Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

WikiLinkGraphs: A Complete, Longitudinal and Multi-Language Dataset of the Wikipedia Link Networks

Cristian Consonni
DISI, University of Trento
cristian.consonni@unitn.it

David Laniado
Eurecat, Centre Tecnològic de Catalunya
david.laniado@eurecat.org

Alberto Montresor
DISI, University of Trento
alberto.montresor@unitn.it

Abstract

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.

Introduction

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 (Alexa Internet, Inc. 2018). Almost 300k active users per month contribute to the project (Wikipedia contributors 2018c), 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 (MediaWiki developers 2018). Wikipedia is usually a top search-result from search engines (Lewandowski and Spree 2011) and research has shown that it is a first-stop source for information of all kinds, including information about science (Yasseri et al. 2014; Spoerri 2007), and medicine (Laurent and Vickers 2009).

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] (Wikipedia contributors 2018e), when a concept is relevant within an article, the article should include a link to the page corresponding to such concept (Borra et al. 2015). 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 (Yeh et al. 2009) to artificial intelligence (Navigli and Ponzetto 2012), from Semantic Web technologies and knowledge bases (Presutti et al. 2014) to complex networks (Capocci et al. 2006), from controversy mapping (Markusson et al. 2016) to human way-finding in information networks (West and Leskovec 2012).

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

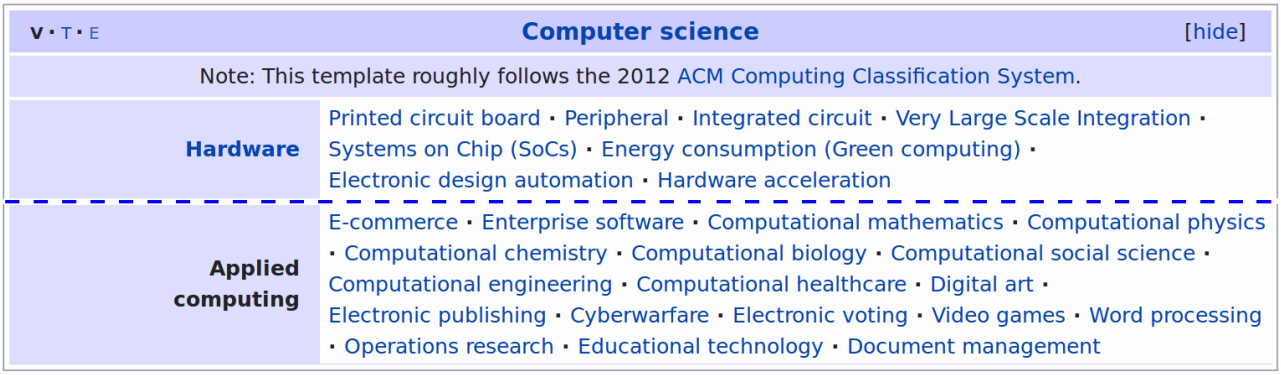
[1] https://www.wikipedia.org

[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.

[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° 878025472 of 12 January 2019, 14:12. The dashed line indicates that a portion of template has been stripped for reasons of space.

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 (Hill and Shaw 2014). 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.

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.

The WIKILINKGRAPHS Dataset

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

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* (Wikipedia contributors 2018a), 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.

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

[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.

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 (Wikipedia contributors 2018h), 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 (Wikipedia contributors 2018g). 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        [^\[]*?
9    )
10 )?
11 \]\]
```

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

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;
- `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.

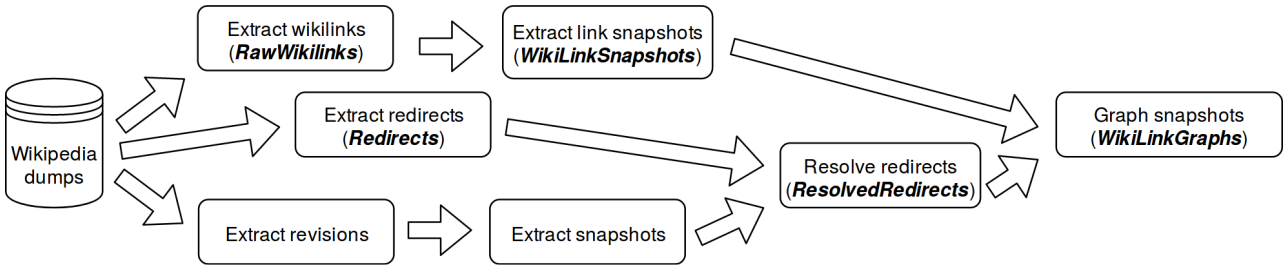


Figure 2: The process to produce the WIKILINKGRAPHS dataset from the Wikipedia dumps. In bold and italics the name of the intermediate datasets produced.

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.

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.

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

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.

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.

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

[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.

are only interested in the fact that the two pages are linked. Wikipedia policies about linking (Wikipedia contributors 2018e) 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.

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`). These editions are the top-9 largest editions per number of articles, which also had more than $1,000$ active users (Wikipedia contributors 2018c). 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 (Wikipedia contributors 2018c), 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 , 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`.

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].

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 (Zachte 2018), 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 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.

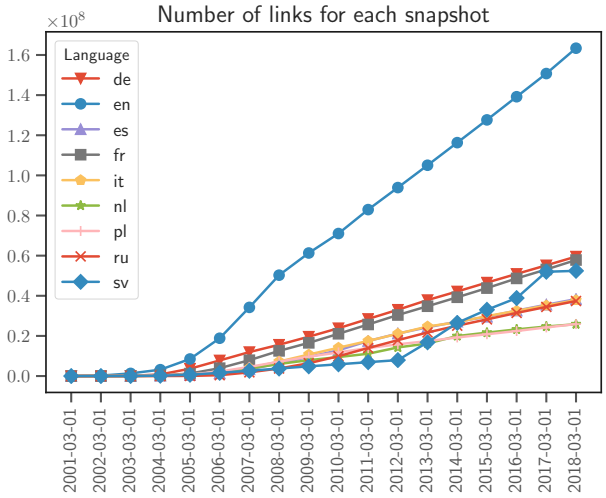


Figure 3: Overview of the growth over time of the number of links in each snapshot in the WIKILINKGRAPHS dataset.

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.

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 (MediaWiki 2019), 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 ta-

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

[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

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.

ble 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.

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 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.

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 (Brin and Page 1998). Pagerank borrows from bibliometrics the fundamental idea that being linked-to is a sign of relevance (Franceschet 2010). 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." (Wikipedia contributors 2018e)

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.

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.

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

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.

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).

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.

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 (Karimov et al. 2018). 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.

Link Recommendation. West, Paranjape, and Leskovec (2015) have studied the problem of identifying missing links in Wikipedia using web logs. More recently, Wulczyn et al. (2016) 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.

Link Addition and Link Removal. The problem of predicting the appearance of links in time-evolving networks has received significant attention (Lü and Zhou 2011); the problem of predicting their disappearance, on the other hand, is less studied. Preusse and collaborators (Preusse et al. 2013) 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. (Armada de Oliveira, Cerqueira Revoredo, and Ochoa Luna 2014) 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.

Anomaly Detection. A related problem is the identification of spurious links, i.e., links that have been erroneously

#	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	Jezyk 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 and Polish Wikipedia article titles are transliterated.

observed (Guimerà and Sales-Pardo 2009; Zeng and Cimini 2012). An example of the application of this approach is the detection of links to spam pages on the Web (Benczur et al. 2005). Similarly, the disconnection of nodes has been predicted in mobