, if so it can be visited and we call *circuit* recursively.

After visiting all the neighbors of v, we check if the current node can be unblocked. Unblocking happens when v is part of a path that formed a cycle. The unblock(G, v) function at Line 17 is the same as the one defined by Johnson [59] and we omit here for reasons of space. If we unblock a node v, we unblock all the parent nodes that could lead to v, stored in $B[v]$. In this way, we are able to explore alternative paths that form a cycle.

3.4.3 SCORE COMPUTATION

Function UpdateScore$(score, S)$ updates the score of the nodes recorded in a stack S of length k, by adding $\sigma(k)$ to the score of every node $v \in S$. Several scoring functions σ can be used; in general, a scoring function should capture the idea that longer cycles contribute less.

We use an exponentially decaying function $\sigma_{\exp}(k) = e^{-k}$ where the length of the cycle is k.

We have chosen the denominator to be exponential in the number of nodes. We will present some data to support this choice in the Experimental Evaluation Section where we show that the number of cycles increases more than exponentially with cycle length for our dataset. Intuitively, an exponentially-dacaying scoring function limits the possibility that short cycles become neglectable compared to long cycles in the computation of CycleRank, especially for higher values of K. Other scoring function can be considered, based on the problem at hand and according to structural properties of the network.

Algorithm 4 CIRCUIT

Input: G: a directed graph $G(V, E)$
Input: v: a node $v \in V$
Input: r: the reference node $r \in V$
Input: K: a positive integer, $K \in \mathbb{N}^+$
Input: S: a stack of nodes
Output: $flag$: a boolean

```
1: function CIRCUIT(G, v, r, K, S)
2:     flag ← false
3:     if S.size() < K then
4:         S.push(v)
5:         blocked[v] ← true
6:         for each w ∈ v.adj() do
7:             if w = r then
8:                 UPDATESCORE(score, S)
9:                 flag ← true
10:            else if ¬w.blocked then
11:                if CIRCUIT(G, w, r, K, S) then
12:                    flag ← true
13:                end if
14:            end if
15:        end for
16:        if flag then
17:            UNBLOCK(G, v)
18:        else
19:            for each w ∈ v.adj() do
20:                if v ∉ w.B then
21:                    w.B.push_back(v)
22:                end if
23:            end for
24:        end if
25:        S.pop()
26:    end if
27:    return flag
28: end function
```

Algorithm 5 UPDATESCORE

Input: *score*: a stack representing a cycle
Input: S: a stack representing a cycle

```
1: function UPDATESCORE(score, S)
2:     ℓ ← LEN(S)
3:     for each v ∈ S do
4:         score[v] = score[v] + σ(ℓ)
5:     end for
6: end function
```

3.5 EXPERIMENTAL EVALUATION

This section is organized as follows: in Section 3.5.1 we describe the dataset that we have used for our experimental evaluation; Section 3.5.2 describes alternative approaches that we will use to compare to our proposed approach in addition to *Personalized PageRank*; and in Section 3.5.3 we provide some details about the implementation of each algorithm. Section 3.5.4 provides some example results and their qualitative description for each algorithm and in Section 3.5.5, we provide a detailed quantitative evaluation with three different evaluation measures based on different ground truth data. Finally, in Section 3.5.6 we compare the execution time of our proposed approach against the alternatives.

3.5.1 DATASET DESCRIPTION

For our analysis, we used the WikiLinkGraphs dataset, consisting of the network of internal Wikipedia links for the 9 largest language editions [1]. The dataset has been developed by us and it is publicly available on Zenodo.[1] The graphs have been built by parsing each revision of each article to track links appearing in the main text, discarding links that were automatically inserted by templates. The dataset contains yearly snapshots of the network and spans 17 years, from the creation of Wikipedia in 2001 to March 1st, 2018. For the experiments in this Chapter we focused on the WikiLinkGraphs snapshot from English Wikipedia taken on March, 1st 2018. This graph has $N = 13,685,337$ nodes and $E = 163,380,007$ edges.[2]

Figure 4 presents the number of cycles by length for a sample of 100 nodes chosen randomly from our dataset. For each page we plot a triplet of points corresponding to the number of simple cycles of length $k = 2$, 3, and 4 respectively, that go through that node. We shift this triplet of points by a random offset along the horizontal axis for ease of reading.

Table 8 presents the top-10 pages by indegree and outdegree in the graph. We can see from the table that indegree dominates outdegree by several orders of magnitude. This implies that ranking the top pages by degree (undirected) is *de facto* equivalent to ranking them by indegree.

1 https://zenodo.org/record/2539424 – DOI: 10.5281/zenodo.2539424.

2 Wikipedia contains also special pages known as *redirects*, i.e. alternative articles titles. These pages appear in the graph as nodes with a single outgoing edge and typically no incoming edges. Our dataset consolidates alternative titles in a single node; for this reason, the count of nodes in our graph differs from official count of the English Wikipedia.

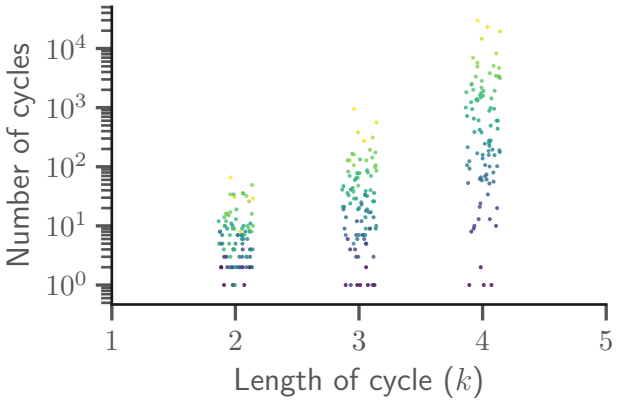


Figure 4: Number of cycles (log scale) by length for a sample of 100 random nodes. For each node in the sample, we have computed the number of cycles of length $k = 2, 3, 4$. Points representing the values for a single page are shifted on the x-axis by a random offset and colored with the same color. The color gradient depends on the value at $k = 4$.

The difference between the top-1000 pages by indegree and the top-1000 pages by degree is of just 34 pages.

Table 9 presents the top-10 results by global *PageRank* score. Global *PageRank* and in-degree are highly correlated, regardless of the value of the damping parameter. When $\alpha = 0.85$ the two rankings have a Kendall correlation coefficient of $\tau = 0.60$ (over all pages with in-degree greater than zero, $n = 8,305,031$); if we limit the two rankings to the top 1,000 articles, they are still highly correlated with $\tau = 0.56$.

3.5.2 ALTERNATIVE APPROACHES

We describe briefly some alternative approaches that we will compare *CycleRank* with: beyond *Personalized PageRank*, we will consider the personalized versions of *CheiRank* and *2DRank*, which are all based on *Personalized PageRank*. Given that we will only consider personalized versions from now on, for the sake of brevity we drop the specifier "personalized" when mentioning the algorithms.

3.5.2.1 *CheiRank*

CheiRank[3] is a ranking algorithm first proposed by Chepelianskii [55], that consists in applying the *PageRank* algorithm on the transposed

3 This algorithm was named later named *CheiRank* by Zhirov, Zhirov, and Shepelyansky [56] for its assonance which the name of the original author and because the name *CheiRank* in Russian sound similar to a phrase which translates to "whose rank".

Table 8: Top-10 pages by indegree and outdegree over the most recent snapshot of the WikiLinkGraphs dataset (2018-03-01)

#	indegree		outdegree	
	article	**degree**	**article**	**degree**
1	United States	332,557	List of current U.S. state legislators	8,019
2	Animal	164,549	List of least concern birds	7,907
3	Association football	146,836	List of people from Illinois	7,827
4	India	126,107	List of birds of the world	6,849
5	World War II	124,806	List of stage names	6,677
6	Arthropod	122,742	List of cities, towns and villages in Kerman Province	5,839
7	Germany	121,705	List of film director and actor collaborations	5,804
8	Insect	118,628	Index of Telangana-related articles	5,747
9	Canada	115,779	Index of Andhra Pradesh-related articles	5,684
10	New York City	107,831	List of municipalities of Brazil	5,585

Table 9: Top-10 pages by (global) *PageRank* over the most recent snapshot of the WikiLinkGraphs dataset (2018-03-01)

#	**PageRank,** $\alpha = 0.30$		**PageRank,** $\alpha = 0.85$	
	article	**score** ($\times 10^{-4}$)	**article**	**score** ($\times 10^{-4}$)
1	United States	4.64	United States	14.14
2	Animal	3.13	World War II	6.54
3	Arthropod	2.49	United Kingdom	6.18
4	Association football	2.45	Germany	5.57
5	Insect	2.42	The New York Times	5.27
6	Germany	2.16	Association football	5.25
7	List of sovereign states	2.11	List of sovereign states	5.23
8	India	2.03	Race and ethnicity in the United States Census	5.00
9	Moth	1.85	India	4.91
10	National Register of Historic Places	1.66	Canada	4.68

graph G^T, i. e. all link directions are inverted. This corresponds to transposing the adjacency matrix when computing *PageRank* and results in computing the conjugated Google matrix G^*.

CheiRank is analogous to *PageRank*, but it assigns a higher score to nodes with higher outdegree. In the Wikipedia dataset that we are using there are list articles that have several thousands outgoing links, as shown in Table 8. As expected, the articles with the highest global *CheiRank* score are list articles having high outdegree: out of the top 100 results by global *CheiRank* with $\alpha = 0.30$, 87 have the word `List`, `Lists`, or `Index` in the title. In the following, we are not showing results for *CheiRank* since it suffers from analogous limitations as *PageRank*, and it never resulted on-par with the most performing algorithms in our experiments.

3.5.2.2 *2DRank*

2DRank combines *CheiRank* and *PageRank* [56]; it ranks all nodes in a graph, but it does not produce a score as *PageRank* or *CheiRank* do. Instead, given the rankings $\nu^{(\mathrm{PR})}$ and $\nu^{(\mathrm{ChR})}$ produced by *PageRank* and *CheiRank* respectively, *2DRank* takes the minimum position in which a given node appears in both ranking and builds a new ranking. This process can be visualized in the two-dimensional cartesian plane: xOy, we build a series of squares with one vertex in the origin, two sides formed by the cartesian axes and the other two drawn at integer values. Thus, the first square is identified by $(0;0)$, $(0;1)$, $(1;1)$, and $(1;0)$; the second by $(0;0)$, $(0;2)$, $(2;2)$, and $(2;0)$, and so on. By interpreting the position of an item in the *PageRank* (p) and *CheiRank* rankings (p^*) as the coordinates of a point $P(p, p^*)$, this point will fall on one of the edges of the squares drawn before. The position of a node in *2DRank* is given by assigning a progressive number to each item, starting from the points that lay on inner squares; if two points lay on the same square the algorithm chooses the one closest to either axis first.

Figure 5 shows the computation of 2DRank for a toy graph of 7 nodes.

3.5.3 IMPLEMENTATION AND REPRODUCIBILITY

We implemented *CycleRank* in C++. For *PageRank* and *CheiRank* we used the `igraph` library,[4] *2DRank* was computed directly from *PageRank* and *CheiRank* using a Python script. All code is available under an open-source license at: https://github.com/CycleRank/cyclerank.

4 https://igraph.org/c/

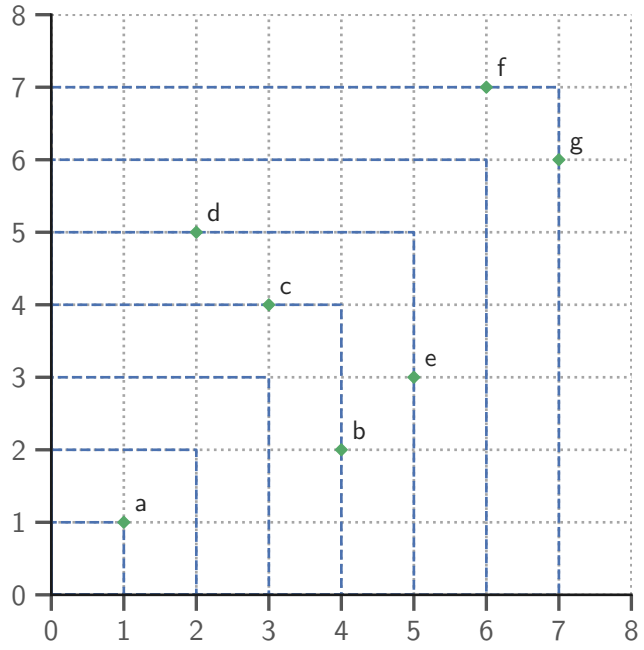


Figure 5: Toy example of the computation of 2Drank for a graph with 7 nodes ranked respectively: $\nu_P = [a, d, c, b, e, f, g]$ and $\nu_C = [a, b, d, e, c, f, g]$. The final ranking is $\nu_{2D} = [a, b, c, d, e, f, g]$.

3.5.4 QUALITATIVE COMPARISON

We first present the results of a comparison between *Personalized PageRank* and *CycleRank* directed over a variety of terms. Then, we focus on the article *"Fake news"* and we explore the capabilities of *CycleRank* more in-depth by performing directed a longitudinal analysis over two snapshots of the Wikipedia link graph at a distance of one year, and a cross-language analysis over 8 languages.

3.5.4.1 *Comparison of CycleRank and PageRank*

Tables 10 and 11 present a comparison between the top-10 results with the highest scores obtained with the *Personalized PageRank* and *CycleRank*, computed with different reference nodes over the wikilink graph of the English Wikipedia taken as of March 1st, 2018. These results highlight the limitation of *Personalized PageRank* that we have described in Section 3.2: in the top positions we see articles such as *"United States"*, *"The New York Times"*, *"World War II"* and *"Germany"*; these articles act as attractors for the unconstrained random walk of *Personalized PageRank* since they have a very high in-degree and have among the highest values of the *PageRank* score in the overall network. Indeed, they are respectively in 1st (*"United States"*), 5th (*"The New York Times"*), 2nd (*"World War II"*) and 4th position (*"Germany"*) in the overall *PageRank* ranking for that network.

However, there are much fewer paths that connect these articles back to the reference nodes. As a result, these articles appear in much lower positions in the ranking produced by the *CycleRank* algorithm: for example, using as a reference node r = "Fake news" they appear respectively in 15th (*"United States"*), 8th (*"The New York Times"*), 147th (*"World War II"*), and 100th (*"Germany"*) position; with r = "Right to be forgotten" only *"The New York Times"* appears in 29th position; with r = "Online identity" only *"United States"* appears in 54th position; finally with r = "Internet privacy" *"United States"* and *"The New York Times"* appear respectively in 185th and 179th position.

In all the other cases these articles receive a *CycleRank* score of zero and do not appear in the rankings. In this way, *CycleRank* leaves space to articles whose content is more strongly associated with the reference topic to appear at higher positions in the ranking.

3.5.4.2 *Case Study: "Fake news"*

This section illustrates the results of longitudinal and cross-language analyses obtained with the *CycleRank* algorithm taking *"Fake news"*

as a starting point. We have performed this analysis for all the topics pertaining to the scope of the ENGINEROOM EU project, but here we present just the results for *"Fake news"* for reasons of space.

Network visualization.

Figure ?? shows a visualization of the network centered around the article "Fake news", where node size reflects the *CycleRank* score so that bigger nodes (and labels) represent concepts that are more relevant to the reference node. The reduced network, obtained as explained in Section 3.4, is used for this purpose: all the concepts which do not share any loop shorter than $K = 4$ are removed so that only concepts having a *CycleRank* score greater than 0 are included in the visualization. For readability reasons, node label is shown only for articles having a *CycleRank* score of at least 20.

Figure 6: Graph induced by the nodes with non-zero *CycleRank* score with reference node $r =$ "Fake news" and $K = 4$ on English Wikipedia. over the snapshot of March 1st, 2018. The network is visualized after applying the ForceAtlas2 algorithm. Colors represents clusters calculated with the Louvain algorithm. The dimension of the nodes and their labels depends on the *CycleRank* score; labels are only shown for *CycleRank* values of at least 20.

A spatialization algorithm is used to this aim to place the nodes in a way that minimizes the distance between nodes that are connected to each other. For this, we rely on the GEPHI software[5] [60], and on the ForceAtlas2 algorithm for placing nodes. The algorithm simulates a physical system with forces attracting and repelling nodes. A repulsive force drives nodes apart, while connections introduce an attractive force that brings nodes closer to each other [61]. In this way, the position of each node in the resulting visualization reflects its connections to the other nodes, and clusters of nodes well connected with each other emerge visually in the network.

Edges, representing hyperlinks between articles, are depicted in clockwise direction according to an established convention. Colors represent clusters of densely connected articles, identified with the Louvain method [62].

Longitudinal analysis.

Table 12 presents the *CycleRank* scores calculated over the snapshots of the wikilink graph taken on March 1st, 2017 and March 1st, 2018. We analyze only two years because the article exists with its current meaning since January 15th, 2017[6]. On one hand, we see a kind of increasing politicization of the debate around fake news, with a rising importance of topics related to the US elections won by Donald Trump. On the other hand, we observe the rise of Facebook up to the third position on 2018, indicating the rapid increase in the importance of the company, in 2017 the article about the company was ranked in 17th position.

Cross-language analysis.

Tables 13 and 14 present the *CycleRank* ranking produced by considering the article *"Fake news"* in English Wikipedia and the corresponding articles in other 7 languages available in the WikiLinkGraphs dataset.[7]

First, we point out that *CycleRank* is able to find results that are pertaining to the local Wikipedia edition, for example, results from

5 https://gephi.org/

6 Prior to that date, the article had a more general meaning which has been moved to the page *"Fake news (disambiguation)"*. The page *"Fake news"* was originally created on April 21st, 2005 and it was used initially as a redirect to *"News propaganda"* and then as a «half-article, half-disambiguation page» - as an editor noted at the time - which described the origin of the term and the then-current concurrent meaning of satirical news with reference to television programs such as *"Saturday Night Live"* and *"The Daily Show"* (https://en.wikipedia.org/w/index.php?title=Fake_news_(disambiguation)&oldid=25951928).

7 Results from Spanish Wikipedia are omitted because the article about *"Fake news"* was created on January, 2nd 2018 and only one article - besides *"Fake news"* itself - received a *CycleRank* score greater than zero.

German Wikipedia include *"Tagesschau.de,"* and *"Der Freitag"*, two local news outlets; from French Wikipedia *"Emmanuel Macron"* (France's Prime Minister), from Polish Wikipedia we find in the top-10 *"Związek Socjalistycznych Republik Radzieckich"* (URRS), *"Kryzys krymski"* (Crimea Crisis), and *"NATO"*; and the results from Russian Wikipedia include *"Vrag naroda"* ("Enemy of the people").

To compare results across languages, we have tagged related results in each table with coloured markers; the color-coding of each group of concepts mirrors the colors of the clusters calculated on the networks as shown in Figure ??:

1. (purple) groups terms related to disinformation (*"Desinformation,"* *"Propaganda,"* *"Désinformation,"*, *"Disinformazione,"* *"Dezinformacja"*), hoaxes and rumors (*"Hoax,"* *"Rumeur,"* *"Bufala"*), and clickbait (*"Clickbait,"* *"Klikbejt,"* *"Klickbete"*);
2. (green) groups terms related to news outlets and publications (*"Der Freitag,"* *"CNN,"* *"The New York Times,"* *"Izvestija,"* *"The Insider,"* etc.) and to journalism in general (*"Journalistiek"*, *"Nieuws,"* *"Tabloid,"* *"Zhjoltaja pressa,"* *"Gula pressen,"* etc.);
3. (cyan) indicates articles about *"Facebook,"* and *"Social media"*;
4. (orange) indicates articles about *"Donald Trump,"* *"United States presidential election 2016,"* and *"Donald Trump presidential campaign 2016"*;

These groups span across languages as these are common elements that characterize the context of the topic *"Fake news"* across all the cultures expressed by the languages that we have examined.

Finally, we point out how certain aspects of *"Fake news"* are especially relevant in some languages without being specifically related to the corresponding culture, such as *"Verifica dei fatti"* (fact checking), *"Debunker,"* and *"Spin doctor"* in Italian Wikipedia; *"Framing"* in Dutch Wikipedia; and *"Källkritik"* (source criticism), and *"Psykologisk krigföring"* (psychological warfare) in Swedish Wikipedia.

3.5.4.3 *Comparison of CycleRank and 2DRank*

Tables 15 and 16 present a comparison between the top-10 results with the highest scores obtained by *CycleRank*, *PageRank* and *2DRank*, computed on the WikiLinkGraphs snapshot of the English Wikipedia of March 1st, 2018 with reference nodes *"Computer science"* and *"Freddie Mercury"*, respectively.

These results highlight the limitations of *PageRank* that we have described in Section 3.2: in the top positions in the rankings produced with $\alpha = 0.85$, we see articles such as *"United States"* and *"World War*

page	Fake news		Right to be forgotten		Online identity	
#	**CycleRank**	**PageRank**	**CycleRank**	**PageRank**	**CycleRank**	**PageRank**
1	Fake news	Fake news	Right to be forgotten	Right to be forgotten	Online identity	Online identity
2	CNN	United States	Freedom of speech	The New York Times	Transgender	Social networking service
3	Facebook	The New York Times	Right to privacy	Freedom of speech	Identity (social science)	Identity (social science)
4	United States presidential election, 2016	World War II	Internet privacy	Google Spain v AEPD and Mario Costeja González	Social networking service	Reputation
5	Social media	The Washington Post	Privacy law	International human rights law	Avatar (computing)	Identity theft
6	Propaganda	The Guardian	Google	European Union	Online chat	Facebook
7	Donald Trump presidential campaign, 2016	President of the United States	General Data Protection Regulation	The Guardian	Digital identity	Google
8	The New York Times	Germany	Internet	European Commission	Online identity management	Twitter
9	Fake news website	Washington, D.C.	Censorship	Data Protection Directive	Social software	Blog
10	Pope Francis	HuffPost	Information privacy	United States	Reputation	Authentication

Table 10: Top-10 articles with the highest *CycleRank* and *PageRank* scores computed from the articles *"Fake news," "Right to be forgotten,"* and *"Online identity"* on English Wikipedia, over the most recent snapshot of the WikiLinkGraphs dataset (2018-03-01).

page	Algorithmic bias		Internet privacy		General Data Protection Regulation	
#	**CycleRank**	**PageRank**	**CycleRank**	**PageRank**	**CycleRank**	**PageRank**
1	Algorithmic bias	Algorithmic bias	Internet privacy	Internet privacy	General Data Protection Regulation	General Data Protection Regulation
2	Machine learning	European Union	Google	IP address	Data Protection Directive	Data Protection Directive
3	Artificial intelligence	Machine learning	Facebook	Firefox	Information privacy	ePrivacy Regulation (European Union)
4	Ethics of artificial intelligence	Artificial intelligence	Tor (anonymity network)	The New York Times	Right to be forgotten	European Union
5	Google	Cambridge, Massachusetts	Privacy	Social networking service	Personally identifiable information	European Commission
6	Internet of things	Database	Internet censorship	Ixquick	National data protection authority	European Parliament
7	Algorithm	Harvard University Press	HTTP cookie	Zombie cookie	Privacy	Directive (European Union)
8	Facebook	Google	Internet	Google Street View	Jan Philipp Albrecht	Regulation (European Union)
9	Cybernetics	Facebook	Proxy server	Internet	Privacy law	Council of the European Union
10	Complex system	Data Protection Directive	Computer security	Tor (anonymity network)	Privacy by design	EIDAS

Table 11: Top-10 articles with the highest *CycleRank* and *PageRank* scores computed from the articles *"Algorithmic bias," "Internet privacy,"* and *"General Data Protection Regulation"* on English Wikipedia, over the most recent snapshot of the WikiLinkGraphs dataset (2018-03-01).

#	2017	2018
1	Fake news	Fake news
2	Social media	CNN
3	Satire	Facebook
4	Fake news website	United States presidential election, 2016
5	Yellow journalism	Social media
6	Mainstream media	Propaganda
7	News satire	Donald Trump presidential campaign, 2016
8	Phishing	The New York Times
9	CNN	Fake news website
10	Donald Trump	Pope Francis

Table 12: Top-10 articles with the highest *CycleRank* score computed from the page *"Fake news"*, over two snapshot from March, 1st 2017 and March 1st, 2018. The article with its current meaning exists since January 15th, 2017.

#	de	en	fr	it
1	Fake News	Fake news	Fake news	Fake news
2	Barack Obama	② CNN	③ Donald Trump	① Disinformazione
3	② Tagesschau.de	④ Facebook	Élection présidentielle française de 2017	Post-verità
4	③ Donald Trump	③ United States presidential election, 2016	④ Facebook	① Bufala
5	① Desinformation	Social media	Ère post-vérité	Debunker
6	③ Donald Trumps Präsidentschafts-wahlkampf 2015/16	① Propaganda	Emmanuel Macron	① Manipolazione dell'informazione
7	② Der Freitag	③ Donald Trump presidential campaign, 2016	Guerre civile syrienne	Verifica dei fatti
8	③ Präsidentschaftswahl in den Vereinigten Staaten 2016	② The New York Times	① Désinformation	① Clickbait
9	Postfaktische Politik	Fake news website	① Rumeur	Spin doctor
10	Hillary Clinton	Pope Francis	② Conspiracy Watch	② Candido (rivista)

Table 13: Top-10 articles with the highest *CycleRank* score computed from the page *"Fake news"* or equivalent in the given language, over the most recent snapshot of the WikiLinkGraphs dataset (2018-03-01), for German (`de`), English (`en`), French (`fr`), and Italian (`it`) Wikipedia. Circled numbers mirror the clusters presented in Figure ??: (1, purple) terms related to disinformation; (2, green) terms related to news outlets and publications; (3, cyan) terms related to Facebook and social media (4, orange) terms related to Donald Trump and the 2016 presidential election in the United States.

#	nl	pl	ru‡	sv
1	Nepnieuws	Fake news	Fal'shivye novosti	Fejknyheter
2	④ Facebook	① Propaganda	① Klikbejt	① Klickbete
3	② Journalistiek	① Dezinformacja	② Zhjoltaja pressa	Sensationalism
4	④ Sociale media	Związek Socjalistycznych Republik Radzieckich	Piccagejt	② Gula pressen
5	③ Donald Trump	Kryzys krymski	Vrag naroda	Källkritik
6	① Desinformatie	④ Media społecznościowe	Respublikanskaja partija (SShA)	① Hoax
7	① Hoax	② Środki masowego przekazu	② Tabloid	Psykologisk krigföring
8	③ Amerikaanse presidentsverkiezingen 2016	② Dziennikarz	② CNN	Andra världskriget
9	Framing	② Informacja	② Izvestija	Google
10	② Nieuws	NATO	② The Insider	② Joseph Pulitzer

Table 14: Top-10 articles with the highest *CycleRank* score computed from the page *"Fake news"* or equivalent in the given language, over the most recent snapshot of the WikiLinkGraphs dataset (2018-03-01), for Dutch (`nl`), Polish (`pl`), Russian (`ru`), and Swedish (`sv`) Wikipedia. Circled numbers mirror the clusters presented in Figure ??: (1, purple) terms related to disinformation; (2, green) terms related to news outlets and publications; (3, cyan) terms related to Facebook and social media (4, orange) terms related to Donald Trump and the 2016 presidential election in the United States. (‡) Russian Wikipedia article titles are transliterated.

II"; these articles act as attractors for the unconstrained random walks of *PageRank* since they have a very high in-degree and have among the highest values of the *PageRank* score in the overall network, as shown in Tables 8 and 9. Indeed, they are respectively in 1st (*"United States"*) and 2nd (*"World War II"*) position in the overall *PageRank* ranking for the network. This problem is only partially mitigated by lowering the damping factor to $\alpha = 0.30$

However, there are much fewer paths that connect these articles back to the reference nodes. As a result, these articles appear in much lower positions in the ranking produced by the *CycleRank* algorithm: for example, for *"Computer science"* they appear respectively in 400th (*"United States"*), and 172th (*"World War II"*) position. In this way, *CycleRank* leaves space to articles whose content is more strongly associated with the reference topic to appear at higher positions in the ranking.

Similarly, the *PageRank* results for *"Freddie Mercury"* suffer from the same problem: *"United States"*, *"London"* (17th position in the global ranking), *"United Kingdom"* (3rd position in the global ranking) and *"BBC"* (53rd position in the global ranking) all appear in the top-10 results for the *PageRank* with $\alpha = 0.85$. This bias is only partially resolved by lowering the damping factor.

2DRank seems to mostly solve this particular issue, but still includes spurius results such as *"Charles Messina"* and *"Panchgani"*, that are only partially related to *"Freddy Mercury"*.

3.5.5 QUANTITATIVE COMPARISON

To compare our proposed approach against existing algorithms, we need a way to evaluate how good a ranking is with respect to some ground truth. In general, we cannot directly compare the ranking functions rf_r, because they may vary wildly in absolute values; furthermore, some algorithms we will compare to do not define a ranking function but just produce the final ranking.

We provide three different comparison strategies that we encapsulate in three different measures. Each measure is based on a suitable dataset that we use as a ground truth against which we evaluate the performance of each algorithm. At high level, we want to evaluate the following three facets of each ranking algorithm:

1. to what extent it is able to maintain the relative ranking of most-clicked links from a given article (*ClickStream* evaluation);
2. to what extent it is able to rank in the top positions articles highlighted by editors in the "See Also" section (*See-Also* evaluation).

Computer science

#	**CycleRank,** $K = 3$		**CycleRank,** $K = 4$		**PageRank,** $\alpha = 0.30$		**PageRank,** $\alpha = 0.85$		**2DRank,** $\alpha = 0.30$	**2DRank,** $\alpha = 0.85$
	article	**score**	**article**	**score**	**article**	**score** ($\times 10^{-4}$)	**article**	**score** ($\times 10^{-4}$)	**article**	**article**
1	List of computer scientists	13.78	List of computer scientists	175.08	Computing	12.48	Mathematics	15.57	Association for Computing Machinery	Association for Computing Machinery
1	Algorithm	6.36	Algorithm	158.78	Computational science	12.07	Computer	14.42	Programming language	Programming language
3	Artificial intelligence	5.96	Artificial intelligence	154.30	Gottfried Wilhelm Leibniz	12.00	Computing	12.71	Theoretical computer science	Edsger W. Dijkstra
4	Theoretical computer science	5.36	Mathematics	134.13	Mathematics	10.92	Association for Computing Machinery	11.77	Artificial intelligence	Artificial intelligence
5	Mathematics	4.37	Programming language	108.42	Association for Computing Machinery	10.91	Gottfried Wilhelm Leibniz	11.71	Algorithm	Machine learning
6	Programming language	4.32	Theoretical computer science	100.13	Algorithm	10.58	Computational science	11.30	Programming language theory	Algorithm
7	List of pioneers in computer science	4.08	List of pioneers in computer science	88.08	Artificial intelligence	9.97	Algorithm	11.21	Edsger W. Dijkstra	Programming language theory
8	List of important publications in computer science	4.02	Alan Turing	79.11	IBM	9.93	United States	10.23	List of computer scientists	Compiler
9	Edsger W. Dijkstra	3.82	Logic	75.90	Computational complexity theory	9.91	World War II	10.04	Data science	Theoretical computer science
10	Alan Turing	3.67	Outline of software engineering	73.35	Logic	9.84	IBM	9.60	Machine learning	List of pioneers in computer science

Table 15: Top-10 articles as ranked by *CycleRank*, *PageRank*, and *2DRank* with *"Computer science"* as reference node. The article *"Computer science"*, which would appear in the first position by definition, is omitted.

Freddie Mercury

#	**CycleRank,** $K = 3$		**CycleRank,** $K = 4$		**PageRank,** $\alpha = 0.30$		**PageRank,** $\alpha = 0.85$		**2DRank,** $\alpha = 0.30$	**2DRank,** $\alpha = 0.85$
	article	**score**	**article**	**score**	**article**	**score** ($\times 10^{-4}$)	**article**	**score** ($\times 10^{-4}$)	**article**	**article**
1	Queen (band)	13.78	Queen (band)	484.34	Queen (band)	10.79	Queen (band)	13.40	Queen (band)	Queen (band)
2	Brian May	6.41	Brian May	147.90	The Freddie Mercury Tribute Concert	8.92	London	12.07	Roger Taylor (Queen drummer)	Brian May
3	Roger Taylor (Queen drummer)	4.72	Elton John	97.48	HIV/AIDS	8.78	United States	11.51	Brian May	Roger Taylor (Queen drummer)
4	John Deacon	3.52	Roger Taylor (Queen drummer)	92.80	Queen II	8.65	Rock music	8.99	Made in Heaven	Queen II
5	Made in Heaven	3.37	Michael Jackson	73.33	Mercury Phoenix Trust	8.52	HIV/AIDS	8.99	Queen II	Made in Heaven
6	We Will Rock You (musical)	3.17	John Deacon	64.73	Middlesex	8.48	The Freddie Mercury Tribute Concert	8.73	John Deacon	John Deacon
7	Elton John	3.12	Live Aid	61.51	Parsi	8.45	Roger Taylor (Queen drummer)	8.51	We Will Rock You (musical)	Bohemian Rhapsody
8	The Freddie Mercury Tribute Concert	3.07	Bohemian Rhapsody	60.24	Zoroastrianism	8.39	United Kingdom	8.30	Bohemian Rhapsody	We Will Rock You (musical)
9	Bohemian Rhapsody	3.02	Brit Awards	53.69	Good Old-Fashioned Lover Boy	8.35	BBC	8.24	Greatest Hits (Queen album)	Greatest Hits (Queen album)
10	Greatest Hits (Queen album)	2.97	The Freddie Mercury Tribute Concert	48.50	Millennium stamp	8.33	Queen II	8.07	Panchgani	Charles Messina

Table 16: Top-10 articles as ranked by *CycleRank*, Personalized PageRank, and 2Drank with *"Freddie Mercury"* as reference node. The article *"Freddie Mercury"*, which would appear in the first position by definition, is omitted.

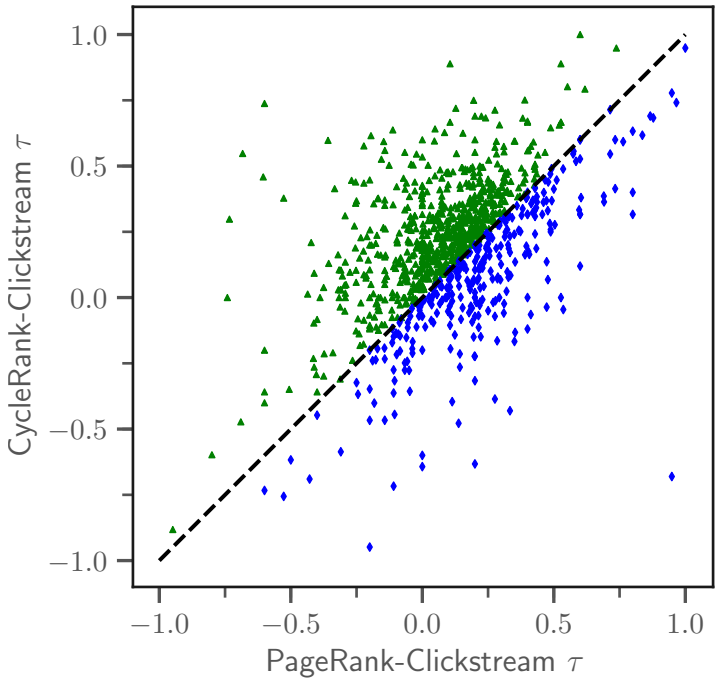


Figure 7: Comparison of the Kendall τ correlation coefficients of the *ClickStream* ranking with the rankings produced by *PageRank* (x coordinate) and *CycleRank* (y coordinate) over a sample of random articles. If for a given article $y > x$ then the correlation between the *CycleRank* and *ClickStream* rankings is higher than the correlation between the *PageRank* and *ClickStream* rankings (green triangles), if $y \leq x$ is vice-versa (blue diamonds).

3. to what extent it tends to give prominence to global "superstars", i. e. nodes which are very popular in the overall network as measured by their high indegree (*"Indegree"* evaluation).

Our evaluation measures are based on the ones used in the information retrieval literature. In particular, we follow in the footstep of Schwarzer and collaborators [53] who have also used *ClickStream* and *See-Also* for evaluating performance across a large set of topics.

In the following subsections, we describe in detail each measure, the dataset used as ground truth and the results of the experiments we performed. We also present examples to illustrate qualitatively the results.

3.5.5.1 *ClickStream Evaluation*

The idea of this measure is to test the ability of each algorithm to maintain the relative relevance of a set of topics with respect to the *ClickStream* dataset [63], which we use as a ground truth. In other words, interpreting clicks on links by Wikipedia readers as a measure

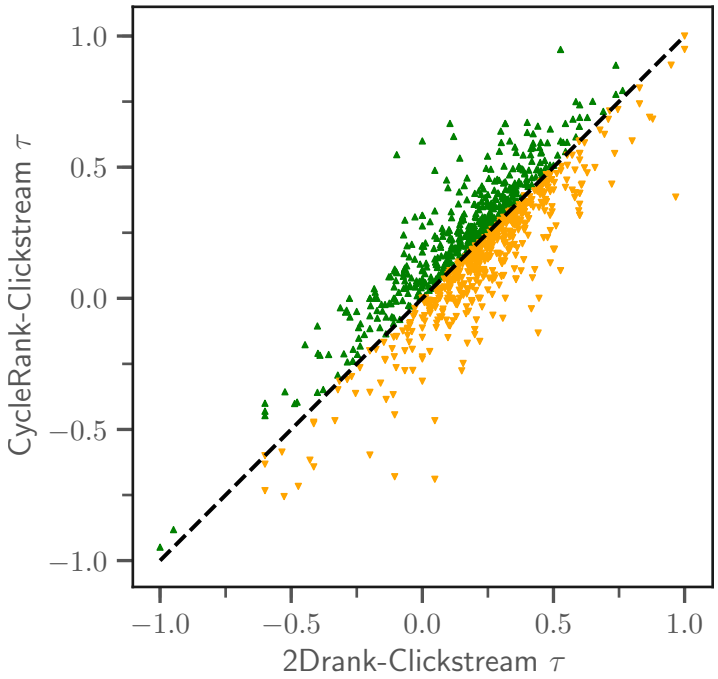


Figure 8: Comparison of the Kendall τ correlation coefficients of the *ClickStream* ranking with the rankings produced by *2DRank* (x coordinate) and *CycleRank* (y coordinate) over a sample of random articles. If for a given article $y > x$ then the correlation between the *CycleRank* and *ClickStream* rankings is higher than the correlation between the *2DRank* and *ClickStream* rankings (green triangles), if $y \leq x$ is vice-versa (blue diamonds).

of the relative importance of each link in an article, we aim to measure whether the algorithms are able to maintain this relative ranking.

We have chosen the February 2018 release [8] of the dataset because it is the closest in time to our WikiLinkGraphs snapshot. This dataset contains counts of `(source, target)` article pairs extracted from the request logs made to Wikipedia's servers over one month. This data reflects the number of times a Wikipedia visitor has reached the `target` article from the `source` article. The fact that a given `(source, target)` pair appears in the clickstream implies the existence of a link in the `source` page pointing to the `target` page; these links may appear as *wikilinks* in the article source or come from templates. The count is the cumulative number of times that users have navigated from `source` to `target` via links. Note that `(source, target)` pairs with a count of 10 or fewer observations are not present in the dataset. In this way, this data provides an aggregated view on how Wikipedia articles are reached by users and what links they click on, producing it gives a weighted network of articles, where each edge

8 https://dumps.wikimedia.org/other/clickstream/2018-02/

weight corresponds to how often people navigate from one page to another.

The *ClickStream* dataset also contains special sources to represent, for example, pages in other Wikimedia projects or external search engines;[9] we filter those out. The dataset in total comprises over 25M pairs, of which over 15.4M are links between pages.[10]

From the *ClickStream* data we can derive an ordered list of articles, which we can consider as a ranking: our evaluation strategy consists in computing Kendall's rank correlation coefficient between the *ClickStream* ranking and the ranking of the same pages produced by the algorithms under consideration. We formalize this evaluation strategy as follows: let $\mathcal{C}$ be the *ClickStream* dataset, i.e. a set of triplets: (v_s, v_t, c) where $v_s, v_t \in V$ are respectively the source and target articles and $c \in \mathbb{N}, c \geq 10$ is the count for the pair (v_s, v_t); we define $W_r \subseteq V$ as the set of nodes that appear in the *ClickStream* dataset with source r, i. e.

$$W_r = \{w \in V | (r, w, c) \in \mathcal{C}\}.$$

for some count c.

We then use the counts in the *ClickStream* dataset to define a $rf_r^{\mathcal{C}}$ over the set W_r. Given a target $w \in W_r$ if the count for the pair (r, w) is c, i. e. if $(r, w, c) \in \mathcal{C}$ then:

$$rf_r^{\mathcal{C}}(w) = c$$

The ranking function defined above produces a ranking $\nu_r^{\mathcal{C}} = [w_1, w_2, \ldots, w_q]$ of the nodes in W_r. Ties are broken at random. The ranking will be the ground truth for evaluating the performance of each algorithm for node r.

Let ν_r be a ranking of the nodes in V produced by one of the algorithms under consideration when r is the reference node. We restrict this ranking to only the pages that appear in the *ClickStream* data $\nu_r|_{W_r}$ and then we build a list of q pairs from the rankings: $[(v_1, w_1), (v_2, w_2), \ldots, (v_q, w_q)]$.

Given two pairs (v_i, w_i) and (v_j, w_j) where $i < j$, these pairs are said to be concordant if the ranks for both elements agree: i. e., if both $v_i > v_j$

9 More precisely, the dataset contains the counts of `(referer, resource)` pairs extracted from Wikipedia's webserver logs. A `referer` is an HTTP header field that identifies the webpage that linked to the resource being requested, a `resource` is the target of the request.

10 Each pair is also accompanied by a `type` field that describes the pair: possible values for type are `link`, if the referer and request are both articles and the referer links to the request; `external`, if the referer host is not `en(.m)?.wikipedia.org`; `other`: if the referer and request are both articles but the referer does not link to the request. We are interested only in the pairs of type `link`.

Computer science					
article (W_r)	c	$\nu_r^{\mathcal{C}}$	$\nu_r^{\mathbf{(CR)}}$	$\nu_r^{\mathbf{(PR)}}$	$\nu_r^{\mathbf{(2D)}}$
Computation	1371	1	56	65	77
Algorithm	876	2	2	6	5
Programming language theory	794	3	17	63	6
Computer graphics (computer science)	648	4	43	134	31
Computational complexity theory	647	5	33	9	108
Human–computer interaction	550	6	47	68	50
Computer scientist	480	7	59	20	62
Outline of computer science	452	8	204	298	173
Computer programming	451	9	62	18	160
Programming language	414	10	6	12	2
$\tau(\nu_r^{\mathcal{C}}, \nu_r)$			0.3333	−0.0222	0.2444

Table 17: *ClickStream* data for the article *"Computer science"* (c is the click count, $\nu_r^{\mathcal{C}}$ is the ranking induced by the count) and rankings produced by *CycleRank* with $K = 3$ (CR) and *PageRank* with $\alpha = 0.30$ (PR) after filtering. The Kendall correlation coefficients between *ClickStream* and the rankings produced by the algorithms presented in the table are computed only over the 10 items displayed.

and $w_i > w_j$, or analogously if $v_i < v_j$. If $v_i = v_j$ or $w_i = w_j$ two pairs are neither concordant nor discordant. Otherwise they are discordant.

The quality of the ranking ν_r, is then defined as Kendall's rank correlation coefficient:

$$\tau(\nu_r^{\mathcal{C}}, \nu_r) = \frac{\pi_+ - \pi_-}{\binom{q}{2}}$$

where π_+ and π_- are the number of concordant and discordant pairs, respectively. We say that a ranking ν_r^1 is better than a ranking ν_r^2 if its rank correlation with the *ClickStream* ranking is higher: $\tau(\nu_r^1) > \tau(\nu_r^2)$.

Table 17 presents an example of how this evaluation metric works for the article *"Computer science"*: the table shows the *ClickStream* data and the induced ranking $\nu_r^{\mathcal{C}}$, as well as the rankings produced by the *CycleRank* ($\nu_r^{(\mathrm{CR})}$), *PageRank* ($\nu_r^{(\mathrm{PR})}$), and *2DRank* ($\nu_r^{(\mathrm{2D})}$) algorithms over the same articles. Regardless of the absolute position of these articles, we measure how these rankings agree with the one given by the *ClickStream* data, a negative value means that the ranking is discordant with the *ClickStream* ranking.

Figures 7 and 8 present the results of the *ClickStream* evaluation over a sample of 1,000 random articles. We have built this sample by se-

lecting random Wikipedia articles that have at least 5 entries in the *ClickStream* data, i. e. they have at least 5 links to other Wikipedia articles. In the figures, the (x, y) coordinates of each point are the values of Kendall's rank correlation coefficient with *ClickStream* data for *PageRank* and *CycleRank* (Figure 7); and for *2DRank* and *CycleRank* (Figure 8). Thus, if points have their y coordinate greater than the x coordinate, i.e. they are above the dashed axis $Y = x$ in the figure, it means that *CycleRank* is outperforming the other approach for that article. This is the case for *CycleRank* and Personalized Pagerank where 68.8% of articles in Figure 7 are above the axis, as well for *CycleRank* and *2DRank* where 50.2% of articles in Figure 7 are above the axis.

algorithm		**CycleRank**	
	parameters	$K = 3$	$K = 4$
PageRank	$\alpha = 0.30$	30.2/68.8	37.6/59.0
	$\alpha = 0.85$	33.4/65.5	36.2/59.7
2Drank	$\alpha = 0.30$	48.2/50.2	58.4/37.4
	$\alpha = 0.85$	48.2/50.4	56.9/37.7

Table 18: Results of the *ClickStream* evaluation over a sample of 1,000 random articles for *CycleRank*, *PageRank*, and *2DRank* for different values of their parameters: maximum cycle length K for *CycleRank* and damping factor α for *PageRank* and *2DRank*.

Table 18 presents the results of the *ClickStream* evaluation for various values of the parameters.

3.5.5.2 *See-Also Evaluation*

We measure the ability of an algorithm to identify relevant articles by using links in the *See-Also* section of a Wikipedia article as a ground truth. Following Wikipedia policies [64], the section *See-Also* contains a list of internal links to related Wikipedia articles. These lists may be ordered logically, chronologically or alphabetically, and there is no guarantee that the same criterion is used across multiple pages. For this reason, we treat these lists as unordered, that is we do not treat the pages listed in these sections as being ranked by relevance, but just as a set of related pages.

More formally, let $W_r \subseteq V$ be a set of ground-truth nodes that are relevant with respect to a reference node r, and let ν_r be a ranking of the nodes in V. The quality of the ranking ν_r is defined as

$$\xi(\nu_r, W_r) = \sum_{w \in W} \xi(\nu_r, w)$$

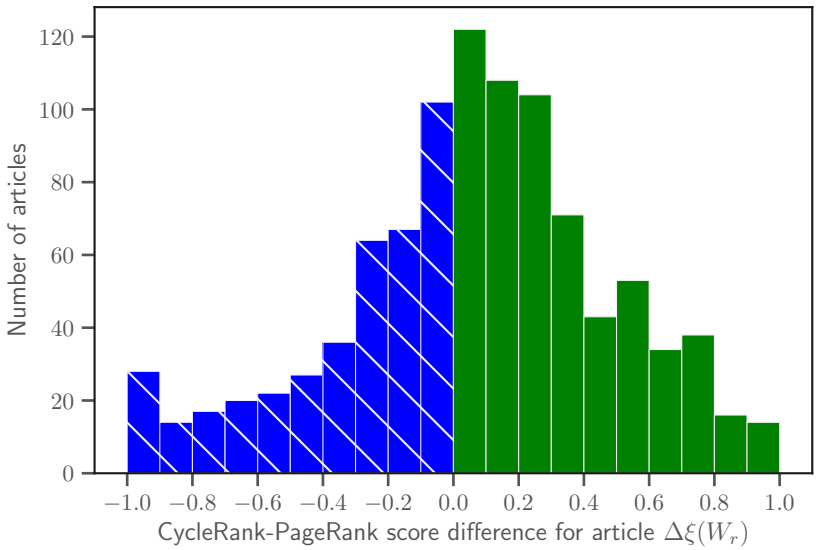


Figure 9: Distribution of $\Delta\xi(W_r)$ between *CycleRank* and *PageRank*. When values are positive (solid green bars) *CycleRank* is able to find *See-Also* articles in a higher position than Personalized Pagerank for a given article; when values are negative (blue bars with with white hatch) is vice-versa.

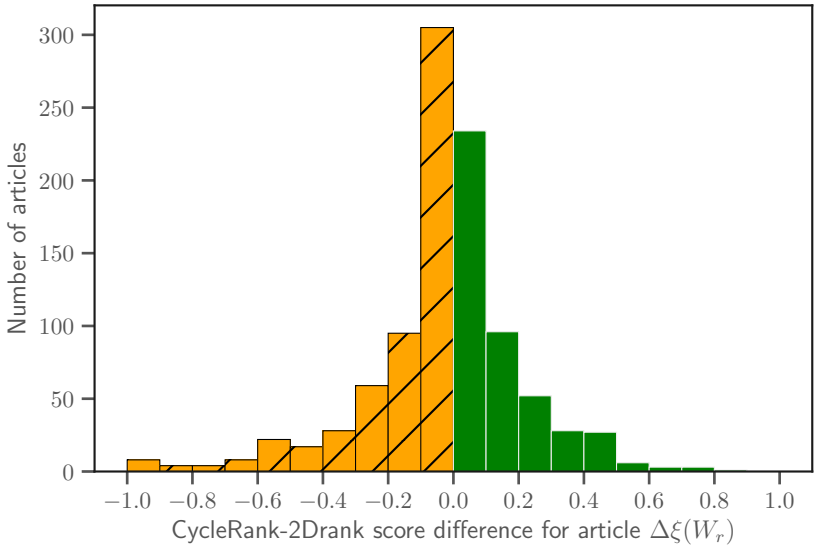


Figure 10: Distribution of $\Delta\xi(W_r)$ between *CycleRank* and *2DRank*. When values are positive (solid green bars), *CycleRank* is able to find *See-Also* articles in a higher position than *2DRank* for a given article; when values are negative (orange bars with with black hatch) is vice-versa.

where

$$\xi(\nu_r, w) = \frac{1}{i} \quad \text{if } w = v_i \in \nu_r \wedge w \in W$$

We say that a ranking ν_r^1 is better than a ranking ν_r^2 if $\xi(\nu_r^1) > \xi(\nu_r^2)$.

Computer science					
article (W_r)	$\nu_r^{(\mathbf{CR})}$	$\nu_r^{(\mathbf{PR})}$	$\Delta\xi$ **CR-PR** ($\times 10^{-4}$)	$\nu_r^{(\mathbf{2D})}$	$\Delta\xi$ **CR-2D** ($\times 10^{-4}$)
Academic genealogy of computer scientists	13	220	723.8	26	384.6
Association for Computing Machinery	16	6	−1041.7	2	−4375.0
Computer Science Teachers Association	207	231	5.0	402	23.4
Engineering informatics	447	228	−21.5	399	−2.7
Informatics	70	106	48.5	87	27.9
List of academic computer science departments	74	232	92.0	7862	133.9
List of computer scientists	2	110	4909.1	9	3888.9
List of important publications in computer science	9	167	1051.2	18	555.6
List of pioneers in computer science	8	16	625.0	538	1231.4
List of unsolved problems in computer science	92	148	41.1	41	−135.2
Outline of software engineering	12	217	787.3	25	433.3
Technology transfer in computer science	206	223	3.7	380	22.2
Turing Award	14	49	510.2	17	12.6
$\sum \Delta\xi(W_r)$			7733.8		2314.4

Table 19: The first 10 articles appearing in the *See-Also* section of the *"Computer science"* article. We use them to compare *CycleRank* with $K = 3$ (CR), *PageRank* with $\alpha = 0.30$ (PR), and *2DRank* with $\alpha = 0.30$ (2D). For each article, the table reports the position in which it appears in the ranking produced by each algorithm, and the corresponding difference in scores $\Delta\xi$. The $\sum \Delta\xi(W_r)$ is calculated only over the 10 items displayed.

Table 19 presents, as an example, the results of the *See-Also* evaluation for the article *"Computer Science"* in English Wikipedia: the first column lists the name of the articles appearing in the *See-Also* section of the article *"Computer science"* W_r; the second, third and fifth columns show the position of each article in the ranking produced by *CycleRank*, *PageRank*, and *2DRank* respectively; the fourth and sixth columns show the difference in evaluation score for each article between *CycleRank* and *PageRank* $\Delta\xi(\mathrm{CR} - \mathrm{PR})$, and *CycleRank* and *2DRank* $\Delta\xi(\mathrm{CR} - \mathrm{2D})$. When *CycleRank* ranked a page in a higher position than the other approach this difference is positive, otherwise it is negative.

Figures 9 and 10 show the distribution of the differences in evaluation score over the same sample of 1,000 articles used above. Figure 9 compares *CycleRank* and *PageRank*, i. e. the plot is the distribution

$\Delta\xi(\text{CR} - \text{PR})$, while Figure 10 presents analogous results for *CycleRank* and *2DRank*, $\Delta\xi(\text{CR} - \text{PR})$.

algorithm	**parameters**	$\xi(\nu_r, W_r)$
CycleRank	$K = 3$	578.4
	$K = 4$	536.3
PageRank	$\alpha = 0.30$	515.7
	$\alpha = 0.85$	467.6
2Drank	$\alpha = 0.30$	600.9
	$\alpha = 0.85$	484.3

Table 20: Results of the *See-Also* evaluation over a sample of 1,000 random articles for *CycleRank*, *PageRank* and *2DRank* for different values of their parameters: maximum cycle length K for *CycleRank* and damping factor α for *PageRank* and *2DRank*.

Table 20 presents the results of the *See-Also* evaluation over a sample of 1,000 random articles. We have built this sample with the following characteristics: we selected Wikipedia articles that have at least 3 links to other existing Wikipedia articles[11], in this way we ensure that the pages used in the sample have enough links. As the table shows, the best performance is achieved by *2DRank* with $\alpha = 0.30$, while the second-best algorithm is *CycleRank* with $K = 3$. This result can be justified by considering that both algorithms take into account both incoming and outgoing links for every page. Even with $K = 4$ *CycleRank* outperforms *PageRank* with $\alpha = 0.30$, and $\alpha = 0.85$ and *2DRank* with $\alpha = 0.85$; however lowering the value of K ensures that results are closer to the reference node. This behavior is similar to what is seen in *PageRank* and *2DRank* where smaller value of the damping factor α give a higher score to nodes that are closer to the reference node.

3.5.5.3 *Indegree Evaluation*

We measure the extent to which an algorithm tends to give prominence to global "superstars", i. e. nodes which are very popular in the overall network as measured by their high indegree. In this section we use the same measure ξ that we have used in the previous *See-Also* evaluation, with two modifications: first, we use the top-100 articles by indegree as test set W_r and in this case lower is better: a ranking ν_r^1 is better than a ranking ν_r^2 if $\xi(\nu_r^1) < \xi(\nu_r^2)$. For fairness, we cut-off all rankings produced at 1,000 results.

11 Even if it is discouraged by Wikipedia policies, in principle a Wikipedia editor could insert in the *See-Also* section a link to a non-existing article.

Computer science

ν^i	***CycleRank*, $K = 3$** article	$\nu_r^{(\mathbf{CR})}$	$\xi(\nu_r^{(\mathbf{CR})}, w)$ $(\times 10^{-4})$
5	World War II	361	27.70
$\sum \xi(\nu_r^{(\mathrm{CR})}, W_r)$			27.70
	PageRank, $\alpha = 0.30$ article	$\nu_r^{(\mathbf{PR})}$	$\xi(\nu_r^{(\mathbf{PR})}, w)$ $(\times 10^{-4})$
1	United States	258	38.76
4	India	678	14.75
5	World War II	24	416.67
7	Germany	451	22.17
10	New York City	357	28.01
11	United Kingdom	265	37.74
12	England	322	31.06
13	London	458	21.37
14	Australia	715	13.99
17	Italy	519	19.27
$\sum \xi(\nu_r^{(\mathrm{PR})}, W_r)$			643.77

Table 21: Positions in which the top-100 articles by indegree appear in the rankings produced by *CycleRank* (top), *PageRank* (bottom), with *"Computer science"* as reference node and their score $\xi(\nu_r, w)$. Results for *PageRank* and *2DRank* are limited to the top-1000 positions. The $\sum \xi(W_r)$ is calculated only over the 10 items displayed.

Table 21 presents the position in which the top-100 articles by indegree appear in the top-1000 positions of the rankings produced *PageRank* (top), *2DRank* (middle), and *CycleRank* (bottom) with *"Computer science"* as reference node.

Figure 11 shows the results of the indegree evaluation for *PageRank*, *2DRank*, and *CycleRank*. We see that *CycleRank* is able to obtain a lower score meaning that it includes fewer pages with high indegree in high position in the rankings it produces.

3.5.6 PERFORMANCE ANALYSIS

Finally, Table 22 evaluates the performance of *CycleRank* with respect to the alternative approaches. Times are computed by averaging over the sample of 1,000 articles used in the *See-Also* evaluation. Each job was executed on HPC cluster nodes equipped with 48 Xeon 5118 processors and 32 GB of RAM, using only one core and one processor.

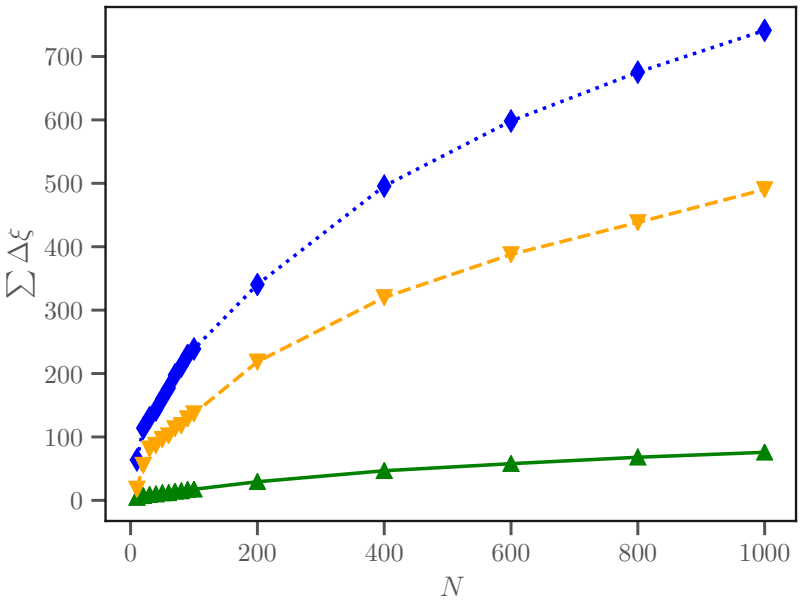


Figure 11: Evaluation scores for *PageRank* (dotted blue line with diamond markers), *2DRank* (dashed orange line with triangle-down markers), and *CycleRank* (straight green line with triangle-up markers) taking the top-N articles by indegree. A lower score is better.

Results presented in Table 22 do not take into account the time needed to read the input graph, which required 60 seconds on average (a value much larger than the time need by *CycleRank* to execute). Execution times for *2DRank* are not obtained directly, but by summing the execution times of *PageRank* and *CheiRank*.

Computing *CycleRank* is two orders of magnitude faster than computing *PageRank* and *CheiRank*. Since our proposed approach is based on enumerating all cycles going through the reference node, we note that in the worst case–a complete graph–the computational complexity increases exponentially with cycle length, making its computation challenging for higher values of K in very dense graphs. However, we have shown that *CycleRank* can produce good results even for small values of K and has a significant time advantage with respect to *PageRank* and *2DRank*.

3.6 CONCLUSIONS

This Chapter introduced *CycleRank*, a novel algorithm based on cyclic paths that can be used to find the most relevant nodes in the Wikipedia link network related to a topic. Given a reference node in a directed graph, the algorithm finds all simple cycles that go through the reference node and assigns a score to each node that belong to these cycles.

algorithm	parameter	time (s)
CycleRank	$K = 3$	4 ± 1
	$K = 4$	10 ± 16
PageRank	$\alpha = 0.30$	260 ± 13
	$\alpha = 0.85$	928 ± 80
CheiRank	$\alpha = 0.30$	258 ± 20
	$\alpha = 0.85$	374 ± 50
2DRank	$\alpha = 0.30$	> 518
	$\alpha = 0.85$	> 1302

Table 22: Execution time comparison of *CycleRank*, *PageRank*, *CheiRank*, and *2DRank* for different values of the parameters: maximum cycle length K for *CycleRank*, and damping parameter α for *PageRank*, *CheiRank*, and *2DRank*. All times are expressed in seconds.

The algorithm possesses one parameter which is the maximum cycle length that we are interested to find.

An extensive comparison between *CycleRank*, *PageRank*, and its variant *2DRank* has been performed, based on three quantitative measues. First, we have used the *ClickStream* dataset to show that the rankings produced by *CycleRank* align better with readers' behaviour. Second, we have used the *See-Also* links that appear in Wikipedia articles to show that *CycleRank* is able to rank related articles in high position; only 2DRank with $\alpha = 0.30$ obtains a better result for this test. Finally, we have shown that *CycleRank* is more robust to pages with high indegree. Furthermore, we have shown that our algorithm is faster than the alternatives, offering order-of-magnitude speed-ups with respect to library implementation of *Personalized PageRank*.

In other words, *CycleRank* is a viable alternative to *PageRank*, especially in the case of graphs where the role of inlinks and outlinks is comparable; we believe that it can be easily applied to similar contexts, such as knowledge bases.

The version of the algorithm presented in this work is based on a simple exponentially-decaying scoring function; while we have empirically validated the results on our WikiLinkGraphs dataset for other types of scoring functions, we have not presented these results here for reason of space. However, many variations and extensions could be explored.

We have assumed that the starting point for the algorithm is a single reference node. However, as in the case of *Personalized PageRank*, it would be possible to take a group of articles as seed. Then, all cycles around each of the seed nodes could be considered. Alternatively, in-

stead of counting cycles, one could count all paths from any node in the seed to any other node in the seed.

Another possible variant would be to specify two different nodes (or groups of nodes) as source and target, and to consider all paths from the source to the target within K steps. In this way, the measure would not only represent the relevance of other nodes with respect to a given reference node, but the (directed) relationship between two nodes or groups of nodes. This would help to answer questions such as: "Which are the most relevant concepts connecting Artificial Intelligence and Human rights, and which are the most relevant concepts on the other way round"?

We think that *CycleRank* provides a foundation that could be further explored to provide a family of algorithms adapted for different graphs and use cases. The suitability of different solutions could also be studied focusing on the structural properties of the network, such as its link density or clustering coefficient.

Part III

APPLICATIONS

We explore applications of *CycleRank*. First, we describe the ENGINEROOM EU project whose objective was identifying and evaluating the key enabling technologies and topics that will underpin the Next Generation Internet. The idea for *CycleRank* emerged from this project when we faced the challenge to find which were the most relevant Wikipedia pages connected to the Wikipedia article about the *Internet*. Finally, we present another contribution using Wikipedia article daily accesses to estimate how the information about diseases propagates.

4

NEXT GENERATION INTERNET - ENGINEROOM

In this chapter, we present our contribution to the ENGINEROOM EU project: this project focused on identifying and evaluating the key enabling technologies and topics characterizing the *Next Generation Internet*. The goal of this project was to identify early signals of new trends and technologies related to the internet and map the ecosystems and networks surrounding these key topics, evaluating their social, legal, technological, ethical and economic contexts. The project revolved around "umbrella topics" that will be explored in the following.

Our contribution to the project was developed during our stay at Eurecat - Centre Tecnòlogic de Catalunya in Barcelona (Spain) under the supervision of Dr. David Laniado. The project was developed by a consortium of three partners comprising Nesta (London, United Kingdom), Eurecat, and DELab at the University of Warsaw (Warsaw, Poland) and it was supported by the European Union's Horizon 2020 research and innovation programme under the EU ENGINEROOM project, with Grant Agreement nº 780643.

In this chapter, we analyze NGI related issues by shifting our focus from individual concepts to the connections between them. This allows us to describe and determine the characteristics of the representations of different topics through their connections to one another and through their context, that can mutate over time and across languages.

To reach this goal, we relied on the network of internal links, or *wikilinks*, between Wikipedia articles. The visual representation of such connections may be seen as a giant concept network emerging from the links established by Wikipedia's user community. Such a graph is not static, rather it is continuously growing and evolving, reflecting the endless collaborative processes behind it.

In order to study the framing of topics such as *"Online privacy"*, *"Online identity,"* or *"Right to be forgotten"* over time and across languages, we looked at their contexts as they emerge from the Wikipedia link networks. We aimed to achieve a more fine-grained and accurate result, following a more robust approach and building networks around specific topics. We selected 12 topics representing emerging social issues related to the *Next Generation Internet*, covering the umbrella topics, and developed a novel methodology to determine the specific charac-

teristics of each of them based on the observation of the most relevant concepts associated with them over time and across different languages. Our study was multilingual, we have used data from 9 Wikipedia editions: German (`de`), English (`en`), Spanish (`es`), French (`fr`), Italian (`it`), Dutch (`nl`), Polish (`pl`), Russian (`ru`), and Swedish (`sv`)

Equipped with the complete graphs of internal links connecting Wikipedia articles, we can explore and analyze parts of the network around specific topics to characterize their definition as emerging from the collaborative process. For each of the 12 Wikipedia articles corresponding to the topics selected for this analysis we studied the network of related concepts in Wikipedia as a mind map and used the *CycleRank* algorithm to study their context as defined by the most relevant keywords. We studied how this context varies over time looking at yearly snapshots of the network and across cultures comparing the networks derived from the nine largest editions of Wikipedia which also correspond to nine major European languages.

This chapter is organized as follows: in the next section, we describe the mapping between the 12 umbrella topics defined by the ENGINEROOM EU project and Wikipedia articles. Then we focus more in-depth on one of the topics: *"Internet governance,"* we present a visualization of the network around the corresponding Wikipedia article and a longitudinal and a cross-language analysis of its context. Additional results are presented in Appendix A. Finally, in the last section, we draw the conclusions of our work.

4.1 KEYWORD SELECTION

We started from the 10 umbrella topics defined in the ENGINEROOM EU project and selected one Wikipedia article related to each topic. The criteria for choosing said articles were:

- the size of the articles, as they had to be reasonably large (at least in the English Wikipedia);
- the lifespan of the articles, possibly existing since some years and in several of the nine selected languages;
- their structure, representing a well-defined and broad enough concept;
- their relevance for the considered topic as they had to present emerging issues relevant for our analysis.

In addition to the topics defined previously in the project, we include two further ones that we deemed of particular relevance to the NGI

Topic	Wikipedia article
Ethical AI/ML	Algorithmic bias
Cybersecurity	Computer security
Cyber-Mobbing	Cyberbullying
Fake News	Fake news
GDPR	General Data Protection Regulation
Sustainability	Green computing
Decentralisation	Internet governance
Opt out	Internet privacy
Accessibility	Net neutrality
Identity/trust	Online identity
Data Sovereignty	Open-source model
Right to be forgotten	Right to be forgotten

Table 23: Selection of reference keywords/Wikipedia articles for the analysis, connected to the umbrella topics.

initiative and the European Commission: *"GDPR"* and *"Right to be forgotten"*.

Table 23 shows the correspondence between topics and Wikipedia articles in the English Wikipedia.

4.2 CROSS-LANGUAGE KEYWORD MAPPING

Table 24 shows the correspondence between keywords and Wikipedia articles for each of the nine selected language editions that we considered. Correspondences were retrieved using *interwiki links* created by the community and connecting the same concepts across languages[1], for these reason for languages other than English the correspondences are not always complete. To overcome this, Hence, additional support was provided by members of the language communities as they were asked to revise the mapping with respect to the context of their Wikipedia language edition and to find corresponding articles when interwiki links were missing. Also, we have chosen the articles that best reflected the meaning and content of the English articles and for this reason, we have excluded articles from other languages that had a very different angle

1 Interwiki links are links to pages of Wikipedia in a different language, *https://en.wikipedia.org/wiki/Interwiki_links*.

de	en	es	fr	it	nl	pl	ru	sv
	Algorithmic bias	Sesgo algorítmico	Biais algorithmique					
Informationssicherheit	Computer security	Seguridad informática	Sécurité des systèmes d'information	Sicurezza informatica	Computerbeveiliging	Bezpieczeństwo teleinformatyczne	Komp'juternaja bezopasnost'	Datasäkerhet
Cyber-Mobbing	Cyberbullying	Ciberacoso	Cyberharcèlement	Cyberbullismo	Cyberpesten	Cyberprzemoc	Internet-travlja	Nätmobbning
Fake News	Fake news	Fake news	Fake news	Fake news	Nepnieuws	Fake news	Fal'shivye novosti	Fejknyheter
Datenschutz-Grundverordnung	General Data Protection Regulation	Reglamento General de Protección de Datos	Règlement général sur la protection des données	Regolamento generale sulla protezione dei dati	Algemene verordening gegevensbescherming	Ogólne rozporządzenie o ochronie danych	Obshhij reglament po zashhite dannyh	Dataskyddsförordningen
Green IT	Green computing	Green computing	Informatique durable	Green computing				Grön IT
Internet Governance	Internet governance	Gobernanza de Internet	Gouvernance d'Internet	Internet_Governance		Zarządzanie Internetem	Upravlenie Internetom	
Datenschutz im Internet	Internet privacy	Privacidad en Internet	Vie privée et informatique	Privacy			Privatnost' v Internete	
Netzneutralität	Net neutrality	Neutralidad de red	Neutralité du réseau	Neutralità della rete	Netneutraliteit	Neutralność sieci	Setevoj nejtralitet	Nätneutralitet
	Online identity	Identidad en línea	Identité numérique	Identità virtuale	Identiteit (eigenheid)	Tożsamość internetowa	Setevaja identichnost'	Användarnamn
Open Source	Open-source model	Código abierto	Open source	Open Source	Open source	Otwarte oprogramowanie		Öppen källkod
Recht auf Vergessenwerden	Right to be forgotten	Derecho al olvido	Droit á l'oubli	Diritto all'oblio	Recht om vergeten te worden	Prawo do bycia zapomnianym	Pravo na zabvenie	Rätten att bli glömd

Table 24: Selection of the reference Wikipedia articles for the analysis, connected to the umbrella topics.

about a subject because this would have introduced spurious results in the analysis.

The article *"Algorithmic bias"* exists only for Spanish and French, beyond English and, at the time of data collection (March 1st, 2018), the article was still underdeveloped and poorly linked. So, there is no cross-language analysis provided for this article, but it was still kept in this list as it was deemed a relevant emerging topic and the analysis was performed for English.

4.3 NETWORK VISUALIZATION

For each keyword, we show a visualization of the network obtained through on the *CycleRank* results. We have selected all the articles with a non-zero *CycleRank* score running with $K = 4$. A spatialization algorithm is used to place the nodes in a way that minimizes the distance between nodes that are connected. For this, we rely on the Gephi software[2] [60], and on the ForceAtlas2 algorithm for placing nodes. The algorithm simulates a physical system with forces attracting and repelling nodes. A repulsive force drives nodes apart, while connections introduce an attractive force that brings nodes closer to each other [61]. In this way, the position of each node in the resulting visualization reflects its connections to the other nodes, and clusters of nodes well connected with each other emerge visually in the network.

The convention for the visualizations is consistent for all networks:

- *Nodes represent articles*, with node size proportional to the *CycleRank* score. To improve readability, in the static visualizations node labels are also represented with size proportional to the *CycleRank* score, and labels are not shown for low values of the score.
- *Edges represent hyperlinks between articles.* The network is directed, and each edge represents a hyperlink in one of the two possible directions; according to an established convention, edges are represented in the clockwise direction. So, in case of a reciprocal connection, there will be a loop between two articles.
- *Colors represent clusters of densely connected articles*, identified with the Louvain method (Emmon et al, 2006). We expect these groups to represent thematic areas spontaneously emerging from the hyperlink network.

2 *https://gephi.org/*

Edges, representing hyperlinks between articles, are depicted in the clockwise direction according to an established convention. Colors represent clusters of densely connected articles, identified with the Louvain method [62].

An interactive version of these network visualizations has also been produced and are available on GitHub.[3] They can be explored zooming in and out, selecting a node to see only the network of its connections (ego-network), or selecting a given group of nodes. They were created with `Sigma.js`, and the Sigma.js export plugin for Gephi.[4]

4.4 INTERNET GOVERNANCE

In this section, we present the results obtained for the keyword *"Internet governance"*. Figure 12 represents the network of articles around the article *"Internet governance"* with *CycleRank* score greater than zero. We see that this network is characterized by 6 clusters:

1. (purple) is a cluster of concepts related digital rights: *"Internet,"* *"Digital divide,"* *"Lawrence Lessig,"* *"Internet Censorship,"* *"Electronic Frontier Foundation"*;
2. (green) groups organizations, institutions and people involved in Internet governance: *"Internet Society,"* *"ICANN,"* *"Vint Cerf,"* *"Internet Standard"*;
3. (cyan) groups organizations and technologies related to the Internet as a telecommunication network: *"Internet Engineering Task Force,"* *"IPv4,"* *"IPv6,"* *"Anycast"*;
4. (orange) is a cluster of concepts related to the internet from a historical, scientific and academic point of view: *"History of the Internet,"* *"ARPANET,"* *"DARPA,"* *"Association for Computing Machinery"*;
5. (dark green) groups technologies and protocols: *"Internet protocol suite,"* *"Port (computer networking),"* *"File Transfer Protocol,"* *"Internet Relay Chat"*;
6. (magenta) is a cluster of concepts related to domain names and their management: *"Internet Assigned Numbers Authority,"* *"Domain Name System,"* *"Domain name"*.

3 *https://github.com/NGI4eu/engineroom-networks*.
4 *https://blogs.oii.ox.ac.uk/vis/2012/11/15/build-your-own-interactive-network/*.

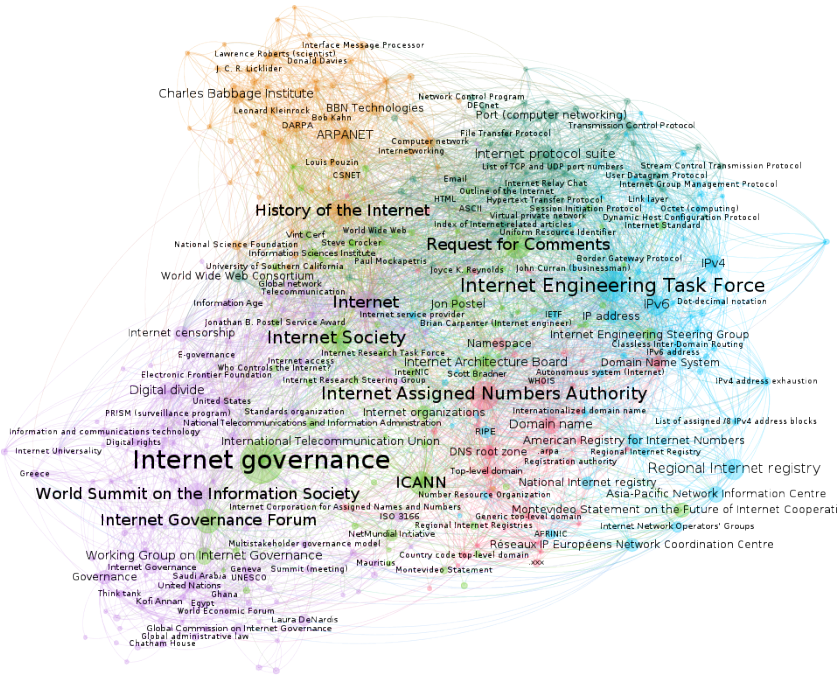


Figure 12: Graph induced by the nodes with non-zero *CycleRank* score with reference node $r =$ *"Internet governance"* and $K = 4$ on English Wikipedia, over the snapshot of March 1st, 2018. Colors represents clusters calculated with the Louvain algorithm. The dimension of the nodes and their labels depends on the *CycleRank* score.

4.4.1 LONGITUDINAL ANALYSIS

Tables 25 and 26 present the results for the longitudinal analysis for the article *"Internet governance"* in English Wikipedia over the years 2007-2018. Contrary to the case of *"Fake news"* this article did not see particular changes in its context over the years and the results as a whole reflect the growth of Wikipedia in general. The context of *"Internet governance"* contains some expected articles like: *"Internet Society"*, *"Internet"*, *"ARPANET,"* and *"ICANN."* These pages appear through the whole life o the article *"Internet governance"*, which was created on November 13th, 2005.[5]

4.4.2 CROSS-LANGUAGE ANALYSIS

Table 27 presents the results of the cross-language analysis for the article *"Internet governance"*. Corresponding articles are present in 3 other language editions besides English: German, French, and Russian. Across languages we find in the top-10 results organizations and insti-

5 https://en.wikipedia.org/w/index.php?title=Internet_governance&oldid=28236513.

tutions related to Internet governance: *"Weltgipfel zur Informationsgesellschaft"* (*"World summit on Information Society"*), *"Internet Governance Forum,"* and *"Internet Corporation for Assigned Names and Numbers"* (*"ICANN"*) in German Wikipedia; *"Sommet mondial sur la société de l'information"* (*"World summit on Information Society"*) in French Wikipedia; and *"Obshhestvo Interneta"* (*"Internet Society"*), *"Sovet po arhitekture Interneta"* (*"Internet Architecture Board"*), *"Konsorcium Vsemirnoj pautiny"* (*"World Wide Web Consortium"*), and *"ICANN"* in Russian Wikipedia. All in all, the framing of Internet governance results to be quite stable, both over time and across language editions.

4.5 CONCLUSIONS

In this chapter, we have presented a general approach to characterize a topic by inspecting its connections to other topics. For this purpose, we leveraged the network of links in Wikipedia, which we used a giant concept map to study the framing of different topics by looking at their context in the network. We created the WikiLinkGraphs dataset, representing the networks of articles for the top 9 language editions of Wikipedia, and we developed a novel computational method, *CycleRank*, to identify the most relevant concepts with respect to a given topic (reference node), as the ones laying more often on circular paths including the reference node.

Once we created the dataset and defined the methodology, we could dive into the analysis of 12 topics representing key social issues related to the *Next Generation Internet*. For each topic, we presented a visualization of the network around it, induced by the *CycleRank* results, and the ranking of the top-10 articles having the highest *CycleRank* scores, comparing results along years and across languages.

The results obtained for the other umbrella topics are available in Appendix A. All the results, including the whole ranking of the *CycleRank* scores obtained for all nodes, and the networks files and visualizations in static and interactive version are also available on GitHub.[6]

We believe that the value of this work is manifold:

- The analysis of the 12 topics presented offers an in-depth inspection of the framing of major NGI-related social issues, and of their variations over time and across cultures. Apart from the textual description in which we have highlighted some observations, further insights can be extracted by looking at the longitudinal and

6 *https://github.com/NGI4eu/engineroom-networks*.

rank	2007	2008	2009	2010	2011	2012
1	Internet governance	Internet governance	Internet governance	Internet governance	Internet governance	Internet governance
2	World Summit on the Information Society	Internet Society	Internet Society	Internet Society	Internet Society	Internet Society
3	ICANN	World Summit on the Information Society	World Summit on the Information Society	Internet	Internet	Internet
4	Working Group on Internet Governance	Internet	Internet Governance Forum	World Summit on the Information Society	History of the Internet	History of the Internet
5	Internet Governance Forum	Internet Governance Forum	Internet	History of the Internet	Charles Babbage Institute	Charles Babbage Institute
6	International Telecommunication Union	Working Group on Internet Governance	Working Group on Internet Governance	Internet Governance Forum	World Summit on the Information Society	World Summit on the Information Society
7	American Registry for Internet Numbers	ICANN	ICANN	Working Group on Internet Governance	Internet Governance Forum	Internet Governance Forum
8	Country code top-level domain	ARPANET	Scott Bradner	Regional Internet registry	Regional Internet registry	Regional Internet registry
9	Geneva	International Telecommunication Union	ARPANET	ICANN	ARPANET	ARPANET
10	History of the Internet	History of the Internet	History of the Internet	Jon Postel	ICANN	Working Group on Internet Governance

Table 25: Top 10 most relevant concepts by *CycleRank* score with respect to *"Internet governance,"* along different yearly snapshots of the WikiLinkGraphs dataset (2007-2012).

rank	2013	2014	2015	2016	2017	2018
1	Internet governance	Internet governance	Internet governance	Internet governance	Internet governance	Internet governance
2	Internet Society	Internet Society	Internet Engineering Task Force	Internet Engineering Task Force	Internet Engineering Task Force	Internet Engineering Task Force
3	Internet	Internet Governance Forum	Internet Society	Internet Assigned Numbers Authority	Internet Assigned Numbers Authority	Internet Assigned Numbers Authority
4	History of the Internet	Internet	Internet	Internet Society	Internet Society	Internet Society
5	Internet Governance Forum	World Summit on the Information Society	History of the Internet	Internet Governance Forum	Internet Governance Forum	Request for Comments
6	World Summit on the Information Society	History of the Internet	Internet Governance Forum	ICANN	Internet	Internet
7	Charles Babbage Institute	Regional Internet registry	ICANN	Internet	ICANN	Internet Governance Forum
8	Regional Internet registry	ICANN	World Summit on the Information Society	History of the Internet	History of the Internet	History of the Internet
9	ICANN	Charles Babbage Institute	Regional Internet registry	World Summit on the Information Society	World Summit on the Information Society	ICANN
10	Governance	Working Group on Internet Governance	Charles Babbage Institute	Regional Internet registry	Regional Internet registry	World Summit on the Information Society

Table 26: Top 10 most relevant concepts by *CycleRank* score with respect to *"Internet governance,"* along different yearly snapshots of the WikiLinkGraphs dataset (2013-2018).

rank	de	en	fr	ru
1	Internet governance	Internet governance	Gouvernance d'Internet	Upravlenie Internetom
2	Weltgipfel zur Informationsgesellschaft	Internet Engineering Task Force	Internet	Internet
3	Working Group on Internet Governance	Internet Assigned Numbers Authority	Révolution numérique	Obshhestvo Interneta
4	Internet Governance Forum	Internet Society	Technologies de l'information et de la communication	Internet-provajder
5	Wolfgang Kleinwächter	Request for Comments	Métadonné	Setevoj nejtralitet
6	Internet Corporation for Assigned Names and Numbers	Internet	Sommet mondial sur la société de l'information	Internet-cenzura
7	Informationsgesellschaft	Internet Governance Forum	Technique	Istorija Interneta
8	Internet	History of the Internet	Serveur informatique	Sovet po arhitekture Interneta
9	Jugendmedienschutz	ICANN	Innovation	Konsorcium Vsemirnoj pautiny
10	Cyberkrieg	World Summit on the Information Society	Économie du savoir	ICANN

Table 27: Top 10 most relevant concepts by *CycleRank* score with respect to *"Internet governance,"* across different Wikipedia language editions.

cross-cultural comparison of the top-most relevant concepts with respect to each topic.

- The network visualizations produced, especially in their interactive version available on GitHub, can be explored for getting insights about specific issues and for answering specific questions, zooming in and out for moving the focus among different levels of granularity.
- Finally, the full availability of all the data generated and of all the code employed to obtain the results allows for easy replicability of the study for other topics beyond the ones shown in this work.

In summary, we believe that this work offers relevant insights on the framing of key social issues related to the Next Generation Internet, and at the same time it opens up to further research, either replicating the analysis for other topics, or building on the datasets and code provided to extend our methods or develop new ones.

5

DISENTANGLING SOCIAL CONTAGION AND MEDIA DRIVERS IN THE EMERGENCE OF HEALTH THREATS AWARENESS

In recent years, thanks to the increasing abundance of (near) real-time, high-quality data on populations; human mobility patterns; and pathogens' biology; the use of data-driven computational models for the study of epidemic outbreak response has gained considerable traction in the public health community [65]. Novel digital data streams, such as search engine queries and social media, in combination with machine learning, statistical and mechanistic disease models, have considerably advanced outbreak forecast science. In this context, predictive epidemic modeling is emerging as an inter-disciplinary field that has been recently used in real time to support responses to recent outbreaks such as the 2009 H1N1 pandemic, the 2014 West Africa Ebola outbreak, the Zika epidemic in the Americas in 2016, and the Highly Pathogenic Avian influenza A(H7N9) in 2014 and 2015.

In view of the huge potential of predictive modeling, we must also be aware of the challenge inherent to the real-time modeling of the feedback loop between the disease progression and the response of social systems. Recent examples (Ebola, MERS, etc. [66, 67]) have shown that the spreading of infectious diseases strongly depends on the social adaptive behavior that characterizes the reaction of the population to the awareness and the perceived risk in the face of the epidemic [66, 67, 68, 69]. In other words, the predictive power of mathematical and computational models heavily relies on our understanding of the population awareness of the disease and the ensuing behavioral changes such as social distancing, travel limitations, etc. [70, 71, 72, 73, 74].

Several mathematical and computational models of the feedback mechanisms between disease spread and the effects due to the awareness of the epidemic in the population has been put forward in the modeling community [73, 74, 75, 76, 77, 78, 79]. In particular, two main mechanisms have been invoked in the spreading of awareness and fear of epidemics: i) the massive flow of news from mass media that possibly acts as an exogenous force on the global population [77, 80, 81]; and ii) the social contagion due to the word of mouth and peer influ-

ence [73, 74, 75, 76]. However, little is known about the relative contribution of the two mechanisms in shaping the spread of information in the population [77, 82, 83], and to what extent individuals' response is affected by mechanisms of social reinforcement [84]. As a matter of fact, the richness of models and mathematical approaches has not yet been transferred to any operational forecasting framework; mostly due to the lack of a quantitative data-driven characterization of the different mechanisms affecting the spreading of the awareness in the population [77, 85].

Here, we propose a modeling framework able to disentangle the contribution due to media drivers and the social contagion in the awareness building of infectious diseases. The developed model assumes that individuals become aware of an epidemic either by means of media communications on different stories related to the ongoing epidemic (through newspapers, websites, broadcasts, etc.) or as a consequence of conversations with other individuals. We test our modeling approach for several major recent epidemics by accurately fitting a proxy of the level of awareness in the population; i.e. the time series of Wikipedia page views [86, 87] related to the disease originating the epidemic. The underlying assumption here is that among the individuals that develop awareness, a fraction of them initiate an information-seeking behavior using Wikipedia to address this need. More precisely, we fit the proxy signal by using a non-linear, time-dependent mechanistic model; explicitly accounting for the combination of temporal changes in the media volume, as recorded in the Google News platform [88] and treated as an exogenous factor of the system, and an endogenous contagion process driven by risk perception and spontaneous social interactions between individuals. The proposed non-linear model is able to measure the effect of the endogenous progression of awareness in the population and quantify its contribution relative to the news and the media drivers. We perform a thorough information theoretical model selection analysis [89] showing that the proposed model outperforms supervised machine learning approaches based only on the news volume, and non-linear models accounting only for the endogenous social contagion component. Our approach outperforms the other considered models also in out-of-sample predictive tests. Interestingly, the mechanistic modeling proposed here allows the estimate of specific parameters such as the doubling time and transmissibility of the awareness through the different routes of information considered. Thus, this work has the potential to open the path to the inclusion of the diffusion of awareness in a wide range of biological and social contagion models, allowing the measurement of specific parameters of social contagion and the design of optimized awareness campaigns.

5.1 RESULTS AND DISCUSSION

We consider as a proxy of the awareness of major infectious disease epidemics the time series of Wikipedia page views for several health threats in the US and Italy. In particular, the 2014 Ebola Outbreak in West Africa [66, 90] and the spread of the Zika virus in the Americas during 2015 [91] are considered illustrative case studies to investigate the impact of *Public Health Emergencies of International Concerns* on the individuals' awareness in both US and Italian populations. Meningitis in Italy and Influenza in US between 2016 and 2017 are considered in order to investigate, respectively, the effect of local and persisting epidemiological threats [80, 92].

The level of media attention to the spreading of an infectious disease is certainly contributing to the population awareness of the disease. This can be readily observed by comparing the time series of the Wikipedia accesses and the volume of news on the epidemic measured from the Google News platform, a quantitative proxy of the media attention over time. As shown in Fig. 13 for the exemplary cases of the 2014 Ebola, the Wikipedia accesses and news volume time series have an extremely good correlation when no or negligible time lag is considered between the two signals (see the SI for a detailed analysis). A closer inspection to the curves, however, shows that, generally, the awareness tends to increase and decrease at a faster rate before and after the peak of the news cycle, respectively. For instance, in the US, after a huge volume of Wikipedia accesses and news about Ebola during week 42 of 2014, Wikipedia page views dropped by 64% in the following week, while media coverage decreased by 22% only, remaining significantly high up to mid-November. Similar temporal patterns were observed when comparing news coverage on Zika and individuals' online behavior in US (see SI), Guatemala and Brazil [93]. These features are the fingerprint of acceleration and saturation phenomena that can be ascribed to endogenous word of mouth processes. Here, as word of mouth processes, we can consider all personal communications among individuals, including both real-world and on-line contacts.

In our model, we assume that the overall strength of media coverage in a given day is proportional to the overall amount of news in that day, which include the contribution of both more and less relevant communications. In support of such assumption, it is worth noting that major peaks in the Google News time series associated with Ebola follows the occurrence of some relevant events for the two considered countries, as the detection and death of the first Ebola case within the US borders, and the first transmission recorded in EU (Fig. 13A,B). In order to test the representativeness and appropriateness of Google News data to model the media attention over time, we compared the

temporal dynamics of the number of Ebola related videos per day from Fox News and MSNBC [77] with the corresponding Google News signal (Fig. 13C). We found that the two signals are strongly correlated (Pearson correlation coefficient 0.92, p-value < 0.001), therefore suggesting that Google News represents a good proxy of the variation in strength of media coverage over time.

In order to model mechanistically the awareness spreading in the population we use a SIRS transmission mechanism [94]. In the transmission model, susceptible individuals (S) represent people unconcerned or unaware about an epidemic threat that can get informed (I) either through the word of mouth, as a consequence of a social contagion based on peer-to-peer individual interactions with informed individuals (I), or directly from the media (M). Potential development of immunity (R) against the exposure to new information and possible mechanism of immunity waning is fully considered with the SIRS model. The model is calibrated by assuming that at each time stamp t, a fraction k of individuals becoming aware of the disease seeks further information on Wikipedia, which is a proxy of awareness acquired through other means. At each time stamp, the model keeps track of all the newly become aware individuals I_t^{New}; i.e. the number of individual transitions $S \to I$. The Wikipedia page views (W_t) at time t is then given by the following relation

$$W_t = \kappa I_t^{\mathrm{New}} + W_0$$

where W_0 represents the number of Wikipedia page views in absence of media attention. The number of individuals I_t^{New} who become aware of the epidemic threat at day t is determined by a dynamical process where the rate for the individual transition $S \to I$ into the aware state is defined as $\lambda_t^{\mathrm{SN}} = \lambda_t^S + \lambda_t^N$, where the terms λ_t^S and λ_t^N represent the social and the news contributions to the force of contagion, respectively. The force of contagion is itself depending on the number of aware individuals I_t and the volume of news M_t at time t. This social-news contagion (SN) model is compartmental (i.e. individuals with the same characteristics are represented by a unique dynamic variable), and takes into account also individuals not yet aware of the disease, and individuals aware but not actively interested in seeking or spreading information on the disease. Mechanisms of waning and re-emergence of the awareness are also considered in the model. For each epidemic, the model's parameters are calibrated separately by using a Markov Chain Monte Carlo (MCMC) approach [95]. The explicit mathematical definition of the model, and the MCMC calibration are detailed in the Material and Methods section.

In Fig. 14, we report the results obtained by using the calibrated model to fit the Wikipedia time series. Remarkably, the model is able to cap-

ture the large fluctuations induced by the media cycle and the quick rise and decay of attention more typical of the social contagion processes [96]. In order to evaluate the performance of the SN-model, we have considered three other modeling approaches. The first one is a classic regression based on a supervised machine learning approach (L) that models the Wikipedia signal by considering as explanatory variables the news released during the preceding days (see details in the Material and Methods section). We then considered two alternative contagion models, the S-model and the N-model, containing only the λ_t^S and λ_t^N force of contagion, respectively. A closer inspection of these models (reported in the SI file) shows that each one of these alternative models has clear limitations. For instance, the N-model appears to not reproduce accurately the decay of interest because of saturation in the awareness process [93]. On the contrary, the S-model accounts for the quick rise of awareness but does not captures fluctuations that can be traced back to the media cycle. In order to put on rigorous ground the comparison across models, we report in Table 28 the basic metrics describing the goodness of fit of the various models (additional metrics in SI). We observe that the SN-model outperforms all other models on the basic quantities such as the mean absolute percentage error (MAPE), the Pearson correlation coefficient, and the coefficient of determination (R^2). In Fig. 15, we also report the relative error of each model as a function of time for the case of Ebola in US and Italy, showing that along the entire time window the SN-model is consistently better performing than the other models. In addition, since we are comparing models with different numbers of parameters, we performed an information theoretic multi-model analysis [66, 89, 97] that clearly shows a very low likelihood that any models could better explain the data with respect to the SN-model (see Table 28).

Finally, to assess the predictive power of different modeling approaches here considered, model performances of the SN-model were compared with those obtained with the L-model, by using only the first 80% of data points (train set) for model calibration and by testing model compliance with the remaining 20% of data (test set). The goodness of fit associated with the two models was compared in terms of MAPE, R^2, and Pearson correlation coefficient, and also compared with the goodness of fit in the last 20% of data points obtained when using 100% of data for model calibration. The carried out analysis suggests that in all these cases, SN-model performs better than the supervised machine learning approach (L, see Table 2).

Our analysis shows that in order to model the spreading of awareness in a population, both influence of media and social contagion are relevant mechanisms. A model accounting for both these routes to develop awareness proves to better reproduce and forecast the dynamics of Wikipedia accesses over time. The model has the added benefit of being

mechanistic in describing the impact of media communications and peer-to-peer contagion processes. This allows us to quantify the specific contribution to the process from the two drivers considered and to measure some key features characterizing the spread of awareness in a population.

Our estimates show that, among the different epidemiological scenarios considered, the fraction of individuals that acquired awareness from media broadcast ranges from 30% to 60% (Fig. 4). This result strongly suggests that both media and the word of mouth represent crucial components of the awareness dynamics. Our results are compliant with recent estimates showing that 20% of visits on Wikipedia are triggered by conversations with other individuals and 30% by the media coverage [82].

The estimated proportion of aware people who seek information on Wikipedia (Fig. 16) is expected to mirror the level of interest and concern raised during an epidemic. This proportion was found higher for epidemics declared *Public Health Emergencies of International Concerns* by the World Health Organization, i.e. Ebola in 2014 (4.6% US, 1.4% Italy) and Zika virus in 2015 (0.9% US, 0.2% Italy), with respect to well known infections as Meningitis (0.08% in Italy between 2016 and 2017) and Influenza (0.14% in US between 2016 and 2017). These estimates suggest that a higher concern is triggered by the emergence of pathogens representing relatively new threats, associated with higher mortality rates [81, 93]. Although the occurrence of severe cases in higher proximity may increase the public attention to a specific disease, the estimated higher impact of Ebola and Zika in US with respect to Italy may be driven by the use of Wikipedia pages written in English by other countries (only 41% of access to Wikipedia pages in English are from US; 91% of access to Italian pages are from Italy [98]).

Model estimates of the media transmission rate during different epidemic scenarios (Fig. 16) provide insights on the impact of media communications on different topics across countries. This quantity represents the average number of individuals who get informed by the release of one single news within 24 hours, in a completely uninformed population. Average estimates of this quantity suggest that the impact of media in US might have been more intense (or effective) during the epidemics of Zika in 2015 and Ebola in 2014 with respect to what has occurred during the 2016-17 Influenza season (see Fig. 16) [80, 81, 93]. In Italy, a higher media transmission rate was estimated for Ebola in 2014 and Meningitis in 2016-17 with respect to what observed during the 2015 Zika epidemic. Interestingly, for all the scenarios considered but for Influenza in US, the accumulation of media communications over time resulted as an amplifier of the impact of news released afterwards (see SI). The negligible effect of past media coverage on the effectiveness of new media communications estimated for Influenza be-

tween 2016 and 2017 may be explained by the perceived lower severity of Influenza infection with respect to others and the occurrence of annual regular Influenza epidemics [80].

By comparing estimates of the doubling time characterizing the social contagion mechanism during different epidemics (Fig. 16; i.e., how fast the number of individuals informed though the word of mouth doubles), we found that the spread of awareness through peer-to-peer individual interactions was more than three times faster during 2014 Ebola in the US and 2016-2017 Meningitis in Italy. These two events possibly represent the two threats that have mostly changed the disease risk perception in the public [80, 81, 93, 92].

Finally, as for the Ebola awareness dynamics, the estimated average duration of immunity against the exposure to new information; driving possible waning of individuals' awareness; and representing the average time between two visits on a Wikipedia page by the same individual, resulted around 80 days for US and 110 days for Italy. Obtained estimates roughly correspond to the period of time between the two major peaks in the Wikipedia and Google News signals and suggest that two dominant sets of Ebola stories impacted the population in the two countries during 2014. On the opposite, for Zika, and Influenza in US, the estimated duration of immunity exceeds the time frames considered in our analysis, suggesting that the awareness dynamics associated with these two epidemics has been influenced by news stories that has occurred within a relatively narrow period of time. For Meningitis in Italy, the average duration of immunity was found around 50 days.

5.2 CONCLUSIONS

The approach presented here provides a modeling framework for investigating time-series related to the spread of awareness of health threats in a population. Our proposed model goes beyond the usual statistical analysis with respect to dependent exogenous variables; indeed, we introduce a mechanistic modeling approach based on the idea that along the news and media drivers, peer-to-peer social contagion plays a major role in the emergence of awareness.

It is worth remarking a number of limitations for the presented approach. First, the models considered for social contagion driven by the word of mouth assume a homogeneous mixing in the population. However, the influence of individuals may be highly heterogeneous and significant different contributions of opinion leaders with respect to the ones of less active individuals may affect the spread of awareness in a population [83, 99]. Clustering of opinions around a given topic within specific population groups or geographical areas is also likely to occur

in response to local events [74] or as a consequence of different cultural backgrounds, social norms, and social reinforcement mechanisms [84]. Second, the influence of media is here estimated by using only news released by news websites. Although these are likely representative of the overall media attention, we did not consider the heterogeneous contribution of different mass media (e.g. television, radio, newspapers) in shaping the spread of awareness and the impact of news was considered regardless the content of different communications and the reputation of who provides the information. More refined models may explicitly take into account the heterogeneity in the infectiousness associated with different peers (both for media communications and social interactions) instead of adopting a mean field approximation. All these aspects may affect the impact of media on the public, the persistence of concern generated by news, and may also be relevant to explain the spread of fake news and misperception [73, 80, 81, 92, 100, 101]. Furthermore, media attention is here considered as an exogenous factor of the considered dynamics, taken as given regardless potential feedback responses between the onset of new stories and people need of more information on a given topic. Finally, our analysis does not provide any indication on whether and how the public has changed their behavior in response to the perceived risk of infection [66, 69, 70, 71, 72, 73, 74, 79]. Further efforts are, therefore, required to better understand similar mechanisms.

This study, however, represents a first important attempt to qualitatively and quantitatively describe the role played by the media and the word of mouth in influencing individual awareness and risk perception during different epidemic threats. In particular, the proposed model is able to disentangle the contribution of news influence and social contagion in driving Wikipedia page views in the case of six different public health threats. Furthermore, it provides better explanatory and forecasting power than alternative models considering only one of the driving mechanisms. Most importantly, the model allows the measurement of parameters defining the contagion process such as the fraction of aware people, and the relative contributions of the different contagion processes. The possibility of gathering quantitative information on these parameters is a first step in the operationalization of epidemic models that include the spread of awareness to diseases and the ensuing behavioral changes of the population. The proposed framework is fairy general and can be applied in other contexts related to the diffusion of information and knowledge.

5.3 MATERIAL AND METHODS

DATA DESCRIPTION

The daily numbers of visits to Wikipedia pages on specific diseases were obtained by publicly available data [86, 87] and used to model temporal changes in the number of individuals seeking information on a specific epidemic threat. Pages in Italian and English, accessed between 2014 and 2017, were used for Italian and US users, respectively. Wikipedia data were preferred to other public available datasets (e.g., Google Trends Data, etc.) for the following reasons: a) Wikipedia visits may better reflect the need to acquire in-depth information about a diseases, instead of capturing recurrent accesses to an updated source of information on what is occurring during an epidemic; and b) access of individuals is provided in absolute numbers, therefore, allowing a comparison of time-series associated with different countries and diseases. Accesses during 2016 on any Wikipedia page in two languages were considered to assess, separately for the two countries, potential differences in the individual use of Wikipedia across different days of the week (e.g., weekdays/weekend). Deviations with respect to weekly means were used to remove spurious fluctuations in Wikipedia accesses (details can be found in SI). Temporal changes of media attention to a given epidemic were modeled by using Google News platform [88]. Specifically, a proxy of daily media response to potential threats was defined by the amount of articles released by websites based in the country containing the name of the considered epidemic in its headline. A cross-correlation at different time lags was performed to assess potential synchrony in the two types of signals.

SUPERVISED MACHINE LEARNING APPROACH

We conduct a multi-linear regression analysis (L) in order to test whether the dynamics of Wikipedia accesses mirror changes of the media attention to a given disease. We consider the number of Wikipedia page views at time t as the response variable and the amounts of news released at different times prior t as potential explanatory variables. Specifically a set of linear models was defined as follows:

$$W_t = \sum_{i=0}^{T} \alpha_i M_{t-i} + W_0$$

where W_t represents the daily number of Wikipedia page views at time t, α_i are the regression coefficients; M_t is the number of news released at time t, T defines the number of days before t, in which media communication can affect Wikipedia accesses; and W_0 reflects accesses to Wikipedia in the absence of media coverage on a given disease. Best values for T were obtained as a result of a multi-model information approach, based on the Akaike information criterion [89]. Details are reported in the SI.

MODELING AWARENESS AS A CONTAGION PROCESS

Word of mouth and media communications are both considered as plausible routes of transmission. The spread of information through the word of mouth was modeled by assuming homogeneous-mixing in the population and a force of infection $\lambda_t^S = \frac{\beta_S}{N} I_t$, where β_S represents the transmission rate for the social contagion associated with peer-to-peer conversations between individuals.

As for the rate at which individuals becomes aware thanks to media communication, we assume that the impact of news released at time t can be inflated by news released in the previous days. The force of infection from media communications is, therefore, defined as follows:

$$\lambda_t^N = \frac{\beta_N}{N} M_t \left(\prod_{i=t_0}^{t-1} 1 + M_i e^{-\rho(t-i)} \right)$$

where β_N is the transmissibility potential related to media coverage at time t, while the term in parentheses defines the amplifying mechanism due to past media communications. Specifically, the contribution of past communications to the amplification of the strength of news released afterwards is assumed to exponentially decay with time passed since their release: smaller the value of ρ, longer the influence of past communications. For large value of ρ, the model considers only media communication released at time t, i.e. $\lambda_t^N = \frac{\beta_N}{N} M_t$. Alternative amplification mechanisms are considered in the SI. The two routes of transmission are considered alone (models S and N) or combined in a nested model (SN).

Briefly, in the proposed models, transitions between classes can be described by the following system of ordinary differential equations:

$$\begin{aligned}
\frac{dS}{dt} &= -\lambda(t)\,S(t) + \nu R(t)\,; \\
\frac{dI}{dt} &= \lambda(t)\,S(t) - \gamma I(t)\,; \\
\frac{dR}{dt} &= \gamma I(t) - \nu R(t)\,,
\end{aligned}$$

where $\lambda(t)$ represents the force of contagion at time t; $\frac{1}{\gamma}$ is the average time period, in which an aware individual can spread the information through word of mouth; $\frac{1}{\nu}$ is the average duration of immunity against the exposure to new information. Such a period of time can be interpreted as the average time between two visits on a Wikipedia page by the same individual. The proposed SIRS model, for specific choices of free model parameters, can degenerate into a SIR model (no immunity waning), and even into a SI model (no immunity).

MODEL CALIBRATION AND GOODNESS OF FIT

Each model was calibrated separately. Free model parameters were estimated by using a Markov Chain Monte Carlo approach [95] applied to the negative binomial Likelihood of the observed Wikipedia page views for each scenario. Goodness of fit was assessed through a wide set of statistical measures including, among others, the Akaike and Deviance Information Criteria (AIC, DIC), the mean absolute percentage error (MAPE) and the Pearson correlation coefficient (R^2) [66, 89, 97]. Details can be found in SI.

		MAPE	Pearson	R2	AIC	Probability of information loss
Ebola US	L	84.895	0.666***	0.352	3709.75	< 0.001
	N	56.883	0.832***	0.659	3635.1	< 0.001
	S	70.214	0.584***	-8.853	3648.48	< 0.001
	SN	39.597	0.844***	0.551	3556.51	–
Ebola Italy	L	109.63	0.61***	0.289	2831.26	< 0.001
	N	80.224	0.706***	-4.744	2792.92	< 0.001
	S	97.871	0.729***	0.509	2808.03	< 0.001
	SN	46.521	0.913***	0.809	2681.16	–
Zika US	L	72.545	0.787***	0.523	3259.58	< 0.001
	N	52.509	0.932***	0.846	3173.02	< 0.001
	S	33.393	0.89***	0.775	3080.87	< 0.001
	SN	30.482	0.929***	0.841	3059.46	–
Zika Italy	L	203.127	0.782***	0.492	2017.67	< 0.001
	N	374.899	0.874***	0.71	2119.44	< 0.001
	S	58.985	0.805***	0.041	1911.98	< 0.001
	SN	56.774	0.871***	0.746	1853.94	–
Meningitis Italy	L	71.238	0.569***	0.311	3392.09	< 0.001
	N	57.034	0.792***	0.588	3333.81	< 0.001
	S	74.089	0.524***	0.161	3369.43	< 0.001
	SN	51.203	0.739***	0.526	3280.18	–
Influenza US	L	14.066	0.781***	0.589	4907.52	< 0.001
	N	16.772	0.719***	0.503	4997.7	< 0.001
	S	17.231	0.706***	0.49	5000.99	< 0.001
	SN	10.138	0.88***	0.773	4693.88	–

*** p-value < 0.001

Table 28: Basic metrics describing the goodness of fit associated with different epidemics and models including the mean absolute percentage error (MAPE), the Pearson correlation coefficient, the coefficient of determination (R^2) and the Akaike information criterion (AIC). Values of R^2 were computed on the basis of equation $y = ax$, thus allowing for negative R^2 values. AIC was used to estimate the probability that information loss is minimized when we consider an alternative model to the model having the lowest AIC value.

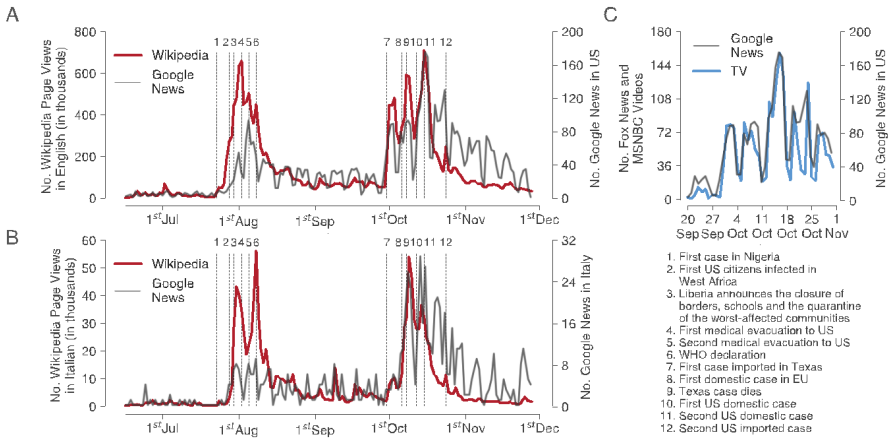


Figure 13: Proxies of health threat awareness and media attention for the illustrative case of Ebola epidemic. **A)** Red line represents the daily number of page views on Wikipedia articles related to Ebola infection during 2014 from the English version of Wikipedia [86]. Grey line represents the daily number of news released on Ebola in the US, as obtained from the Google News platform [88]. Dotted lines indicate noticeable events associated with the West Africa Ebola epidemic. **B)** as A), but for the Italian version of Wikipedia and news released in Italy. **C)** Comparison of two different proxies of media attention to the Ebola epidemic. Grey line represents the daily number of news released on Ebola in the US, as obtained from the Google News platform [88]. Blue line indicates the number of Ebola related videos per day, from Fox News and MSNBC [77].

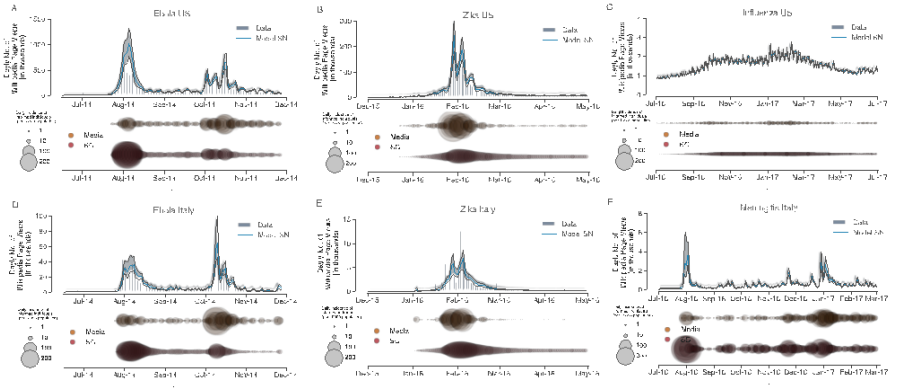


Figure 14: SN-model estimates for Ebola **(A)**, Zika **(B),** Influenza **(C)** in the US and for Ebola **(D)**, Zika **(E),** Meningitis **(F)** in Italy. In each panel, blue bars represent the daily number of Wikipedia page views over time for the considered infection. The blue lines and the shaded areas refer to the average and the 95% CI of estimates as obtained with the SN-model on the daily number of informed individuals seeking information on Wikipedia. Bubble plots represent the median incidences of informed individuals. Yellow and red bubbles refer to incidences of informed individuals by media communications and through social contagion, respectively.

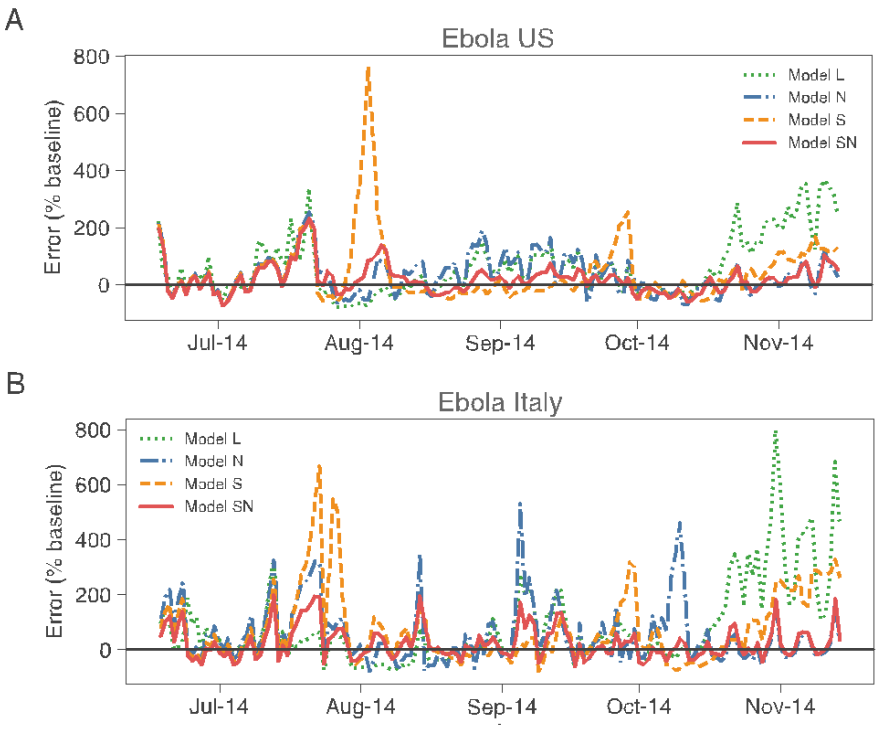


Figure 15: Percentage error between model estimates and data records on the number of Wikipedia page views over time, as obtained for different models when considering the Ebola awareness epidemic in the US **(A)** and in Italy **(B)**.

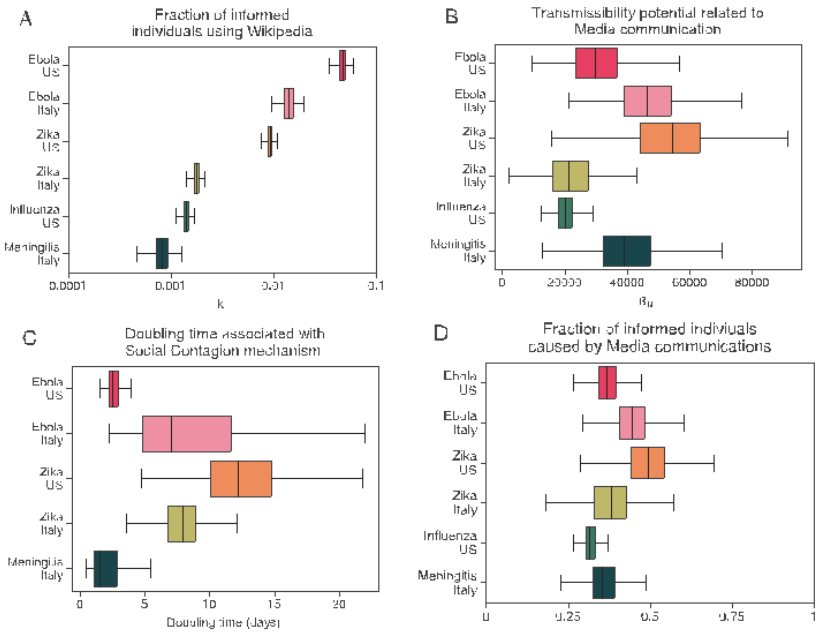


Figure 16: Posterior distribution (2.5%, 25%, 75%, and 97.5% quantiles and mean) of fraction of informed individuals using Wikipedia **(A)**, the transmissibility potential related to media communications **(B)**, the doubling time associated with the social contagion mechanism **(C)**, the fraction of informed individuals due to media communications **(D)**, as obtained for the different epidemic scenarios considered with the SN-model.

Scenario	Model	MAPE		Pearson correlation coefficient		R^2	
		80/20	**Baseline**	80/20	**Baseline**	80/20	**Baseline**
Ebola US	L	258.46	188.22	0.85***	0.85***	-2.08	-0.53
	SN	46.3	31.93	0.91***	0.92***	0.7	0.83
Ebola Italy	L	382.95	291.96	0.77***	0.78***	-3.74	-1.58
	SN	46.91	38.88	0.94***	0.94***	0.81	0.86
Zika US	L	122.38	35.01	0.50*	0.53*	-17.06	-0.72
	SN	53.61	30.23	0.90***	0.87***	-4.59	-1
Zika Italy	L	121.08	47.77	0.57**	0.63**	-1.59	0.34
	SN	42.38	40.64	0.90***	0.90***	0.27	0.64
Meningitis Italy	L	230.92	124.91	0.77***	0.77***	-10.57	-1.99
	SN	69.77	41.53	0.49**	0.67***	-0.27	0.4
Influenza US	L	21.48	19.91	0.26**	0.27**	-1.56	-1.22
	SN	16.92	13.75	0.48***	0.48***	-0.29	0.09

* p-value < 0.1
** p-value < 0.05
*** p-value < 0.001

Table 29: Statistical measures on the performances of the SN-Model and the supervised machine learning approach (L), as obtained for different epidemic scenarios. Each measure was obtained for two distinct calibration procedures. In the first one (labeled as 80/20 in the table), model parameters were calibrated using only 80% of data; in the second (baseline) model parameters were calibrated using 100% of data. In both cases, performances were assessed only in the last 20% data points.

6

CONCLUSIONS

In this thesis we have explored the idea of extracting knowledge from links in Wikipedia.

In the first chapter, we have laid the foundation of our work by building a dataset called WikiLinkGraphs, which makes available the complete graph of links between Wikipedia articles in the nine largest language editions.

This dataset overcomes the limitations of previous work by being more clean, since we have removed templates and resolved redirects, and complete, since we have analyzed over 1 billion revisions to build a temporal graphs starting since the inception of Wikipedia. We provided both yearly snapshots and the raw dataset containing the complete history of every single link within the encyclopedia.

In the second chapter, we introduced *CycleRank*, a novel algorithm based on cyclic paths that can be used to find the most relevant nodes in the Wikipedia link network related to a topic. Given a reference node in a directed graph, the algorithm finds all simple cycles that go through the reference node and assigns a score to each node that belong to these cycles. The algorithm is simple yet flexible: it has only one parameter, K maximum cycle length that we are interested to find, but its behavior can be regulated also by changing the function used to assign scores to nodes. We believe that this flexibility can be leveraged when applying *CycleRank* to graphs with diverse characteristics.

We have carried out an extensive comparison between *CycleRank*, *PageRank*, and its variants such as *CheiRank*, and *2DRank*. We presented both a qualitative analysis over several topics and a quantitative evaluation based on three measures with corresponding ground truth data: *ClickStream*, *See-Also* and top-indegree pages. These measure encapsulate respectively the ability of the algorithm to maintain the relative importance between articles, its ability to find related content, and the ability to avoid incorporating in the results very popular topics that are not well-tailored to the particular under consideration, we have shown that *CycleRank* is more robust than other baselines with respect to this problem. Furthermore, we have shown that our algorithm is faster than the alternatives, offering order-of-magnitude speed-ups with respect to a library implementation of *Personalized PageRank*.

In other words, *CycleRank* is a viable alternative to *PageRank*, especially in the case of graphs where the role of inlinks and outlinks is comparable; we believe that it can be easily applied to similar contexts, such as knowledge bases.

In the third chapter we have show how to use *CycleRank* to characterize a topic through inspecting its connections to other topics. We believe that this work offers relevant insights on the framing of key social issues related to the Next Generation Internet, and at the same time it opens up to further research, either replicating the analysis for other topics, or building on the datasets and code provided to extend our methods or develop new ones.

Finally, in the last chapter we have show how to use other data from the Wikipedia ecosystem, namely visitors log, to understand human-behavior and phenomena like the spreading of diseases.

As future work we plan to show how *CycleRank* can be used to find the most relevant input for a machine learning system. Confronted with the task of estimating the level of influenza-like illness (ILI) in Europe, we select the most relevant articles concerning *Influenza*. Once selected the article we used a dataset with the number of accesses to Wikipedia articles for each day between December 2007 and August 2017 and compared these data to official ILI activity levels provided by the European Centre for Disease Control and Prevention (ECDC). We aim to answer the following question: "Do the pages found by *CycleRank* correlate to the number of cases of ILI measured through the article views?"

As a final remark we would like to highlight that all the code written by us to produce the works presented in this thesis has been made publicly available. This also applies to the dataset we have produced, not just the WikiLinkGraphs dataset but all the supporting datasets as well. We think that this creates a number of research opportunities that we list hear in the hope that they will pick the interest of the broader research community:

- the WikiLinkGraphs dataset is currently made available for the nine largest Wikipedia language editions, however it can be extended to other language editions, or by creating snapshots with a finer temporal granularity. Being a very large network of concepts, the WikiLinkGraphs dataset can benefit several communities such as the Semantic Web, the Artificial Intelligence, and natural language processing research communities.
- the version of *CycleRank* presented in this thesis is based on a simple exponentially-decaying scoring function, we have also tested a linear-decaying scoring function; while we have empirically validated the results, many variations and extensions could be explored.

- we have assumed that the starting point for *CycleRank* is a single reference node. However, as in the case of *Personalized PageRank*, it would be possible to take a group of articles as seed. Then, all cycles around each of the seed nodes could be considered. Alternatively, instead of counting cycles, one could count all paths from any node in the seed to any other node in the seed.
- another possible variant would be to specify two different nodes (or groups of nodes) as source and target, and to consider all paths from the source to the target within K steps. In this way, the measure would not only represent the relevance of other nodes with respect to a given reference node, but the (directed) relationship between two nodes or groups of nodes. This would help to answer questions such as: "Which are the most relevant concepts connecting Artificial Intelligence and Human rights, and which are the most relevant concepts on the other way round"?
- we think that *CycleRank* provides a foundation that could be further explored to provide a family of algorithms adapted for different graphs and use cases. The suitability of different solutions could also be studied focusing on the structural properties of the network, such as its link density or clustering coefficient.

We hope that future research will continue to be shared as openly as possible and we hope to have provided with this thesis a good example of the benefits of Open access and Open Science.

Part IV

APPENDIX

A

THE ENGINEROOM EU PROJECT

In this Appendix, we present additional results from the ENGINEROOM EU project, covered in Chapter 4. For each keyword we present: a longitudinal analysis across several years on English Wikipedia, covering from the creation on the article to 2018; and a cross-language analysis over the most recent data and all 9 languages available.

A.1 ALGORITHMIC BIAS

In the case of Algorithmic bias, the article was not existing (or not sufficiently developed as to have some loops) in English before March 2017, and in the other languages before March 2018. Therefore, we only show results for the latest snapshot of the English Wikipedia in our dataset (March 2018), and do not perform cross-language or longitudinal analysis for this topic.

On the other hand, we observe that in March 2018 the article was already well linked to and from relevant concepts, as it can be observed looking at the network. This is an interesting indicator of the quickly growing attention around this topic.

Figure 17 shows the network of article with *CycleRank* score greater than zero. The main cluster (shown in magenta) containing the article Algorithmic bias itself includes as the most relevant concept Machine learning, and then tech companies such as *"Google," "Facebook," "Apple Inc," "Amazon,"* and general concepts such as *"Internet of Things"*. It also contains the article about the *"GDPR"*. The second concept for importance after Machine learning is *"Artificial intelligence,"* at the center of a cluster (shown in green) including concepts related to Ethics of artificial intelligence. A smaller cluster (shown in cyan) contains general concepts including *"Algorithm," "Computer science," "Cybernetics," "Complex systems," "Social science"*. Finally, a red cluster includes concepts with a low *CycleRank* score from the field of statistics; the most relevant concept in the cluster is *"Credit score"*.

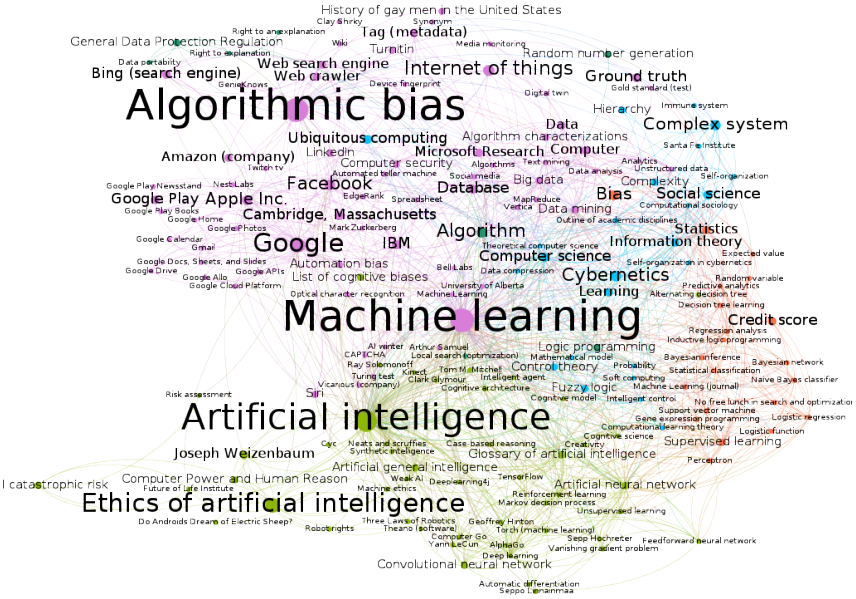


Figure 17: Graph induced by the nodes with non-zero *CycleRank* score with reference node r = *"Algorithmic bias"* and $K = 4$ on English Wikipedia, over the snapshot of March 1st, 2018. Colors represents clusters calculated with the Louvain algorithm. The dimension of the nodes and their labels depends on the *CycleRank* score.

A.2 CYBERBULLYING

Figure 18 represents the network of articles around the article *"Cyberbullying"* with *CycleRank* score greater than zero. We see that this network is characterized by 5 clusters:

1. (purple) groups news media, public figures and what could be defined the "mainstream": *"CNN,"* *"The New York Times,"* *"ABC News,"* *"The Guardian,"* *"Barack Obama,"* and *"Emmanuel Macron"*, *"Pope Francis"*, but also *"Arab Spring"* and *"First Amendment of the United States Constitution"*;

2. (green) is related to bullying and harassment: *"Stalking,"* *"Misogyny,"* *"Suicide of Jadin Bell,"* *"Suicide of Megan Meier,"* *"Suicide of Ryan Halligan,"* *"Suicide of Tyler Clementi"*;

3. (cyan) groups concepts related to the Internet: *"World Wide Web,"* "*Google,*" *"Internet Service provider."* and *"HTML"*;

4. (orange) is similar to the previous but it is specifically dedicated to social media: *"Facebook,"* *"Twitter,"* *"MySpace,"* *"4chan,"* *"Mashable,"* *"Blog,"* *"Internet meme"*;

5. (dark green) groups concepts related to online gaming and games: *"Gamergate controversy,"* *"Youtube,"* *"Video game*

culture," "Call of Duty," "League of Legends," and *"Online game".*

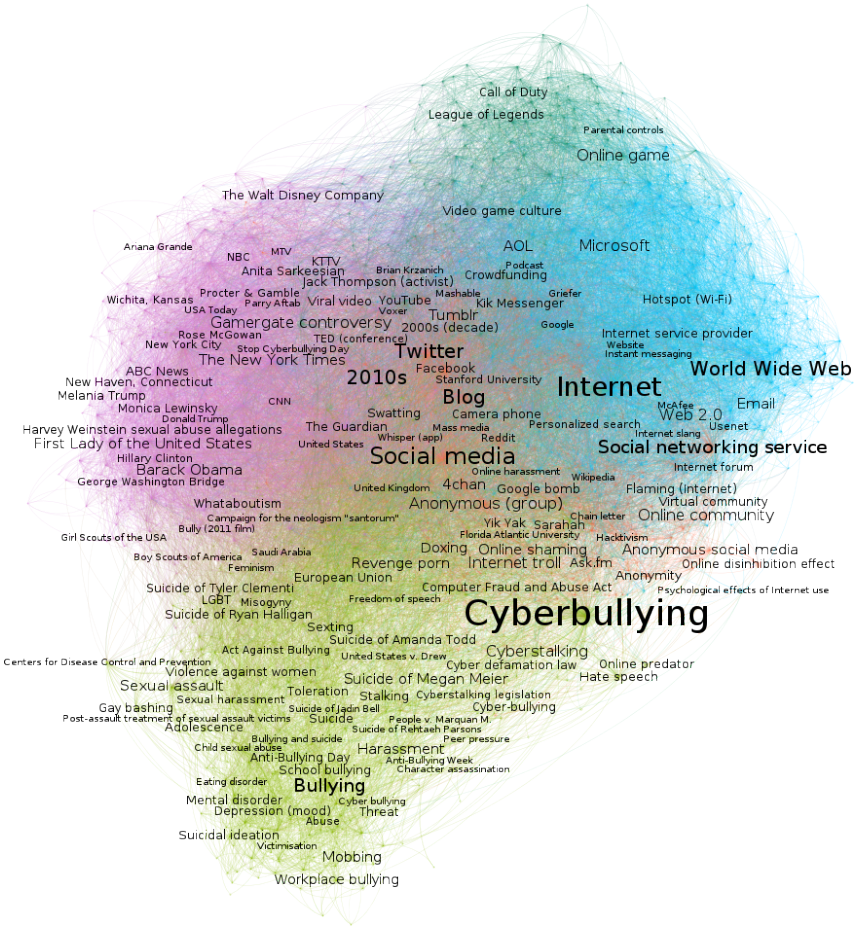


Figure 18: Graph induced by the nodes with non-zero *CycleRank* score with reference node $r =$ *"Cyberbullying"* and $K = 4$ on English Wikipedia, over the snapshot of March 1st, 2018. Colors represents clusters calculated with the Louvain algorithm. The dimension of the nodes and their labels depends on the *CycleRank* score.

A.2.1 LONGITUDINAL ANALYSIS

Tables 36 and 37 show how the context for the concept of *"Cyberbullying"* has changed over time.

We see how the top-10 articles contain references to prominent events and people that led the public discourse about harassment online and cyberbullying:

- 2006: *"Eric Harris and Dylan Klebold,"* the shooters of the Columbine High School massacre;

- 2008: *"Louise Burfitt-Dons,"* a British writer and humanitarian mostly known for her work anti-bullying work and her charity "Act Against Bullying", which appears in the terms of 2009 and 2010;
- 2009: *"Suicide of Megan Meier,"* the case of a 13-year-old girl who committed suicide in 2006. The suicide was attribute to cyberbullying on MySpace and online harassment and the episode led then to a court case that was adjudicated in 2009;
- 2010: *"Suicide of Ryan Halligan,"* the case of a 13-year-old boy who committed suicide in 200 after being bullied in person and online. The case gained prominence in the following years as his father lobbied local politicians to introduce anti-bullying and suicide prevention laws;
- 2012: *"Suicide of Tyler Clementi,"* the case of a 18-years-old boy who committed suicide after an intimate video of him was divulged on Twitter without his consent.

Over time several social media appear in prominent positions: *"MySpace"* (2011), *"Facebook"* (2011, 2013-2017) and *"Twitter"* (2018), suggesting that although platform usage changed, and new platforms emerged during the years, the phenomenon persisted permeating different online spaces.

A.2.2 CROSS-LANGUAGE ANALYSIS

The cross-language analysis results are presented in Table 38. The page *"Cyberbullying"* was present in all the 9 language studied in the snapshot from March 1st, 2018. The most frequent result in the top-10 is *"Facebook"*, highlighting the prominence of the role that the social network is covering with regards to the topic of Cyberbullying. Also Twitter, WhatsApp and Wikipedia appear in the results along with the more generic terms *"SMS"*, *"chat"*, and *"blog"*. Some language editions such as English, German and Swedish seem to highlight more the technological aspects and terms related to the online context, while other such as Spanish or French give more prominence to concepts related to the phenomenon of bullying, such as suicide, intimidation or different kinds of harassment.

rank	2006	2007	2008	2009	2010	2011	2012
1	Cyberbullying	Cyberbullying	Cyberbullying	Cyberbullying	Cyberbullying	Cyberbullying	Cyberbullying
2	Computer crime	Cyberstalking	Act Against Bullying	Cyberstalking	Cyberstalking	Suicide of Megan Meier	Harassment
3	Cyberstalking	Harassment by computer	Chappaqua, New York	Cybercrime	Cybercrime	Cyberstalking	Suicide of Megan Meier
4	Instant messaging	Harassment	Louise Burfitt-Dons	Suicide of Megan Meier	Suicide of Megan Meier	Harassment	Cyberstalking
5	Talker	Computer crime	Bill Clinton	Act Against Bullying	Act Against Bullying	Stalking	Stalking
6	Cyberterrorism	Sexual harassment	Hate speech	Cyberterrorism	Suicide of Ryan Halligan	Facebook	Threat
7	Internet Relay Chat	Bullying	January 31	Stalking	Stalking	Myspace	Act Against Bullying
8	Online chat	Cyberterrorism	2007	Hate speech	Online predator	Suicide of Ryan Halligan	Suicide of Tyler Clementi
9	Eric Harris and Dylan Klebold		Bullying	Chappaqua, New York	Cyberterrorism	Cyberstalking legislation	School violence
10	Bullying		Hate group	Harassment	Anonymous (group	Social software	Bullying

Table 30: Top 10 most relevant concepts by *CycleRank* score with respect to *"Cyberbullying,"* along different yearly snapshots of the WikiLinkGraphs dataset (2006-2012).

rank	2013	2014	2015	2016	2017	2018
1	Cyberbullying	Cyberbullying	Cyberbullying	Cyberbullying	Cyberbullying	Cyberbullying
2	World Wide Web	World Wide Web	Internet	Internet	Internet	Internet
3	Social networking service	Social networking service	World Wide Web	World Wide Web	Social media	Social media
4	Facebook	Facebook	Email	Social networking service	World Wide Web	World Wide Web
5	Cyberstalking	Email	Facebook	2010s	Blog	2010s
6	Anonymous (group)	Cyberstalking	Social networking service	Facebook	2010s	Blog
7	Suicide of Megan Meier	Microsoft	Cyberstalking	Bullying	Facebook	Twitter
8	Suicide of Tyler Clementi	Suicide of Megan Meier	Anonymous (group)	Email	Social networking service	Bullying
9	Microsoft	Online game	Microsoft	Cyberstalking	Anonymous (group)	Social networking service
10	AOL	Adolescence	Social media	1990s	Bullying	Anonymous (group)

Table 31: Top 10 most relevant concepts by *CycleRank* score with respect to *"Cyberbullying,"* along different yearly snapshots of the WikiLinkGraphs dataset (2013-2018).

rank	de	en	es	fr	it	nl	pl	ru	sv
1	Cyber-Mobbing	Cyberbullying	Ciberacoso	Cyberharcèlement	Cyberbullismo	Cyberpesten	Cyberprzemoc	Internet-travlja	Nätmobbning
2	Facebook	Internet	Acoso escolar	Sexisme	Internet	Sociaalnetwerksite	Komunikator internetowy	Travlja	Näthat
3	Soziales Netzwerk (Internet)	Social media	Acoso sexual	Violence	Chat	Facebook	Internet	Mobbing	World Wide Web
4	Mobbing	World Wide Web	Acoso psicológico	Adolescence	Malware	Twitter	1996	Bulling	Internettroll
5	Internet	2010s	Hostigamiento	Pseudonyme	Facebook	Sociale media	GG (komunikator internetowy)	Kibermobbing	Netikett
6	Online-Community	Blog	Acoso laboral	Harcèlement scolaire	Bullismo	Cyberbaiting	SMS	Personal'nyj komp'juter	Flashback Forum
7	Social Media	Twitter	Abuso sexual	Patrick Bruel	WhatsApp	Pesten (gedrag)	Komunikator Tlen.pl	Internet	E-post
8	Chat	Bullying	Suicidio	Najat Vallaud-Belkacem	SMS	Bezemen	Facebook	Samoubijstvo	Mobbning
9	Mobbing in der Schule	Social networking service	Acoso familiar	Intimidation	Blog	Stalking	Filtr rodzinny	Komp'juternaja revoljucija	Webbplats
10	Sexting	Anonymous (group)	Acosador	Harcèlement moral	Suicidio	Wikipedia	Anonimowość	Massovoe ubijstvo v Mjunhene (2016)	Facebook

Table 32: Top 10 most relevant concepts by *CycleRank* score with respect to *"Cyberbullying,"* across different Wikipedia language editions.

A.3 COMPUTER SECURITY

Figure 19 represents the network of articles around the article *"Computer security"* with *CycleRank* score greater than zero. We see that this network is characterized by 7 clusters:

1. (purple) groups concepts related to computer systems in general such as: *"IBM,"* *"Microsoft Windows,"* *"Operating system,"* and *"Linux distribution"*;
2. (green) covers the political side with concepts such as *"Hacktivism"*, *"Cyberwarfare"* *or* *"Cyberterrorism"*, and relevant actors including people, companies, public agencies and states: *"United States,"* *"Russia,"* *"Cyberwarfare in China,"* *"Wikileaks,"* *"Anonymous (group),"* *"Edward Snowden,"* *"Verizon Communications,"* *"Facebook,"* "LinkedIn," *"Federal Bureau of Investigation,"* and *"FBI–Apple encryption dispute"*;
3. (cyan) includes cybersecurity threats and related concepts such as: *"Trojan horse (computing),"* *"Phishing,"* *"Malware,"* *"Ransomware,"* *"Denial of service attack,"* *"Botnet,"* and *"Antivirus software"*;
4. (orange) groups terms related to defense techniques and technologies to protect against cybersecurity threats: *"Encryption,"* *"Transport Layer Security,"* *"Password,"* *"Authentication,"* *"Smart card,"* "One-time password," and *"RSA (cryptosystem)"*;
5. (maroon) comprises concepts related to telecommunications and networks: *"Internet,"* *"Computer network,"* *"Firewall,"* *"MAC address,"* *"Wi-Fi,"* and *"Proxy server"*;
6. (magenta) groups concepts related to consumer electronics, companies and consumer software like operating systems and Internet browsers: *"Microsoft,"* *"Internet Explorer,"* *"Windows XP,"* *"Windows 7,"* *"Apple Inc.,"* *"Google,"* and *"Smartphone"*;
7. (dark green) groups universities, colleges, and other concepts relate to academia: *"Massachusetts Institute of Technology,"* *"Carnegie Mellon University,"* *"Turing Award,"* *"Bell Labs,"* and *"Cornell University"*.

A.3.1 LONGITUDINAL ANALYSIS

Tables 33 and **??** present the results of the longitudinal analysis over the years from 2002 to 2018. The most remarkable result is that until 2007 computer security was centered primarily around computer viruses and

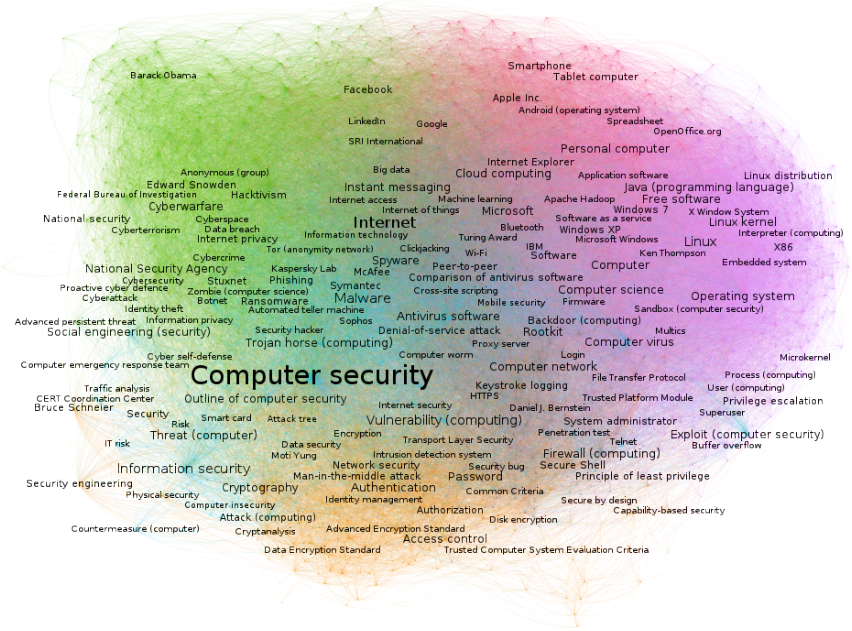


Figure 19: Graph induced by the nodes with non-zero *CycleRank* score with reference node $r =$ *"Computer security"* and $K = 4$ on English Wikipedia, over the snapshot of March 1st, 2018. Colors represents clusters calculated with the Louvain algorithm. The dimension of the nodes and their labels depends on the *CycleRank* score.

ways to contain them, such as *"Trusted system"*. Since 2008 onwards the most prominent topic connected to computer security was *"Internet"* reflecting how the internet as become the channel through which computer threats are propagated. In the same sense, we highlight the appearance of Facebook in 2016 and 2017.

A.3.2 CROSS-LANGUAGE ANALYSIS

The article related to *"Computer security"* appears, albeit in different forms and with difference nuances, in all the 9 languages studied. The results of the cross-language analysis are presented in Tables **??** and **??**. The results cover a wide list of terms related to computer security, without any specific term appearing with higher frequency across languages.

We can see how different language editions give prominence to different concepts; for example, *"Hacker"* is prominent in Spanish, and in the 8th position in Russian; *"Cryptography"* appears in German and Spanish, and has a prominent position in Polish; operating systems appear in prominent position in the Swedish Wikipedia, while the top elements in the ranking for the French edition are mostly very generic concepts.

rank	2002	2003	2004	2005	2006	2007	2008	2009	2010
1	Computer security	Computer security	Computer security	Computer security	Computer security	Computer security	Computer security	Computer security	Computer security
2	Computer science	Security engineering	Computer insecurity	Computer insecurity	Internet	1995	Internet	Computer	Computer
3	Computing	Computer insecurity	Security engineering	Microsoft	Cryptography	Cryptography	Cryptography	Internet	Internet
4	Mathematics	Secure cryptoprocessor	Computing	Cryptography	Microsoft Windows	Microsoft Windows	Computer insecurity	Cryptography	Operating system
5	Cracker	Cryptography	Computer science	Hacker	Computer insecurity	Computer insecurity	Operating system	Microsoft Windows	Computer insecurity
6	Microsoft	Defensive programming	Buffer overflow	Security engineering	Computer	Operating system	Computer	Operating system	Cryptography
7	One-time pad	Trusted system	1995	Computing	Microsoft	Computer	Microsoft Windows	Computer insecurity	Information security
8	Bruce Schneier	Multics	1970s	1995	Computing	Microsoft	Computing	Computer science	Password
9	Computer system	Microsoft	Federal Standard 1037C	Buffer overflow	Buffer overflow	Computer virus	Computer virus	Computing	Computer science
10	Hacking	1970s	Security	Computer science	Computer virus	FreeBSD	Buffer overflow	Password	Buffer overflow

Table 33: Top 10 most relevant concepts by *CycleRank* score with respect to *"Computer security,"* along different yearly snapshots of the WikiLinkGraphs dataset (2002-2010).

rank	2011	2012	2013	2014	2015	2016	2017	2018
1	Computer security	Computer security	Computer security	Computer security	Computer security	Computer security	Computer security	Computer security
2	Computer	Computer	Computer	Internet	Microsoft	Internet	Internet	Internet
3	Internet	Internet	Internet	Computer	Internet	Malware	Computer	Information security
4	Information security	Operating system	Information security	Information security	Computer	IBM	Malware	Malware
5	Operating system	Microsoft	Microsoft	Microsoft	IBM	Computer virus	Computer virus	Vulnerability (computing)
6	Computer insecurity	Information security	Operating system	Operating system	Malware	Information security	Information security	Linux
7	Vulnerability (computing)	Vulnerability (computing)	Vulnerability (computing)	Vulnerability (computing)	Computer virus	Linux	Facebook	Computer virus
8	Cryptography	Computer insecurity	Computer insecurity	IBM	Information security	Operating system	Linux	Computer
9	Malware	Malware	IBM	Computer insecurity	Operating system	Vulnerability (computing)	Vulnerability (computing)	Firewall (computing)
10	Password	IBM	Computer virus	Computer virus	Linux	Facebook	Operating system	Authentication

Table 34: Top 10 most relevant concepts by *CycleRank* score with respect to *"Computer security,"* along different yearly snapshots of the WikiLinkGraphs dataset (2011-2018).

rank	de	en	es	fr	it	nl	pl	ru	sv
1	Informationssicherheit	Computer security	Seguridad informática	Sécurité des systèmes d'information	Sicurezza informatica	Computerbeveiliging	Bezpieczeństwo teleinformatyczne	Komp'juternaja bezopasnost'	Datasäkerhet
2	Computer	Internet	Hacker	Sécurité des données	Internet	Computervirus	Informatyka	Internet	Internet Explorer
3	Datenschutz	Information security	Malware	Logiciel	Sistema operativo	Antimalwaresoftware	Kryptologia	Sistemnyj administrator	Windows Vista
4	Netzwerkprotokoll	Malware	Virus informático	Sécurité	Informatica	Computerkraker	Zapora sieciowa	Informacionnaja bezopasnost'	Linux
5	Computerwurm	Vulnerability (computing)	Agujero de seguridad	Système d'information	Server	Informatica	Linux	Vsemirnaja pautina	Internet
6	Schadprogramm	Linux	Software libre	Fuite d'information	Antivirus	Antivirussoftware	Haker (bezpieczeństwo komputerowe)	Windows NT	Brandvägg
7	Personal Firewall	Computer virus	Criptografía	Politique de sécurité du système d'information	Firewall	Internet Explorer	Administrator (informatyka)	Celostnost' informacii	Mac OS X
8	Kryptographie	Computer	Exploit	Sécurité informatique	Malware	Malware	GSM	Haker	Datavetenskap
9	Wireless Local Area Network	Firewall (computing)	Cortafuegos (informática)	Vulnérabilité (informatique)	Autenticazione	Informatiebeveiliging	Federal Information Processing Standard	Bezopasnost'	Antivirusprogram
10	IT-Sicherheitsaudit	Authentication	Unix	Capital immatériel	Software	Internet	Poczta elektroniczna	Dostupnost' informacii	Dator

Table 35: Top 10 most relevant concepts by *CycleRank* score with respect to *"Computer security,"* across different Wikipedia language editions.

A.4 GREEN COMPUTING

Figure 20 represents the network of articles around the article *"Green computing"* with *CycleRank* score greater than zero. We see that this network is characterized by 6 clusters:

1. (purple) groups concepts related to the environment and energy efficiency: *"Green computing,"* *"Restriction of Hazardous Substances Directive,"* *"Efficient energy use,"* *"Electronic waste,"* *"United States Environmental Protection Agency,"* *"Energy conservation,"* and *"Sustainability"*;
2. (green) includes concepts related to consumer electronics and companies: *"Desktop computer,"* *"Personal computer,"* *"Dell,"* *"Hewlett-Packard,"* and *"Hardware"*;
3. (cyan) groups concepts related to electronics manufacturer and components: *"Intel,"* *"Computer,"* *"Power management,"* *"Central Processing Unit,"* *"Graphics Processing Unit,"* and *"Low-power electronics"*;
4. (orange) includes concepts related to computer software, software companies and software aspects of power management: *"Operating system,"* *"Linux,"* *"Microsoft Windows,"* *"Sleep mode,"* *"Hibernation (computing),"* *"Sun Microsystem,"* and *"Microsoft"*;
5. (dark green) includes big hardware manufacturers and supercomputers: *"IBM,"* *"Supercomputer,"* *"Performance per watt,"* and *"Mainframe computer"*;
6. (magenta) groups concepts related to data centers, the Internet and business aspects related to green computing: *"Server farm,"* *"Data center,"* *"Internet,"* and *"Business Intelligence"*.

A.4.1 LONGITUDINAL ANALYSIS

Tables 36 and 37 present the results for the longitudinal analysis for the keyword *"Green computing"* from 2006 to 2018. Over the course of the years, the top-10 results with the highest *CycleRank* score are mostly terms related to power consumption of personal computers (*"Power management"*, *"Energy Star"*, *"Sleep mode"*, and *"Standby power"*). We note the appearance from 2012 onwards of the article *"Supercomputer"* and, at the same time, the presence of the chip-manufacturing company *"Intel"*. We attribute this change to the fact that the context of the topic "Green computing" has shifted from a personal computer setting to a cloud infrastructure ("Software As A Service"), where the performance and the environmental impact of servers providing cloud services has

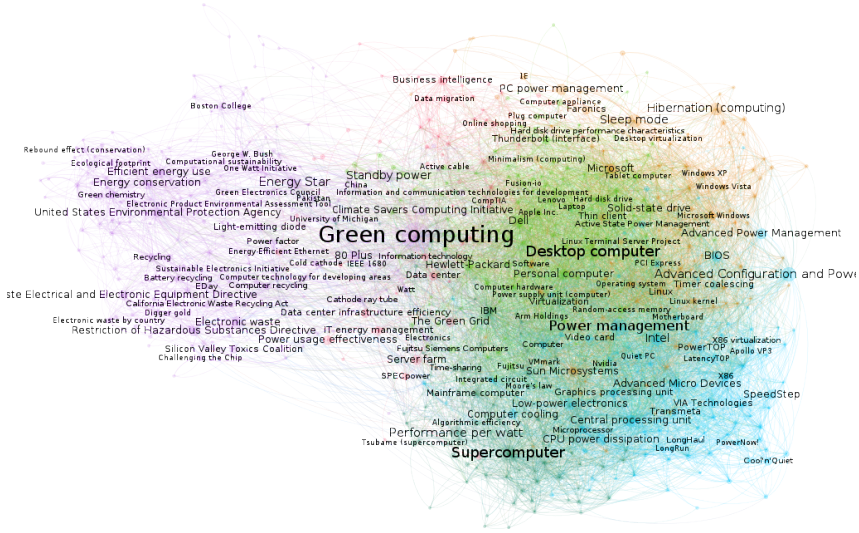


Figure 20: Graph induced by the nodes with non-zero *CycleRank* score with reference node r = *"Green computing"* and $K = 4$ on English Wikipedia, over the snapshot of March 1st, 2018. Colors represents clusters calculated with the Louvain algorithm. The dimension of the nodes and their labels depends on the *CycleRank* score.

gained prominence. In fact, we note the appearance in 2018 of the page *"Performance per watt"*, which is a major issue in the discussion about cloud services and IoT devices.

A.4.2 CROSS-LANGUAGE ANALYSIS

The cross language comparison of results is presented in Table38. Only the German, English, French, Italian, and Spanish editions of Wikipedia have an article dedicated to *"Green computing"*. We see that in German, English, and Spanish Wikipedia the article about *"Green computing"* has strong ties with the context of cloud computing with several pages related to it appearing in the top-10 results; in German: *"Rechenzentrum"* (*"Data center"*), *"Thin Client,"* *"Virtualisierung"* (*"Virtualisation,"*) *"Server,"* and *"Cloud Computing;"* in English: *"Supercomputer,"* *"Performance per watt,"* and *"Intel;"* in Spanish: *"Virtualización"* (*"Virtualisation,"*) *"Apple,"* *"Google,"* and *"Cliente liviano"* (*"Thin client"*).

Italian Wikipedia has a decidedly more environmental focus with prominent articles being: *"Ambientalismo"* (*"Environmentalism,"*) *"Rifiuti di apparecchiature elettriche ed elettroniche"* (*"Electronic waste,"*) *"Riciclaggio dei rifiuti"* (*"Recycling,"*) and *"Normativa comunitaria RoHS"* (*"RoHS Directive"*, a EU directive that restricts the use of six hazardous materials in the manufacture of various types of electronic and electrical equipment).

French Wikipedia has, instead, a more general focus with *"Développement durable"* (*"Sustainable development,"*) *"Cloud computing,"* *"Loi the Moore"* (*"Moore's Law,"*) *"Entreprise"* (*"Enterprise,"*) and *"Technologies de l'information et de la communication"* (*"Information and Communication Technologies"*).

A.5 INTERNET PRIVACY

Figure 21 represents the network of articles around the article *"Internet privacy"* with *CycleRank* score greater than zero. We see that this network is characterized by 7 clusters:

1. (purple) consists of social networking services and companies, together with major mobile platforms (notably, Google is not in this cluster): *"Facebook,"* *"Twitter,"* *"Reddit,"* *"Youtube,"* *"Apple Inc.,"* *"iPhone,"* *"Android (operating system),"* and *"iOS"*;

2. (green) groups concepts related to the WWW and its technologies, and web browsers (notably, Microsoft is included in this cluster): *"World Wide Web,"* *"HTTP cookie,"* *"Web browser,"* *"Adobe Flash Player,"* *"HTML,"* *"Javascript,"* *"Google Chrome,"* *"Mozilla,"* and *"Firefox"*;

3. (cyan) includes terms related mostly to U.S. politics and institutions, with a major focus on former president Barack Obama: *"Barack Obama,"* *"Presidency of Barack Obama,"* *"Supreme Court of the United States,"* and *"United States Senate Committee on the Judiciary"*;

4. (orange) groups technology for privacy protection on the Internet (the term Internet itself is included in this cluster): *"Tor (anonymity network)"*, *"Anonymous P2P"*, *"Proxy server,"* and *"Internet censorship"*;

5. (maroon) collects Google and other web search engines: *"Google,"* *"Web search engine,"* *"DuckDuckGo,"* and *"Bing"*;

6. (magenta) includes the concept of Internet privacy itself together with surveillance, so it includes: *"Privacy,"* *"Mass surveillance,"* *"PRISM (surveillance program),"* *"Edward Snowden,"* and *"National Security Agency"*;

7. (dark green) groups computer security topics and cybersecurity threats: *"Computer security,"* *"Malware,"* *"Trojan horse (computing),"* *"Password strength,"* and *"Spyware"*;

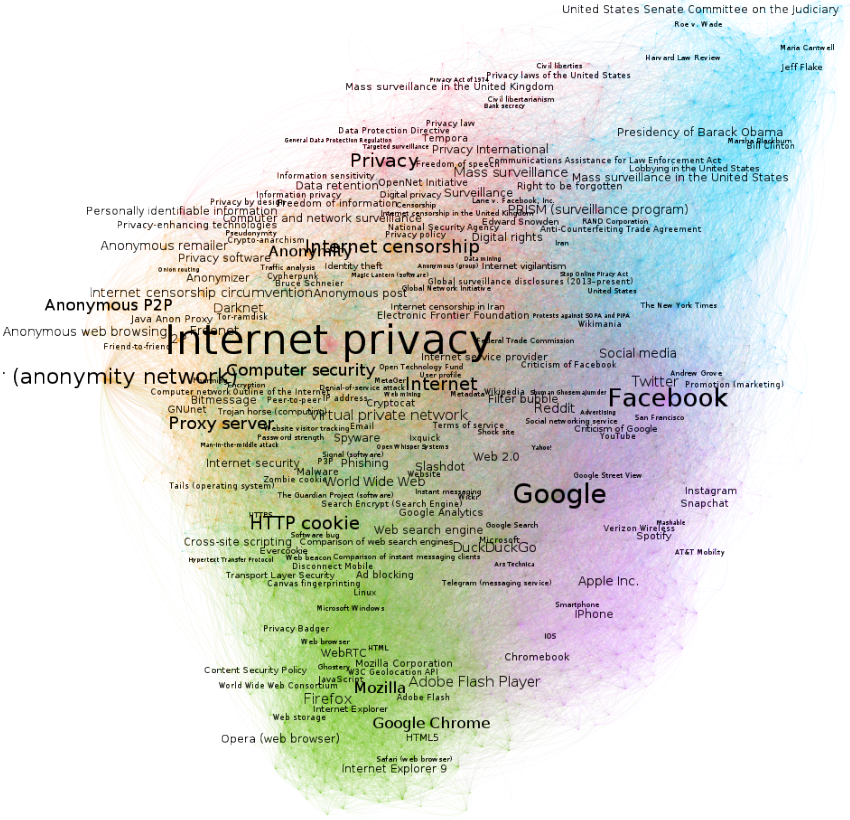


Figure 21: Graph induced by the nodes with non-zero *CycleRank* score with reference node $r =$ *"Internet privacy"* and $K = 4$ on English Wikipedia, over the snapshot of March 1st, 2018. Colors represents clusters calculated with the Louvain algorithm. The dimension of the nodes and their labels depends on the *CycleRank* score.

A.5.1 LONGITUDINAL ANALYSIS

The longitudinal analysis for the keyword *"Internet privacy"* is presented in Tables 23 and 24. This is one of the articles whose evolution over time most reflected the changes in perception towards the topic in the public discourse. We see that over time, from 2005 to 2018, some themes remain at the forefront of *"Internet privacy"* such as *"Anonymity," "Proxy server," "HTTP cookie," "Internet censorship,"* and *"Computer security"*; these topics are closely related to internet privacy. The anonymity network "Tor"[1] is also relevant though the whole period under consideration, but we stress how the article rose to second position, i.e. the most relevant article with respect to *"Internet privacy"* in the period 2013-2016; we think that this could be in part due to coverage of the revelations of Edward Snowden about the U.S: National Security Agency and the role that Tor played in allowing the safe transfer of leaked documents to journalists. However,

1 *https://en.wikipedia.org/wiki/Tor_(anonymity_network)*

rank	2006	2007	2008	2009	2010	2011	2012
1	Green computing	Green computing	Green computing	Green computing	Green computing	Green computing	Green computing
2	Sleep mode	Power management	Power management	Hard disk drive	Hard disk drive	Power management	Supercomputer
3	Power management	Electronic waste	Restriction of Hazardous Substances Directive	Power management	Desktop computer	Desktop computer	Hard disk drive
4	Green chemistry	Sleep mode	Advanced Configuration and Power Interface	Electronic waste	Power management	Hard disk drive	Desktop computer
5	Energy Star	Energy Star	Electronic waste	Power supply unit (computer)	Energy Star	Electronic waste	Power management
6	Advanced Configuration and Power Interface	Advanced Configuration and Power Interface	Waste Electrical and Electronic Equipment Directive	Energy Star	Performance per watt	Restriction of Hazardous Substances Directive	Advanced Configuration and Power Interface
7		Advanced Power Management	Advanced Power Management	Thin client	Advanced Configuration and Power Interface	Advanced Configuration and Power Interface	Cloud computing
8		Restriction of Hazardous Substances Directive	Sleep mode	Standby power	Power supply unit (computer)	Power supply unit (computer)	Electronic waste
9		Recycling	BIOS	Netbook	Standby power	Performance per watt	Energy Star
10		Green chemistry	PowerTOP	Advanced Configuration and Power Interface	Dell	Energy Star	Intel

Table 36: Top 10 most relevant concepts by *CycleRank* score with respect to *"Green computing,"* along different yearly snapshots of the WikiLinkGraphs dataset (2006-2012).

rank	2013	2014	2015	2016	2017	2018
1	Green computing	Green computing	Green computing	Green computing	Green computing	Green computing
2	Supercomputer	Supercomputer	Supercomputer	Supercomputer	Desktop computer	Desktop computer
3	Desktop computer	Desktop computer	Desktop computer	Desktop computer	Supercomputer	Supercomputer
4	Power management	Cloud computing	Power management	Power management	Power management	Power management
5	Cloud computing	Power management	Advanced Configuration and Power Interface	Advanced Configuration and Power Interface	Advanced Configuration and Power Interface	Energy Star
6	Energy Star	Advanced Configuration and Power Interface	Intel	Intel	Energy Star	Advanced Configuration and Power Interface
7	Advanced Configuration and Power Interface	Intel	Energy Star	Energy Star	Intel	Performance per watt
8	Performance per watt	Energy Star	Performance per watt	Performance per watt	Performance per watt	Intel
9	Electronic waste	Performance per watt	Electronic waste	Sleep mode	Sleep mode	Sleep mode
10	Standby power	Electronic waste	Sleep mode	Hibernation (computing)	Hibernation (computing)	Standby power

Table 37: Top 10 most relevant concepts by *CycleRank* score with respect to *"Green computing,"* along different yearly snapshots of the WikiLinkGraphs dataset (2013-2018).

rank	de	en	es	fr	it
1	Green IT	Green computing	Green computing	Informatique durable	Green computing
2	Hardware	Desktop computer	Fit PC	Développement durable	Ambientalismo
3	Informationstechnik	Supercomputer	Linutop	Système d'information	Rifiuti di apparecchiature elettriche ed elettroniche
4	Rechenzentrum	Power management	Virtualización	Informatique	2007
5	Greenpeace	Energy Star	Apple	Cloud computing	Energy Star
6	Thin Client	Advanced Configuration and Power Interface	Google	Loi de Moore	Italia
7	Virtualisierung (Informatik)	Performance per watt	LTSP	Gestion des données de référence	Organizzazione non a scopo di lucro
8	Erneuerbare Energien	Intel	Cliente liviano	Entreprise	Riciclaggio dei rifiuti
9	Server	Sleep mode	MacBook Pro	Technologies de l'information et de la communication	Effetto rimbalzo (economia)
10	Cloud Computing	Standby power		Fracture numérique (géographique)	Normativa comunitaria RoHS

Table 38: Top 10 most relevant concepts by *CycleRank* score with respect to *"Green computing,"* across different Wikipedia language editions.

"Tor" the article has already a prominent position since 2009. Finally, we highlight the jump in relevance of the article about *"Google"* in 2017 and 2018; the company did already appear in the top-10 in 2009-2011. Facebook appears in 3rd position in 2018. The latter company appears consistently in the top-20 since 2011 (included).

A.5.2 CROSS-LANGUAGE ANALYSIS

Table 25 presents the results of the cross-language analysis for the article *"Internet privacy."* We see that the concept of internet privacy has different nuances across cultures, even if all the results are related even among languages. We see that in German Wikipedia, *"Internet privacy"* is strongly related to data and information privacy with articles such as *"Vorratsdatenspeicherung"* (*"Data retention,"*) *"Datenschutz"* (*"Information privacy"*), and *"Überwachung"* (*"Surveillance."*)

In Spanish Wikipedia, we see a novel rendition of the browser wars[2] with *"Mozilla Firefox"* gaining a more prominent position with respect to *"Internet Explorer"*, together with *"Cookie."*

French Wikipedia features both locally-relevant articles such as *"Commission nationale de l'informatique et des libertés (France)"* (*"National commission for Informatics and liberties,"*)[3] and *"Fichage en France"* ("Filing in France", in the sense of the collection by public French authorities of data about French citizens usually in the context of law enforcement), and global-focused articles such as *"Révélations d'Edward Snowden"* (*"Edward Snowden revelations,"*) and *"Surveillance globale"* (*"Global surveillance."*) Finally, there are more general articles like *"Données personnelles"* (*"Personal data,"*) *"Smartphone,"* *"Fuite d'information"* (*"Information leak,"*) and *"Tor."*

In Italian Wikipedia the context of *"Internet privacy"* is mostly oriented towards computer security with broad terms such as: *"Malware,"* *"File,"* *"Antivirus,"* and *"Browser."*

A.6 NET NEUTRALITY

Figure 22 represents the network of articles around the article *"Net neutrality"* with *CycleRank* score greater than zero. We see that this network is characterized by 4 main clusters:

2 *https://en.wikipedia.org/wiki/Browser_wars*

3 The Commission is an independent administrative authority established in France in 1978 and charged with the oversight of technology and its impact on the rights of French citizens (*https://fr.wikipedia.org/wiki/Commission_nationale_de_l%27informatique_et_des_libert%C3%A9s_(Fra*

rank	2005	2006	2007	2008	2009	2010	2011
1	Internet privacy	Internet privacy	Internet privacy	Internet privacy	Internet privacy	Internet privacy	Internet privacy
2	Anonymity	Internet	Internet	Internet	Internet	Internet	Proxy server
3	Proxy server	Anonymity	Anonymity	Proxy server	Proxy server	Tor (anonymity network)	Privacy
4	Internet censorship	Proxy server	HTTP cookie	Computer security	Tor (anonymity network)	Proxy server	Tor (anonymity network)
5	Anonymous remailer	HTTP cookie	Computer security	HTTP cookie	HTTP cookie	Privacy	HTTP cookie
6	Freenet	World Wide Web	Yahoo!	Anonymity	Privacy	HTTP cookie	Anonymity
7	Friend-to-friend	Freenet	Spyware	Spyware	Computer security	Anonymity	Computer security
8	Cyberspace	Tor (anonymity network)	Microsoft	Opera (web browser)	Anonymity	Computer security	Spyware
9	Spamming	Spyware	World Wide Web	Tor (anonymity network)	Google	Spyware	Google
10	Pseudonymity	IP address	AOL	Yahoo!	Spyware	Google	Firefox

Table 39: Top 10 most relevant concepts by *CycleRank* score with respect to *"Internet privacy,"* along different yearly snapshots of the WikiLinkGraphs dataset (2005-2011).

rank	2012	2013	2014	2015	2016	2017	2018
1	Internet privacy	Internet privacy	Internet privacy	Internet privacy	Internet privacy	Internet privacy	Internet privacy
2	Firefox	Tor (anonymity network)	Tor (anonymity network)	Tor (anonymity network)	Tor (anonymity network)	Google	Google
3	Proxy server	Firefox	Proxy server	HTTP cookie	Proxy server	Tor (anonymity network)	Facebook
4	HTTP cookie	HTTP cookie	HTTP cookie	Privacy	HTTP cookie	National Security Agency	Tor (anonymity network)
5	Tor (anonymity network)	Proxy server	Privacy	Proxy server	Privacy	Privacy	Privacy
6	Privacy	Privacy	Computer security	Computer security	Internet censorship	HTTP cookie	Internet censorship
7	Computer security	Anonymity	Internet censorship	Internet censorship	Computer security	Proxy server	HTTP cookie
8	Anonymity	Anonymous P2P	Anonymity	Anonymity	Anonymous P2P	Internet censorship	Internet
9	Spyware	Internet censorship	Anonymous P2P	Anonymous P2P	Anonymity	Internet	Proxy server
10	Internet censorship	Computer security	Adobe Flash Player	Internet	Internet	Computer security	Computer security

Table 40: Top 10 most relevant concepts by *CycleRank* score with respect to *"Internet privacy,"* along different yearly snapshots of the WikiLinkGraphs dataset (2012-2018).

rank	de	en	es	fr	it
1	Datenschutz im Internet	Internet privacy	Privacidad en Internet	Vie privée et informatique	Privacy
2	Vorratsdatenspeicherung	Google	Facebook	Données personnelles	Internet
3	Anonymität im Internet	Facebook	Mozilla Firefox	Commission nationale de l'informatique et des libertés (France)	Anni 1990
4	Internet	Tor (anonymity network)	Internet	Smartphone	Sicurezza informatica
5	Datenschutz	Privacy	Internet Explorer	Révélations d'Edward Snowden	Malware
6	Google	Internet censorship	Cookie (informática)	Fuite d'information	File
7	Rechtsfreier Raum	HTTP cookie	Redes sociales en Internet	Sécurité des données	Diritto
8	Überwachung	Internet	Privacidad	Fichage en France	Antivirus
9	Internetdienstanbieter	Proxy server	Seguridad informática	Tor (réseau)	Browser
10	Google-Konto	Computer security	Navegador web	Surveillance globale	Posta elettronica

Table 41: Top 10 most relevant concepts by *CycleRank* score with respect to *"Internet privacy,"* across different Wikipedia language editions.

1. (purple) groups concepts related to telecommunications and net neutrality as seen in the context of networks: *"Net neutrality,"* *"Internet Service Provider,"* *"Internet access,"* *"Zero-rating,"* *"Telecommunications,"* *"Tom Wheeler,"*[4] *"Ajit Pai,"*[5] *"Verizon Communications,"* and *"AT&T"*;

2. (green) groups concepts related to U.S. politic and other prominent figures, and mainstream media: *"Barack Obama,"* *"Presidency of Barack Obama,"* *"Peter Thiel,"* *"Chelsea Manning,"* *"Edward Snowden,"* *"United States Senate,"* *"The New York Times,"* and *"CNN"*;

3. (cyan) contains concepts mostly related to Internet and social media with digital rights and Internet freedom, together with prominent figures in this regard: *"Internet,"* *"Social media"*, *"World Wide Web,"* *"Facebook,"* *"Freedom of speech,"* *"World Wide Web,"* *"Stop Online Piracy Act,"* *"Aaron Swartz,"* *"Lawrence Lessig,"* and *"Tim Berners-Lee"*;

4. (orange) contains a collection of Internet video streaming services and related topics: *"Netflix,"* *"Comcast,"*[6] *"HBO Go,"* *"Xfinity,"* *"Last Week Tonight with John Oliver,"* and *"Hulu"*;

A.6.1 LONGITUDINAL ANALYSIS

Tables 42 and 43 present the results of the longitudinal analysis for the article *"Net neutrality"* in English Wikipedia. The most remarkable characteristic is that the results over the period 2015-2018 are very stable: at the top we find the former US president *"Barack Obama"*, and the *"Federal Communications Commission"* highlighting the prominence of the presidency of Barack Obama over the Net neutrality regulations in the US and the role of the FCC. We also point out the presence of *"Verizon Communications"* and *"Comcast"*, two of the biggest Internet Service Providers in the United States. Finally, the presence of *"Tim Berners-Lee"* stresses the extent to which the inventor of the World Wide Web has played a significant role in the public discourse about Net neutrality.

4 Former Chairman of the Federal Communications Commission (*https://en.wikipedia.org/wiki/Tom_Wheeler*).

5 Current Chairman of the Federal Communications Commission (*https://en.wikipedia.org/wiki/Ajit_Pai*).

6 Comcast is a telecommunications company, it has ended in this cluster probably because of a well-covered feud with streaming service Netflix over Internet speed throttling (*https://en.wikipedia.org/wiki/Criticism_of_Comcast#Netflix*).

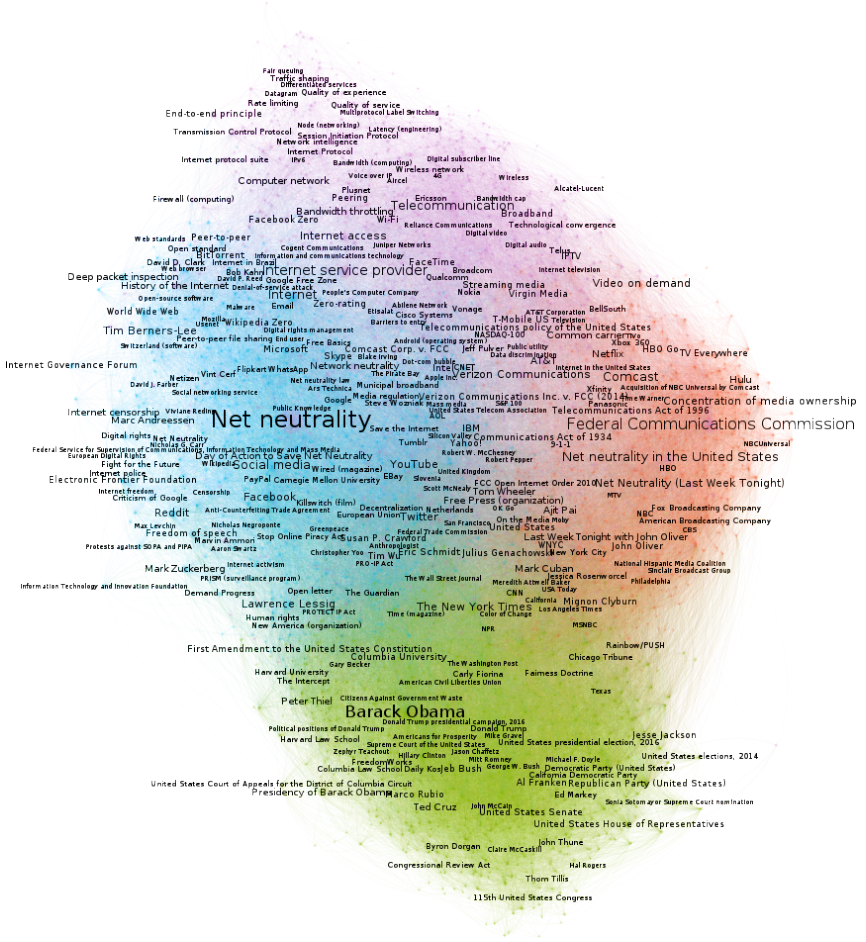


Figure 22: Graph induced by the nodes with non-zero *CycleRank* score with reference node r = *"Net neutrality"* and $K = 4$ on English Wikipedia, over the snapshot of March 1st, 2018. Colors represents clusters calculated with the Louvain algorithm. The dimension of the nodes and their labels depends on the *CycleRank* score.

A.6.2 CROSS-LANGUAGE ANALYSIS

Tables **??** and **??** present the results of the cross-language analysis for the article *"Net neutrality"*. The analysis shows that the debate around net neutrality is dominated by the United States, in fact:

- *"Barack Obama"* appears 3 times;
- the *"Federal Communications Commission"* appear 2 times,
- US tech companies and services (*"Google,"* *"Facebook,"* *"Ebay,"* *"Youtube,"* and *"Skype"*) appear 7 times;

However, there are also local elements in the results, such as:

- In German Wikipedia, several telecommunications companies and Internet service providers appear in the top-10 results:

"Deutsche Telekom," *"Telekom Deutschland,"* *"Vodafone Kabel Deutschland,"* and *"Unitymedia"*;

- In French Wikipedia, different aspects of the net neutrality debate are touched:
  - *"Benjamin Bayart,"*[7] a French engineer, militant for digital rights and president of *French Data Network*[8], the olderst ISPs in France still in business;
  - "La Quadrature du Net,"[9] a French association for the defense of citizens' rights on the Internet, founded in 2008;
  - *"Orange,"*[10] a French telecommunications company;
- In Russian Wikipedia, we have *"Meduza,"*[11] a Riga-based online newspaper and news aggregator written in Russian;

A.7 ONLINE IDENTITY

Figure 23 represents the network of articles around the article *"Online identity"* with *CycleRank* score greater than zero. We see that this network is characterized by 6 clusters:

1. (purple) groups generic terms about users in the context of online services together with some of these services, mostly social networks: *"User profile,"* *"Digital footprint,"* *"Tag (metadata),"* *"Online identity management,"*; and on the service side; *"Facebook,"* *"Twitter,"* *"LinkedIn,"* *"Google+,"* and *"Snapchat"*;
2. (green) is centered around sexual and gender identity: *"Transgender,"* *"Gender,"* *"Minority group,"* *"Gender identity,"* and *"Intersex"*;
3. (cyan) groups concepts related to the internet and privacy in general: *"Internet,"* *"Online chat,"* *"Privacy,"* *"Sockpuppet (Internet),"* and *"Computer security"*;
4. (orange) a cluster with a stronger focus on online participation: *"Avatar (computing),"* *"Internet forum,"* *"Anonymity,"* *"Pseudonymity,"* and *"Virtual community"*;
5. (maroon) contains specific terms related to digital identity such as: *"Online identity,"* *"Digital identity,"* and *"OpenId"*; notably, "Online dating service" is included in this cluster;

7 *https://fr.wikipedia.org/wiki/Benjamin_Bayart*
8 *https://fr.wikipedia.org/wiki/French_Data_Network*
9 *https://fr.wikipedia.org/wiki/La_Quadrature_du_Net*
10 *https://fr.wikipedia.org/wiki/Orange_(entreprise)*
11 *https://ru.wikipedia.org/wiki/Meduza*

rank	2006	2007	2008	2009	2010	2011	2012
1	Net neutrality	Net neutrality	Net neutrality	Net neutrality	Net neutrality	Net neutrality	Net neutrality
2	Vint Cerf	Internet	Internet	Virgin Media	Virgin Media	Virgin Media	Virgin Media
3	Internet	Microsoft	Barack Obama	BitTorrent	Dianne Feinstein	WNYC	WNYC
4	DARPA	2002	United States	Dianne Feinstein	Fiber to the x	Michael F. Doyle	Concentration of media ownership
5	History of the Internet	Google	2007	2008	WNYC	The New York Times	Comcast Corp. v. FCC
6	ICANN	Voice over IP	YouTube	Federal Communications Commission	Concentration of media ownership	Cable television	Michael F. Doyle
7	June 23	Apple Inc.	Google	WNYC	Broadband Internet access	Federal Communications Commission	Anna Eshoo
8	2002	Internet protocol suite	Wi-Fi	New media	Michael F. Doyle	Anna Eshoo	Cliff Stearns
9	Packet switching	Internet Protocol	Yahoo!	Jack Layton	Federal Communications Commission	Edolphus Towns	Etisalat
10	Internet Society	Wi-Fi	2006	Concentration of media ownership	Internet service provider	Concentration of media ownership	Cable television

Table 42: Top 10 most relevant concepts by *CycleRank* score with respect to *"Net neutrality,"* along different yearly snapshots of the WikiLinkGraphs dataset (2006-2012).

rank	2013	2014	2015	2016	2017	2018
1	Tim Berners-Lee	Tim Berners-Lee	Federal Communications Commission	Federal Communications Commission	Federal Communications Commission	Federal Communications Commission
2	Lawrence Lessig	Network neutrality	Telecommunication	Telecommunication	Telecommunication	Internet service provider
3	Net neutrality in the United States	Internet	Internet service provider	Internet service provider	Internet service provider	Net neutrality in the United States
4	Virgin Media	Federal Communications Commission	Internet access	Verizon Communications	Verizon Communications	Telecommunication
5	Anna Eshoo	Net neutrality in the United States	Verizon Communications	Comcast	Comcast	Comcast
6	Julius Genachowski	Lawrence Lessig	Comcast	Internet	Internet	Social media
7	Bandwidth throttling	Virgin Media	Net neutrality in the United States	Social media	Social media	Internet
8	Federal Communications Commission	Peering	United States	Internet access	Tim Berners-Lee	Tim Berners-Lee
9	Tim Berners-Lee	Tim Berners-Lee	Federal Communications Commission	Federal Communications Commission	Federal Communications Commission	Federal Communications Commission
10	Lawrence Lessig	Network neutrality	Telecommunication	Telecommunication	Telecommunication	Internet service provider

Table 43: Top 10 most relevant concepts by *CycleRank* score with respect to *"Net neutrality,"* along different yearly snapshots of the WikiLinkGraphs dataset (2013-2018).

rank	de	en	es	fr	it	nl	pl	ru	sv
1	Netzneutralität	Net neutrality	Neutralidad de red	Neutralité du réseau	Neutralità della rete	Netneutraliteit	Neutralność sieci	Setevoj nejtralitet	Nätneutralitet
2	Internet	Barack Obama	Comcast	Internet	Internet	Mobiel internet	Internet	Internet	Dagens Nyheter
3	Deutsche Telekom	Federal Communications Commission	Partido Pirata	World Wide Web	Barack Obama	Facebook	Stany Zjednoczone	Google (kompanija)	2000-talet
4	Vodafone Kabel Deutschland	Internet service provider	Barack Obama	Union internationale des télécommunications	Banda larga	Internet	Partia Demokratyczna (Stany Zjednoczone)	EBay	Federal Communications Commission
5	Vodafone	Net neutrality in the United States	Neutralidad de la red	Histoire d'Internet	Telecomunicazione	Internetprovider		FTP	Engelska
6	Unitymedia	Telecommunication	Carta abierta	Google	Wi-Fi	Internetcensuur		Telegraf	Internet
7	Telekom Deutschland	Comcast	Lawrence Lessig	La Quadrature du Net	Voice over IP	VoIP via de mobiele telefoon		Meduza	Pakistan
8	IP-Telefonie	Social media	Peering	Benjamin Bayart	YouTube	Vodafone		Institut Katona	USA
9	Paketvermittlung	Internet	2017	Microsoft	Peer-to-peer	YouTube		Cifrovye prava	Body of European Regulators of Electronic Communications
10	Next Generation Network	Tim Berners-Lee	Derechos digitales	Orange (entreprise)	Skype	2008		San-Francisko	Ajit Pai

Table 44: Top 10 most relevant concepts by *CycleRank* score with respect to *"Net neutrality,"* across different Wikipedia language editions.

6. (magenta) groups concepts related to identity in a social sense: *"Reputation," "Identity (social science)," "Impression management," "Persona,"* and *"Identity (philosophy)"*.

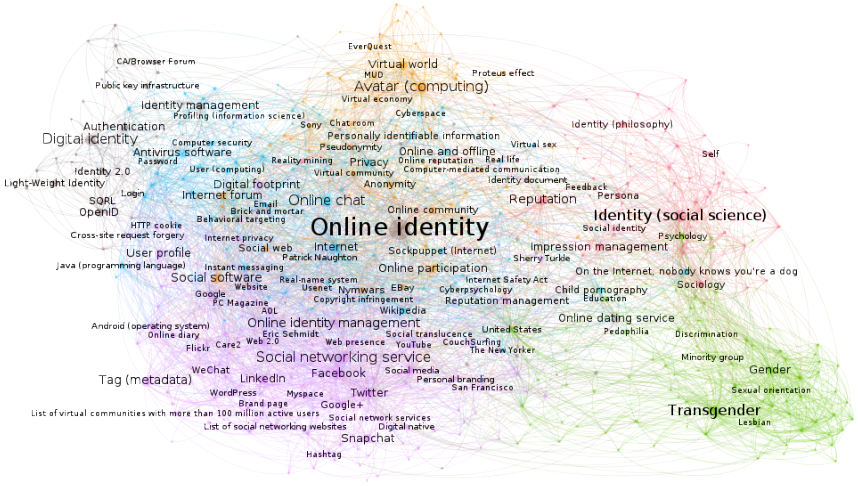


Figure 23: Graph induced by the nodes with non-zero *CycleRank* score with reference node r = *"Online identity"* and $K = 4$ on English Wikipedia, over the snapshot of March 1st, 2018. Colors represents clusters calculated with the Louvain algorithm. The dimension of the nodes and their labels depends on the *CycleRank* score.

A.7.1 LONGITUDINAL ANALYSIS

Tables 45 and 46 present the results of the longitudinal analysis for the article *"Online identity"* in English Wikipedia. The article consistently mostly related over time - since 2016 to 2018 - to *"Online identity"* is *"Reputation"*; we also find articles such as *"Identity,"* and *"Digital identity."* From 2010, the new area of *"Online identity management"* emerges from the results. Another article that appears consistently in the top positions since 2009 is *"Online chat"*. Finally, we see that *"Pseudonymity"* and *"Anonymity"* were present in the period 2006-2008, but they both lost importance subsequently probably due to the rise of social networks and services that discourage anonymity and pseudonymity in favor of users disclosing their real names. It is notable the rise to the second position of *"Transgender,"* highlighting the prominence of the debate about gender and sexual identity even in the context of online identity.

A.7.2 CROSS-LANGUAGE ANALYSIS

Table 47 presents the result of the cross-language analysis for the article *"Online identity."* In French Wikipedia, the topic of online identity

is rendered in the page *"Idéntité numerique"* (lit. *"Digital identity"*), which has a different nuance than *"Online identity"*; in fact the results from French Wikipedia are closer to the ones obtained for *"Internet privacy"*: we see *"Vie privée et informatique,"* (*"Private life and informatics"*), *"Usurpation d'identité"* (*"Identity theft"*), *"Données personnelles"* (*"Personal data"*), *"Commission nationale de l'informatique"* (*"National commission for Informatics and liberties,"*)[12] et des libertés (France)" and *"Biométrie"* (*"Biometry"*).

Finally, In Polish and Russian Wikipedia results are more oriented towards the concept of identity and psychology with *"Tożsamość"* (*"Identity"*), *"Tożsamość (psychologia)"* (*"Identity (psychology)"*), *"Tożsamość osobista"* (*"Personal identity"*), and *"Egosyntoniczność"* (*"Egosyntonic"*[13]) in Polish; and *"Internet-soobshhestvo"* (*"Online community"*), *"Blogosfera"* (*"Blogosphere"*), *"Kiberpsihologija"* (*"Cyberpsycology"*), and *"Vojskunskij, Aleksandr Evgen'evich"* (*"Alexander Evgenievich Voyskunsky"*, a prominent Russian psychologist scholar that studied the impact of the Internet on the human psyche) in Russian.

A.8 OPEN-SOURCE MODEL

Figure 24 represents the network of articles around the article *"Open-source model"* with *CycleRank* score greater than zero. We see that this network is characterized by 6 clusters:

1. (purple) contains terms related to intellectual property: *"Patent,"* *"Tim O'Reilly,"* *"Creative Commons,"* and *"Open content"*;
2. (green) groups terms related to operating systems and related concepts: *"Linux,"* *"Debian,"* and *"Unix"*;
3. (cyan) reflects the dichotomy between open licenses and proprietary software: *"Open source,"* *"Proprietary software,"* *"Microsoft Windows,"* *"Free software,"* *"Bruce Perens,"* and *"Richard Stallman"*;
4. (orange) reflects the dichotomy between open standards and proprietary formats in end-user softwares: *"Application software,"* *"Browser wars,"* *"Android (operating system),"* *"Apple Inc.,"* *"Vendor lock-in,"* and *"Open format"*;
5. (dark green) groups programming languages, formats and middleware: *"Wordpress,"* *"JSON,"* *"Java (programming language),"*

12 See the corresponding footnote in the cross-language analysis for *"Internet privacy."*
13 *https://en.wikipedia.org/wiki/Egosyntonic_and_egodystonic*

rank	2006	2007	2008	2009	2010	2011	2012
1	Online identity	Online identity	Online identity	Online identity	Online identity	Online identity	Online identity
2	Reputation	Reputation	Avatar (computing)	Avatar (computing)	Reputation	Reputation	Reputation
3	Identity	Social software	Reputation	Online chat	Online chat	Online and offline	Online and offline
4	Social software	Digital identity	Social software	Online and offline	Online and offline	Avatar (computing)	Avatar (computing)
5	Anonymity	Virtual community	Digital identity	Reputation	Digital identity	Digital identity	Online chat
6	Pseudonymity	Reputation management	Virtual community	Digital identity	Avatar (computing)	Online chat	Digital identity
7	Pseudonym	Anonymity	Internet forum	Social software	Identity (social science)	Identity (social science)	Identity (social science)
8	Friendster	Pseudonymity	Anonymity	EverQuest	Online identity management	Online identity management	Online identity management
9	Digital identity	Internet forum	World of Warcraft	Virtual community	Social web	Social software	Social software
10	Internet troll	OpenID	Reputation management	Internet forum	Social software	Tag (metadata)	Tag (metadata)

Table 45: Top 10 most relevant concepts by *CycleRank* score with respect to *"Online identity,"* along different yearly snapshots of the WikiLinkGraphs dataset (2006-2012).

rank	2013	2014	2015	2016	2017	2018
1	Online identity	Online identity	Online identity	Online identity	Online identity	Online identity
2	Reputation	Reputation	Reputation	Reputation	Online chat	Transgender
3	Online chat	Online chat	Online chat	Online chat	Identity (social science)	Identity (social science)
4	Online and offline	Online and offline	Online and offline	Identity (social science)	Social networking service	Social networking service
5	Avatar (computing)	Social networking service	Identity (social science)	Social networking service	Digital identity	Avatar (computing)
6	Digital identity	Social software	Social networking service	Online dating service	Facebook	Online chat
7	Identity (social science)	Avatar (computing)	Social software	Social software	Avatar (computing)	Digital identity
8	Tag (metadata)	Identity (social science)	Digital identity	Avatar (computing)	Online dating service	Online identity management
9	Social software	Tag (metadata)	Online identity management	Online identity management	Social software	Social software
10	Online identity management	Digital identity	Tag (metadata)	Digital identity	User profile	Reputation

Table 46: Top 10 most relevant concepts by *CycleRank* score with respect to *"Online identity,"* along different yearly snapshots of the WikiLinkGraphs dataset (2013-2018).

rank	en	fr	pl	ru
1	Online identity	Identité numérique	Tożsamość internetowa	Setevaja identichnost'
2	Transgender	Vie privée et informatique	Tożsamość	Internet-soobshhestvo
3	Identity (social science)	Usurpation d'identité	Pedofilia	Blogosfera
4	Social networking service	Données personnelles	Tożsamość (psychologia)	Vojskunskij, Aleksandr Evgen'evich
5	Avatar (computing)	Sécurité des données	Tożsamość osobista	Kiberpsihologija
6	Online chat	Radio-identification	Egosyntoniczność	
7	Digital identity	Virtuel		
8	Online identity management	Authentification forte		
9	Social software	Commission nationale de l'informatique et des libertés (France)		
10	Reputation	Biométrie		

Table 47: Top 10 most relevant concepts by *CycleRank* score with respect to *"Online identity,"* across different Wikipedia language editions.

"C (programming language)," and *"Python (programming language)"*;

6. (magenta) contains concepts pertaining to open-source hardware: *"Openmoko,"* *"Open-source hardware,"* *"Open-source robotics,"* *"Arduino,"* *"Open design,"* and *"3D printing"*;

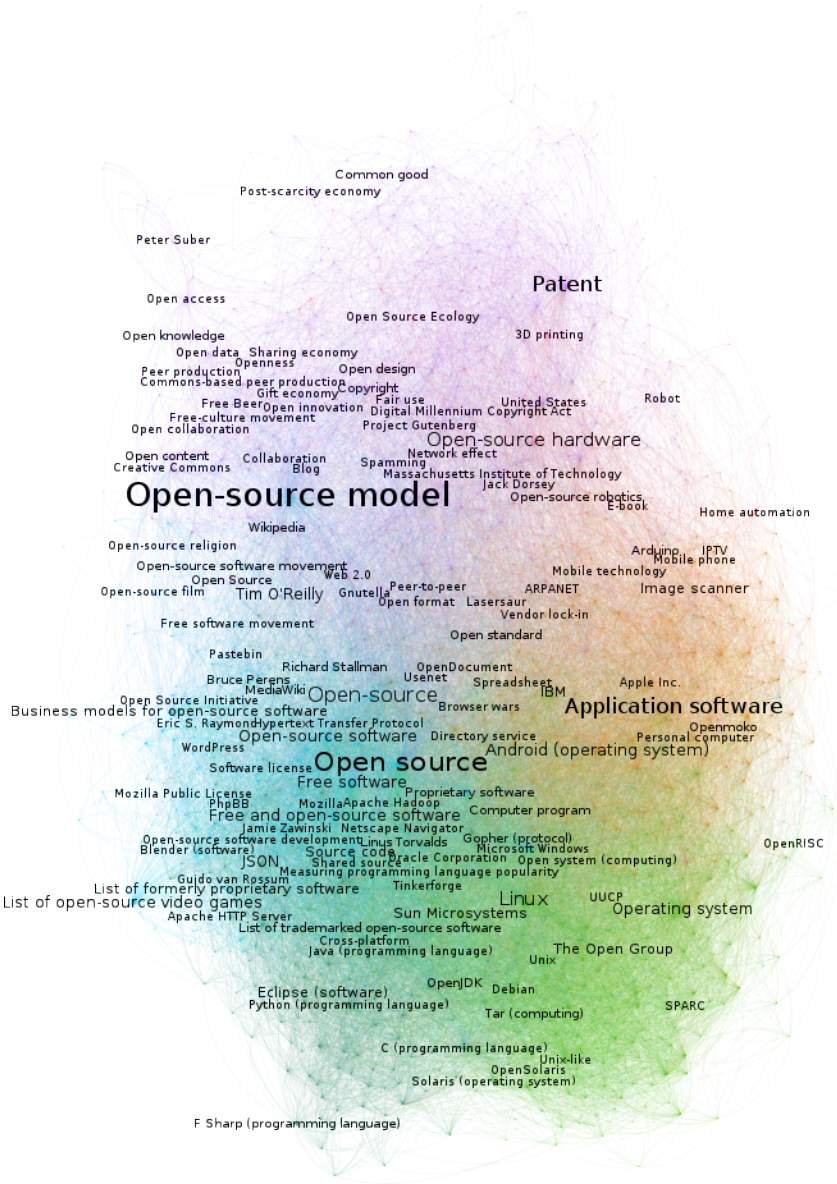


Figure 24: Graph induced by the nodes with non-zero *CycleRank* score with reference node $r =$ *"Online identity"* and $K = 4$ on English Wikipedia, over the snapshot of March 1st, 2018. Colors represents clusters calculated with the Louvain algorithm. The dimension of the nodes and their labels depends on the *CycleRank* score.

A.8.1 LONGITUDINAL ANALYSIS

"Open-source model" was one of the most difficult topics to map. In fact, there exists a whole galaxy of articles centered around *"Open source"*[14] - which is itself a separate article on English Wikipedia - and its related aspect as it is demonstrated by the existence of a disambiguation page[15] called *"Open source (disambiguation)"*. As of this writing, the page *"Open source (disambiguation)"* lists 27 articles, on English Wikipedia alone, containing the keyword *"Open source"* in their title. Note that the aim of this list in not to include all the articles with "open source" in their name, in fact it does not include articles such as *"Business models for open-source software,"*[16] whose topic and title is well-separated from "open source" alone.

The article *"Open-source model"* was created originally on English Wikipedia on October 14th, 2001[17] with the title *"Open source"*. Since this topic is very close to the hearth of the WIkipedian community and Wikipedia describes itself as an "open source" project it is not surprising that this article exists since 2001. The article acquired the current title being moved from *"Open source"* to *"Open-source model"* in March, 2016[18] and the article now-called *"Open source"* refers specifically to software reuse and was created in November, 2018.[19] Note that this article is itself different from *"Open-source software"*[20]. For this reason, we limit the longitudinal analysis over the article *"Open-source model"* to the years 2017 and 2018, the results are presented in Table 48.

From the longitudinal analysis we find different results in the two years under consideration, a sign of the dynamic nature of the links in this article. The articles that are present on both years within the top ten are: *"Linux,"* *"Android (operating system),"*[21] and *"Open-source software."*

A.8.2 CROSS-LANGUAGE ANALYSIS

In selecting the articles corresponding to *"Open-source model"* in the other Wikipedia language editions we faced the same challenges that we encountered on English Wikipedia. Open source is a galaxy in all

14 *https://en.wikipedia.org/wiki/Open_source*
15 A "disambiguation page" in Wikipedia is a page used as for resolving conflicts in article titles that occur when a single term can be associated with more than one topic, making that term likely to be the natural title for more than one article. (https://en.wikipedia.org/wiki/Help:Disambiguation)
16 *https://en.wikipedia.org/wiki/Business_models_for_open-source_software*
17 *https://en.wikipedia.org/w/index.php?title=Open-source_model&oldid=398934973*
18 *https://en.wikipedia.org/w/index.php?title=Open-source_model&oldid=709483345*
19 *https://en.wikipedia.org/w/index.php?title=Open_source&oldid=870042373*
20 *https://en.wikipedia.org/wiki/Open-source_software*
21 The Linux-based operating system for smartphones developed by Google.

the Wikipedias, with multiple articles covering different aspects of the topic. This is a known problem in Wikipedia [49]. With the help of local community members we have chosen the most-fitting articles to pair with *"Open-source model"*, in many cases the choice landed on the local version of the article *"Open source."* Table 49 presents the results of the cross-language analysis. Even though these articles are in principle not only focused on software, and other more specific articles about "Open source software" exist in some languages, we see that software tends to be the main protagonist in all languages. An interesting outlier is the presence of *"Culture libre" ("Free culture")* within the top of the ranking in the French Wikipedia.

A.9 RIGHT TO BE FORGOTTEN

Figure 25 represents the network of articles around the article *"Right to be forgotten"* with *CycleRank* score greater than zero. We see that this network is characterized by 6 clusters:

1. (purple) groups terms related to internet services, which have in comment the fact that they are usually part of the debate about the *"Right to be forgotten"*: *"Wikipedia," "Twitter," "Electronic Frontier Foundation," "Streisand effect," "Censorship by Google,"* and *"Google"*;
2. (green) this cluster is centered around the right to privacy in the United States with articles such as: *"Right to privacy," "United States Constitution," "First Amendment to the United States Constitution,"* and *"Freedom of Speech"*;
3. (cyan) groups concepts related to privacy: *"Privacy," "Privacy law," "Data Protection Directive," "General Data Protection Regulation," "Google Spain v AEPD and Mario Costeja González,"* and *"Internet privacy"*;
4. (orange) groups concepts related to the UK and the European Union, together with media outlets: *"Accountability," "The Guardian," "The New York Times," "Theresa May," "Günther Oettinger,"* and *"Sabine Leutheusser-Schnarrenberger"*;
5. (dark green) groups concepts and states related to media censorship: *"Media freedom in Russia," "China," "Digital rights," "Internet censorship," "Censorship," "Reporters without borders," "Internet access,"* and *"Virtual Private Network"*;
6. (magenta) groups concepts related to human rights: *"Human rights," "Universal Declaration of Human Rights,"* and *"Fundamental rights"*.

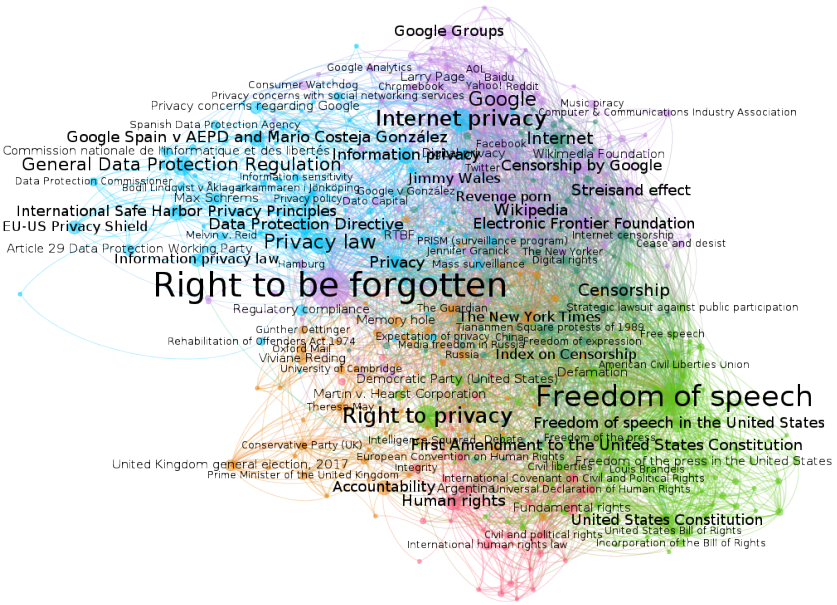


Figure 25: Graph induced by the nodes with non-zero *CycleRank* score with reference node $r =$ *"Right to be forgotten"* and $K = 4$ on English Wikipedia, over the snapshot of March 1st, 2018. Colors represents clusters calculated with the Louvain algorithm. The dimension of the nodes and their labels depends on the *CycleRank* score.

rank	2017	2018
1	Open-source model	Open-source model
2	Linux	Open source
3	Operating system	Application software
4	Open-source software	Patent
5	Free software	Open-source
6	Free and open-source software	Open-source hardware
7	Sun Microsystems	Linux
8	Android (operating system)	Android (operating system)
9	Berkeley Software Distribution	Open-source software
10	IBM	Operating system

Table 48: Top 10 most relevant concepts by *CycleRank* score with respect to *"Open-source model,"* along different yearly snapshots of the WikiLinkGraphs dataset (2017-2018).

rank	de	en	es	fr	it	nl	pl	sv
1	Open Source	Open-source model	Código abierto	Open source	Open source	Open source	Otwarte oprogramowanie	Öppen källkod
2	Linux	Open source	Software libre	Logiciel libre	Software libero	Linux	Linux	Linux
3	Freie Software	Application software	GNU/Linux	Culture libre	Informatica	Lijst van opensourcesoftware	Microsoft Windows	Fri programvara
4	Microsoft	Patent	Unix	Android	Unix	Vrije software	Microsoft	Unix
5	Ubuntu	Open-source	Ubuntu	Debian	Linux (kernel)	Opensourcesoftware	Firefox	Operativsystem
6	Fedora (Linux-Distribution)	Open-source hardware	Android	Google	Browser	Ubuntu (Linuxdistributie)	Wolne oprogramowanie	Källkod
7	LibreOffice	Linux	Internet	Linux	Software	Technologie	Unix	FreeBSD
8	Linux (Kernel)	Android (operating system)	1998	Système d'exploitation	Sistema operativo	Mozilla Firefox	Ubuntu	Mac OS X
9	Software	Open-source software	Richard Stallman	Logiciel	Android	PHP	GNU General Public License	Mozilla Firefox
10	Open-Source-Software in öffentlichen Einrichtungen	Operating system	IBM	Apple	Linux	Software	Kod źródłowy	Google

Table 49: Top 10 most relevant concepts by *CycleRank* score with respect to *"Open-source model,"* across different Wikipedia language editions.

A.9.1 LONGITUDINAL ANALYSIS

The article was created after March 2014. Since the beginning, in 2015, *"Freedom of speech"* is consistently the most relevant concept over years; while it is straightforward why *"Freedom of speech"* is mentioned in the article, given the concerns about how to conciliate it with the right to be forgotten, it is interesting that a link existed also in the opposite direction since 2014. Table 31 presents the results of the longitudinal analysis over the article *"Right to be forgotten"* in English Wikipedia.

A.9.2 CROSS-LANGUAGE ANALYSIS

The cross language analysis of the article *"Right to be forgotten"* produces two main results: in some languages - German, English, Dutch and Russian - the context of this article is mainly composed by articles referring to the Internet, to Google, to Wikipedia and other internet-related pages such as *"Internet censorship"*. Instead in Spanish, Italian and French Wikipedia the focus is more centered around privacy and laws to protect privacy. Freedom of speech is a keyword that appears across several languages even if with different nuances: *"Libertad de expresión"* in Spanish and *"Libertà di manifestazione del pensiero"* in Italian ("Freedom of expression"), *"Svoboda informacii"* and *"Vrijheid van informatie"* ("Freedom of information") in Russian and Dutch, respectively. References to the European *"General Data Protection Regulation"* appear across different languages. Finally, it is notable to point out that the 10th most-related article in Russian Wikipedia related to *"Right to be forgotten"* is *"Gomoseksual'nost"'* (*"Homosexuality"*).

A.10 GENERAL DATA PROTECTION REGULATION (GDPR)

Figure 26 represents the network of articles around the article *"General Data Protection Regulation"* with *CycleRank* score greater than zero. We see that this network is characterized by 6 clusters:

1. (purple) groups articles related to privacy: *"Data Protection Directive,"* *"Privacy law,"* and *"Privacy"*;
2. (green) includes articles related to the *"Right to be forgotten"*: *"Right to be forgotten,"* *"Jan Philipp Albrecht"* and *"Google Spain v AEPD and Mario Costeja González"*; notably, the article *"Mass surveillance"* is also in this cluster;

rank	2015	2016	2017	2018
1	Right to be forgotten	Right to be forgotten	Right to be forgotten	Right to be forgotten
2	Freedom of speech	Freedom of speech	Freedom of speech	Freedom of speech
3	First Amendment to the United States Constitution	Internet privacy	Internet privacy	Right to privacy
4	Freedom of speech in the United States	First Amendment to the United States Constitution	Right to privacy	Internet privacy
5	Censorship	Censorship	Privacy law	Privacy law
6	Human rights	Freedom of speech in the United States	First Amendment to the United States Constitution	Google
7	United States Constitution	Human rights	Freedom of speech in the United States	General Data Protection Regulation
8	Criticism of Google	Privacy law	Censorship	Internet
9	Streisand effect	Streisand effect	Internet	Censorship
10	Internet privacy	Internet	Human rights	Information privacy

Table 50: Top 10 most relevant concepts by *CycleRank* score with respect to *"Right to be forgotten,"* along different yearly snapshots of the WikiLinkGraphs dataset (2017-2018).

rank	de	en	es	fr	it	nl	ru
1	Recht auf Vergessenwerden	Right to be forgotten	Derecho al olvido	Droit à l'oubli	Diritto all'oblio	Recht om vergeten te worden	Pravo na zabvenie
2	Jimmy Wales	Freedom of speech	Revista Latinoamericana de Protección de Datos Personales	Données personnelles	Privacy	Internet	Neprikosnovennost' chastnoj zhizni
3	Sabine Leutheusser-Schnarrenberger	Right to privacy	Derechos	Directive 95/46/CE sur la protection des données personnelles	Chat	Internet van A tot Z	Internet-cenzura
4	Wikipedia	Internet privacy	Unión Europea	Vie privée et informatique	Codice in materia di protezione dei dati personali	Google	Zashhita personal'nyh dannyh
5	Google	Privacy law	Protección de datos personales	Règlement général sur la protection des données	Garante per la protezione dei dati personali	Privacy	Svoboda informacii
6	EU-Datenschutzreform	Google	Dirección Nacional de Protección de Datos Personales	Vie privée	Trattamento dei dati personali	Vrijheid van informatie	Internet
7	Wikimedia Foundation	General Data Protection Regulation	Constitución	G29	Internet		Cenzura
8	Datenschutz-Grundverordnung	Internet	Libertad de expresión	Commission nationale de l'informatique et des libertés (France)	Libertà di manifestazione del pensiero		Svoboda slova
9	Luciano Floridi	Censorship	Canarias	Chartes du droit à l'oubli numérique	Diritto di cronaca		Vikipedija
10	Viktor Mayer-Schönberger (Jurist)	Information privacy	Gobierno	Google	Regolamento generale sulla protezione dei dati		Gomoseksual'nost'

Table 51: Top 10 most relevant concepts by *CycleRank* score with respect to *"Right to be forgotten,"* across different Wikipedia language editions.

3. (cyan) groups articles related to the EU: *"European Union,"* *"European Commission,"* and *"EU–US Privacy Shield"*;

4. (orange) groups articles about EU regulations: *"Regulation (European Union),"* *"Directive (European Union),"* and *"EPrivacy Regulation (European Union)"*;

5. (dark green) groups general concepts about privacy and data: *"Information privacy,"* *"Data protection,"* and *"National data protection authority"*.

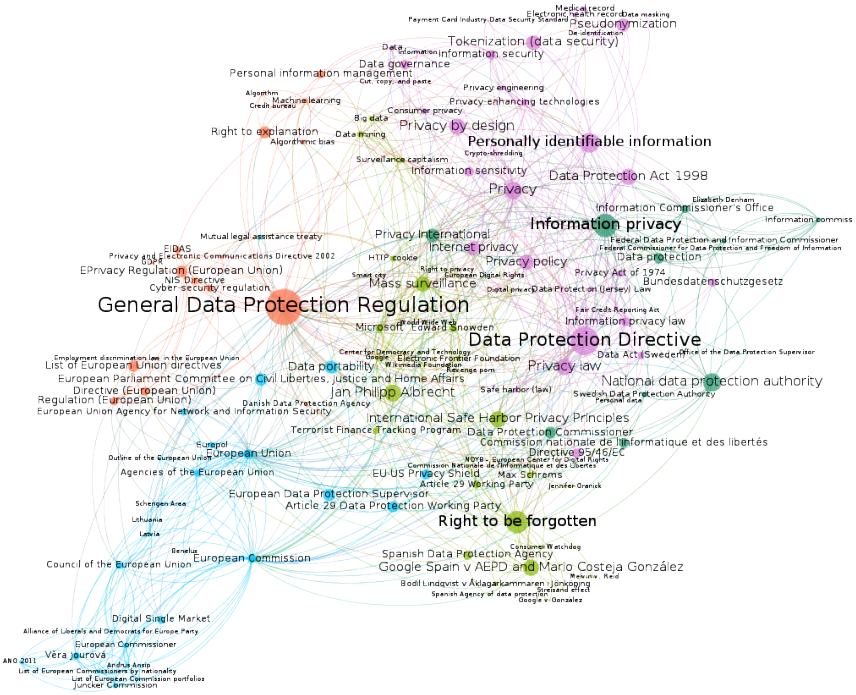


Figure 26: Graph induced by the nodes with non-zero *CycleRank* score with reference node $r =$ *"General Data Protection Regulation"* and $K = 4$ on English Wikipedia, over the snapshot of March 1st, 2018. Colors represents clusters calculated with the Louvain algorithm. The dimension of the nodes and their labels depends on the *CycleRank* score.

A.10.1 LONGITUDINAL ANALYSIS

The article *"General Data Protection Regulation"* was created on January, 3rd 2013. The top 10 results with highest *CycleRank* score over years - presented in Table 52 - show that the context of the article *"General Data Protection Regulation"* includes articles related to privacy such as: *"Privacy,"* *"Privacy law,"* *"Privacy by design"* and *"Privacy policy"*; *"Right to be forgotten"* is also consistently present since 2015. Another prominent article is *"Data Protection Directive"* the directive that regulated the processing of personal data within the European

Union before being superseded by the *"General Data Protection Regulation"*. Interestingly, article *"Mass surveillance"* appears among the top-10 in 2016 and 2017.

A.10.2 CROSS-LANGUAGE ANALYSIS

Tables 53 and 54 present the results of the cross-language analysis for the article *"General Data Protection Regulation"*. We see that across languages recurring concepts are *"Personal data"*: *"Données personnelles"* in French, *"Persoonsgegevens"* in Dutch, *"Dane osobowe"* in Polish. There are several references to laws, officers or institutions for data protection either local or international:

- In German: *"Richtlinie 95/46/EG (Datenschutzrichtlinie),"* *"Bundesbeauftragter für den Datenschutz und die Informationsfreiheit,"* *"Datenschutzbeauftragter,"* and *"Bundesdatenschutzgesetz"*;
- In Spanish: *"LOPD,"* and *"Ley Orgánica de Protección de Datos de Carácter Personal (España)"*;
- In French: *"Délégué à la protection des données,"* and *"Directive 95/46/CE sur la protection des données personnelles"*;
- In Italian: *"Trattamento dei dati personali,"* *"Codice in materia di protezione dei dati personali,"* and *"Garante per la protezione dei dati personali"*;
- In Dutch: *"Richtlijn 95/46/EG"*;

Furthermore, the keyword *"Privacy"* appears in 5 languages (en, es (*"Privacidad"*), fr (*"Vie priveée"*), it, and nl).

Finally, the *"Right to be forgotten"* appears in the top-10 results for English and Italian Wikipedia (*"Diritto all'oblio"*).

rank	2013	2014	2015	2016	2017	2018
1	General Data Protection Regulation	General Data Protection Regulation	General Data Protection Regulation	General Data Protection Regulation	General Data Protection Regulation	General Data Protection Regulation
2	Data Protection Directive	Data Protection Directive	Google Spain v AEPD and Mario Costeja González	Google Spain v AEPD and Mario Costeja González	Data Protection Directive	Data Protection Directive
3			Right to be forgotten	Jan Philipp Albrecht	Right to be forgotten	Information privacy
4			Data Protection Directive	Data Protection Directive	EU-US Privacy Shield	Right to be forgotten
5			European Parliament Committee on Civil Liberties, Justice and Home Affairs	Right to be forgotten	Personally identifiable information	Personally identifiable information
6			Privacy	Personally identifiable information	Jan Philipp Albrecht	National data protection authority
7			Information privacy	Mass surveillance	Google Spain v AEPD and Mario Costeja González	Privacy
8			Information sensitivity	Data breach	Mass surveillance	Jan Philipp Albrecht
9			Bodil Lindqvist v Åklagarkammaren i Jönköping	Privacy	Privacy	Privacy law
10			Jan Philipp Albrecht	Information sensitivity	Privacy policy	Privacy by design

Table 52: Top 10 most relevant concepts by *CycleRank* score with respect to *"General Data Protection Regulation,"* along different yearly snapshots of the WikiLinkGraphs dataset (2013-2018).

rank	de	en	es	fr
1	Datenschutz-Grundverordnung	General Data Protection Regulation	Reglamento General de Protección de Datos	Règlement général sur la protection des donnés
2	Datenschutz	Data Protection Directive	Privacidad	Vie privée et informatique
3	Datenschutzrecht	Information privacy	LOPD	Directive 95/46/CE sur la protection des données personnelles
4	Bundesdatenschutzgesetz	Right to be forgotten	Ley Orgánica de Protección de Datos de Carácter Personal (España)	Données personnelles
5	Richtlinie 95/46/EG (Datenschutzrichtlinie)	Personally identifiable information	Parlamento Europeo	Association française des correspondants à la protection des données à caractère personnel
6	Verordnung (EU)	National data protection authority	Seudonimización	Vie privée
7	Datenschutzbeauftragter	Privacy	Partido Pirata de Alemania	Journée européenne de la protection des données
8	Bundesbeauftragter für den Datenschutz und die Informationsfreiheit	Jan Philipp Albrecht	Partido Pirata Europeo	Directive du 12 juillet 2002 sur la protection de la vie privée dans le secteur des communications électroniques
9	EU-Datenschutzreform	Privacy law		Commission nationale de l'informatique et des libertés (France)
10	Jan Philipp Albrecht	Privacy by design		Délégué à la protection des données

Table 53: Top 10 most relevant concepts by *CycleRank* score with respect to *"General Data Protection Regulation,"* across different Wikipedia language editions (de, en, es, fr).

rank	it	nl	pl	sv
1	Regolamento generale sulla protezione dei dati	Algemene verordening gegevensbescherming	Ogólne rozporz adzenie o ochronie danych	Dataskyddsfŕordningen
2	Privacy	Wet bescherming persoonsgegevens	Dane wrażliwe	Dataskyddsdirektivet
3	Codice in materia di protezione dei dati personali	Autoriteit Persoonsgegevens	Dane osobowe	Personuppgiftslagen
4	Diritto all'oblio	Privacy	Ochrona danych osobowych	
5	Cookie	Persoonsgegevens	Zbiór danych osobowych	
6	Trattamento dei dati personali	Richtlijn 95/46/EG		
7	Diritto di cronaca	Privacywet (België)		
8	Browser	Europees Toezichthouder voor gegevensbescherming		
9	Garante per la protezione dei dati personali	Datalek		
10	Cache	Uitvoeringswet Algemene verordening gegevensbescherming		

Table 54: Top 10 most relevant concepts by *CycleRank* score with respect to *"General Data Protection Regulation,"* across different Wikipedia language editions (it, nl, pl, sv).

ACKNOWLEDGEMENTS

I would like to thank Michele Bortolotti and the team of "Gestione Sistemi" at the University of Trento for their support with the HPC cluster, this was expecially valuable for the work that consitutes Chapter 2 of this thesis.

I would also like to thank the following Wikipedians for their help with their local Wikipedia related to the work presented in Appendix A: Catrin Vimercati and Cornelius Kibelka (de); Patricio Lorente and Eloy Caloca Lafont (es); Nicolas Belett Vigneron (fr); Luca Martinelli (it); Lodewijk Gelauff (nl); Dariusz Jemielniak (pl); Dmitry Rozhkov (ru); and Lennart Guldbrandsson (sv).

I would like to acknowledge the support from Microsoft in the form of computing resources over the platform Microsoft Azure[22] through the "Microsoft Azure for Research Award: Data Science" program, award nº CRM:0518942.

The work in Chapters 2–4 and in Appendix A has been supported by the European Union's Horizon 2020 research and innovation programme under the EU Engineroom project, with Grant Agreement nº 780643.

22 https://azure.microsoft.com/

RINGRAZIAMENTI

«We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.»

— Donald E. Knuth [102]

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

[1] Cristian Consonni, David Laniado, and Alberto Montresor. WikiLinkGraphs: A complete, longitudinal and multi-language dataset of the Wikipedia link networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 598–607, 2019.

[2] Cristian Consonni, David Laniado, and Alberto Montresor. CycleRank, or There and Back Again: personalized relevance scores from cyclic paths on graphs. *Submitted to VLDB 2020*, 2020.

[3] Cristian Consonni, David Laniado, and Alberto Montresor. Discovering Topical Contexts from Links in Wikipedia. Part of The Web Conference 2019, 2019.

[4] Paolo Bosetti, Piero Poletti, Cristian Consonni, Bruno Lepri, David Lazer, Stefano Merler, and Alessandro Vespignani. Disentangling social contagion and media drivers in the emergence of health threats awareness. *Science Advances*, 2019. *Under review at Science Advances.*

[5] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[7] Chris Woolston. A call to deal with the data deluge. *Nature*, 525(7570), 2015.

[8] Economist editorial. The data deluge. *The Economist*, 2010.

[9] Alexa Internet, Inc. The top 500 sites on the web. https://www.alexa.com/topsites, 2019. [Online; accessed 13-March-2019].

[10] Wikipedia contributors. List of wikipedias — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=List_of_Wikipedias&oldid=886713365, 2019. [Online; accessed 13-March-2019].

[11] MediaWiki developers. English Wikipedia, Special:Statistics. https://en.wikipedia.org/wiki/Special:Statistics, 2018. [Online; accessed 28-December-2018].

[12] Dirk Lewandowski and Ulrike Spree. Ranking of wikipedia articles in search engines revisited: Fair ranking for reasonable quality? *Journal of the American Society for Information Science and Technology*, 62(1):117–132, 2011.

[13] Taha Yasseri, Anselm Spoerri, Mark Graham, and János Kertész. The most controversial topics in wikipedia: A multilingual and geographical analysis. In P.Fichman and N. Hara, editors, *Global Wikipedia: International and Cross-Cultural Issues in Online Collaboration*. Scarecrow Press, 2014.

[14] Anselm Spoerri. What is popular on wikipedia and why? *First Monday*, 12(4), 2007.

[15] Michaël R Laurent and Tim J Vickers. Seeking health information online: does wikipedia matter? *Journal of the American Medical Informatics Association*, 16(4):471–479, 2009.

[16] Wikipedia contributors. Wikipedia:manual of style/linking — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Manual_of_Style/Linking&oldid=875531776, 2018. [Online; accessed 28-December-2018].

[17] Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. Societal controversies in wikipedia articles. In Bo Begole, Jinwoo Kim, Kori Inkpen, and Woontack Woo, editors, *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 193–196. ACM, 2015.

[18] Eric Yeh, Daniel Ramage, Christopher D Manning, Eneko Agirre, and Aitor Soroa. Wikiwalk: random walks on wikipedia for semantic relatedness. In *Proceedings of the 2009 workshop on graph-based methods for natural language processing*, pages 41–49. Association for Computational Linguistics, 2009.

[19] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[20] Valentina Presutti, Sergio Consoli, Andrea Giovanni Nuzzolese, Diego Reforgiato Recupero, Aldo Gangemi, Ines Bannour, and Haïfa Zargayouna. Uncovering the semantics of wikipedia pagelinks. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 413–428. Springer, 2014.

[21] Andrea Capocci, Vito DP Servedio, Francesca Colaiori, Luciana S Buriol, Debora Donato, Stefano Leonardi, and Guido Caldarelli.

Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical review E*, 74(3):036116, 2006.

[22] Nils Markusson, Tommaso Venturini, David Laniado, and Andreas Kaltenbrunner. Contrasting medium and genre on Wikipedia to open up the dominating definition and classification of geoengineering. *Big Data & Society*, 3(2):2053951716666102, 2016.

[23] Robert West and Jure Leskovec. Human wayfinding in information networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 619–628. ACM, 2012.

[24] Benjamin Mako Hill and Aaron Shaw. Consider the redirect: A missing dimension of Wikipedia research. In *Proceedings of The International Symposium on Open Collaboration*, page 28. ACM, 2014.

[25] Wikipedia contributors. Help:wikitext — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Help:Wikitext&oldid=866759011, 2018. [Online; accessed 28-December-2018].

[26] Wikipedia contributors. Wikipedia:visualeditor — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:VisualEditor&oldid=869507404, 2018. [Online; accessed 29-December-2018].

[27] Wikipedia contributors. Wikipedia:red link — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Red_link&oldid=858691658, 2018. [Online; accessed 29-December-2018].

[28] Erik Zachte. Wikimedia stats. [Online; accessed 30-December-2018].

[29] MediaWiki. Manual:pagelinks table — mediawiki, the free wiki engine, 2019. [Online; accessed 15-January-2019].

[30] Massimo Franceschet. Pagerank: Standing on the shoulders of giants. *arXiv preprint arXiv:1002.2858*, 2010.

[31] Jeyhun Karimov, Tilmann Rabl, Asterios Katsifodimos, Roman Samarev, Henri Heiskanen, and Volker Markl. Benchmarking distributed stream processing engines. *arXiv preprint arXiv:1802.08496*, 2018.

[32] Robert West, Ashwin Paranjape, and Jure Leskovec. Mining missing hyperlinks from human navigation traces: A case study of wikipedia. In *Proceedings of the 24th international conference on*

World Wide Web, pages 1242–1252. International World Wide Web Conferences Steering Committee, 2015.

[33] Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 975–985. International World Wide Web Conferences Steering Committee, 2016.

[34] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.

[35] Julia Preusse, Jérôme Kunegis, Matthias Thimm, Steffen Staab, and Thomas Gottron. Structural dynamics of knowledge networks. In *ICWSM*, 2013.

[36] Marcius Armada de Oliveira, Kate Cerqueira Revoredo, and Jose Eduardo Ochoa Luna. Semantic unlink prediction in evolving social networks through probabilistic description logic. In *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*, pages 372–377. IEEE, 2014.

[37] Roger Guimerà and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009.

[38] An Zeng and Giulio Cimini. Removing spurious interactions in complex networks. *Physical Review E*, 85(3):036101, 2012.

[39] Andras A Benczur, Karoly Csalogany, Tamas Sarlos, and Mate Uher. Spamrank–fully automatic link spam detection work in progress. In *Proceedings of the first international workshop on adversarial information retrieval on the web*, 2005.

[40] Fabio De Rosa, Alessio Malizia, and Massimo Mecella. Disconnection prediction in mobile ad hoc networks for supporting cooperative work. *Pervasive Computing, IEEE*, 4(3):62–70, 2005.

[41] Wikipedia contributors. History of wikipedia — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=History_of_Wikipedia&oldid=875601169, 2018. [Online; accessed 28-December-2018].

[42] Wikipedia contributors. Wikipedia:five pillars — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Five_pillars&oldid=869228495, 2018. [Online; accessed 28-December-2018].

[43] Wikipedia contributors. Wikipedia:neutral point of view — Wikipedia, the free encyclopedia. https://en.wikipedia.org/

w/index.php?title=Wikipedia:Neutral_point_of_view&oldid=871557947, 2018. [Online; accessed 28-December-2018].

[44] Christian Pentzold, Esther Weltevrede, Michele Mauri, David Laniado, Andreas Kaltenbrunner, and Erik Borra. Digging Wikipedia: The Online Encyclopedia as a Digital Cultural Heritage Gateway and Site. *Journal on Computing and Cultural Heritage (JOCCH)*, 10(1):5, 2017.

[45] Pablo Aragon, David Laniado, Andreas Kaltenbrunner, and Yana Volkovich. Biographical social networks on Wikipedia: a cross-cultural study of links that made history. In *Proceedings of the eighth annual international symposium on Wikis and open collaboration*, page 19. ACM, 2012.

[46] Young-Ho Eom, Pablo Aragón, David Laniado, Andreas Kaltenbrunner, Sebastiano Vigna, and Dima L Shepelyansky. Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. *PloS one*, 10(3):e0114825, 2015.

[47] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. "the sum of all human knowledge": A systematic review of scholarly research on the content of w ikipedia. *Journal of the Association for Information Science and Technology*, 66(2):219–245, 2015.

[48] Robert West and Jure Leskovec. Automatic versus human navigation in information networks. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[49] Yilun Lin, Bowen Yu, Andrew Hall, and Brent Hecht. Problematizing and addressing the article-as-concept assumption in wikipedia. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2052–2067. ACM, 2017.

[50] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[51] David F Gleich, Paul G Constantine, Abraham D Flaxman, and Asela Gunawardana. Tracking the random surfer: empirically measured teleportation parameters in pagerank. In *Proceedings of the 19th international conference on World wide web*, pages 381–390. ACM, 2010.

[52] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. Pagerank as a function of the damping factor. In *Proceedings of the 14th international conference on World Wide Web*, pages 557–566. ACM, 2005.

[53] Malte Schwarzer, Moritz Schubotz, Norman Meuschke, Corinna Breitinger, Volker Markl, and Bela Gipp. Evaluating link-based recommendations for wikipedia. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 191–200. ACM, 2016.

[54] Jaap Kamps and Marijn Koolen. Is wikipedia link structure different? In *Proceedings of the second ACM international conference on Web search and data mining*, pages 232–241. ACM, 2009.

[55] Alexei D Chepelianskii. Towards physical laws for software architecture. *arXiv preprint arXiv:1003.5455*, 2010.

[56] AO Zhirov, OV Zhirov, and DL Shepelyansky. Two-dimensional ranking of wikipedia articles. *The European Physical Journal B*, 77(4):523–531, 2010.

[57] Leo Torres, Pablo Suárez-Serrato, and Tina Eliassi-Rad. Non-backtracking cycles: length spectrum theory and graph mining applications. *Applied Network Science*, 4(1):41, 2019.

[58] Andries E Brouwer and Willem H Haemers. *Spectra of graphs*. Springer Science & Business Media, 2011.

[59] Donald B Johnson. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing*, 4(1):77–84, 1975.

[60] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. In Eytan Adar, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov, and Belle L. Tseng, editors, *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17-20, 2009*. The AAAI Press, 2009.

[61] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9(6):e98679, 2014.

[62] Scott Emmons, Stephen Kobourov, Mike Gallant, and Katy Börner. Analysis of network clustering algorithms and cluster quality metrics at scale. *PLOS ONE*, 11:1–18, 07 2016.

[63] Ellery Wulczyn and Dario Taraborelli. Wikipedia clickstream, Feb 2017.

[64] Wikipedia contributors. Wikipedia:manual of style/layout — 'see also' section — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Manual_of_Style/Layout&oldid=906190242#%22See_also%22_section, 2019. [Online; accessed 15-July-2019].

[65] National Science and Technology Council Pandemic Prediction and Forecasting Science and Technology Working Group. Towards epidemic prediction: Federal efforts and opportunities in outbreak modeling. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/NSTC/towards_epidemic_prediction-federal_efforts_and_opportunities.pdf, 2016. [Online; accessed January-2018].

[66] Cécile Viboud, Kaiyuan Sun, Robert Gaffey, Marco Ajelli, Laura Fumanelli, Stefano Merler, Qian Zhang, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani, et al. The rapidd ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, 22:13–21, 2018.

[67] Chiara Poletto, Pierre-Yves Boëlle, and Vittoria Colizza. Risk of mers importation and onward transmission: a systematic review and analysis of cases reported to who. *BMC infectious diseases*, 16(1):448, 2016.

[68] Neil Ferguson. Capturing human behaviour. *Nature*, 446(7137):733, 2007.

[69] Chris T Bauch and Alison P Galvani. Social factors in epidemiology. *Science*, 342(6154):47–49, 2013.

[70] Gillian K SteelFisher, Robert J Blendon, Mark M Bekheit, and Keri Lubell. The public's response to the 2009 h1n1 influenza pandemic. *New England Journal of Medicine*, 362(22):e65, 2010.

[71] Neil M Ferguson, Derek AT Cummings, Christophe Fraser, James C Cajka, Philip C Cooley, and Donald S Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448, 2006.

[72] BJ Cowling, LM Ho, and GM Leung. Effectiveness of control measures during the sars epidemic in beijing: a comparison of the r t curve and the epidemic curve. *Epidemiology & Infection*, 136(4):562–566, 2008.

[73] Piero Poletti, Marco Ajelli, and Stefano Merler. The effect of risk perception on the 2009 h1n1 pandemic influenza dynamics. *PloS one*, 6(2):e16460, 2011.

[74] Sebastian Funk, Erez Gilad, Chris Watkins, and Vincent AA Jansen. The spread of awareness and its impact on epidemic outbreaks. *Proceedings of the National Academy of Sciences*, 106(16):6872–6877, 2009.

[75] Nicola Perra, Duygu Balcan, Bruno Gonçalves, and Alessandro Vespignani. Towards a characterization of behavior-disease models. *PloS one*, 6(8):e23084, 2011.

[76] Piero Poletti, Bruno Caprile, Marco Ajelli, Andrea Pugliese, and Stefano Merler. Spontaneous behavioural changes in response to epidemics. *Journal of theoretical biology*, 260(1):31–40, 2009.

[77] Sherry Towers, Shehzad Afzal, Gilbert Bernal, Nadya Bliss, Shala Brown, Baltazar Espinoza, Jasmine Jackson, Julia Judson-Garcia, Maryam Khan, Michael Lin, et al. Mass media and the contagion of fear: the case of ebola in america. *PloS one*, 10(6):e0129179, 2015.

[78] Eli P Fenichel, Carlos Castillo-Chavez, M Graziano Ceddia, Gerardo Chowell, Paula A Gonzalez Parra, Graham J Hickling, Garth Holloway, Richard Horan, Benjamin Morin, Charles Perrings, et al. Adaptive human behavior in epidemiological models. *Proceedings of the National Academy of Sciences*, 108(15):6306–6311, 2011.

[79] A Demetri Pananos, Thomas M Bury, Clara Wang, Justin Schonfeld, Sharada P Mohanty, Brendan Nyhan, Marcel Salathé, and Chris T Bauch. Critical dynamics in population vaccinating behavior. *Proceedings of the National Academy of Sciences*, page 201704093, 2017.

[80] Isaac Chun-Hai Fung, Zion Tsz Ho Tse, Chi-Ngai Cheung, Adriana S Miu, and King-Wa Fu. Ebola and the social media. *The Lancet*, 2014.

[81] Gillian K SteelFisher, Robert J Blendon, and Narayani Lasala-Blanco. Ebola in the united states—public reactions and implications. *New England Journal of Medicine*, 373(9):789–791, 2015.

[82] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. Why we read wikipedia. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1591–1600. International World Wide Web Conferences Steering Committee, 2017.

[83] Paul F. Lazarsfeld, Bernard Berelson, and Hazel Gaudet. *The People's Choice. How the Voter Makes up his Mind in Presidential Campaign*. Columbia University Press, 1944.

[84] Damon Centola. The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197, 2010.

[85] Peter Sheridan Dodds and Duncan J Watts. Universal behavior in a generalized model of contagion. *Physical review letters*, 92(21):218701, 2004.

[86] Wikimedia Foundation. Pagecounts-raw. https://wikitech.wikimedia.org/w/index.php?title=Analytics/Archive/Data/

Pagecounts-raw&oldid=1757933, 2018. [Online; accessed January-2018].

[87] Wikimedia Foundation. Pageviews analysis. https://tools.wmflabs.org/pageviews/, 2018. [Online; accessed January-2018].

[88] Google, Inc. Google news. https://news.google.com/, 2018. [Online; accessed January-2018].

[89] Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003.

[90] Marcelo FC Gomes, Ana Pastore y Piontti, Luca Rossi, Dennis Chao, Ira Longini, M Elizabeth Halloran, and Alessandro Vespignani. Assessing the international spreading risk associated with the 2014 west african ebola outbreak. *PLoS currents*, 6, 2014.

[91] Qian Zhang, Kaiyuan Sun, Matteo Chinazzi, Ana Pastore y Piontti, Natalie E Dean, Diana Patricia Rojas, Stefano Merler, Dina Mistry, Piero Poletti, Luca Rossi, et al. Spread of zika virus in the americas. *Proceedings of the National Academy of Sciences*, page 201620161, 2017.

[92] Istituto Superiore di Sanità. Meningite: l'epidemia è solo mediatica. http://www.epicentro.iss.it/problemi/meningiti/EpidemiaMediatica.asp, 2018. [Online; accessed January-2018].

[93] Brian G Southwell, Suzanne Dolina, Karla Jimenez-Magdaleno, Linda B Squiers, and Bridget J Kelly. Zika virus–related news coverage and online behavior, united states, guatemala, and brazil. *Emerging infectious diseases*, 22(7):1320, 2016.

[94] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[95] Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC, 1995.

[96] Cristian Candia, C Jara-Figueroa, Carlos Rodriguez-Sickert, Albert-László Barabási, and César A Hidalgo. The universal decay of collective memory and attention. *Nature Human Behaviour*, page 1, 2018.

[97] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[98] Wikimedia Foundation. Wikimedia traffic analysis report – page views per wikipedia language – break-

down. https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm, 2018. [Online; accessed January-2018].

[99] Duncan J Watts and Peter Sheridan Dodds. Influentials, networks, and public opinion formation. *Journal of consumer research*, 34(4):441–458, 2007.

[100] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[101] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[102] Donald E. Knuth. Computer Programming as an Art. *Communications of the ACM*, 17(12):667–673, 1974.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Final Version as of October 14, 2019 (`classicthesis` version 2.0).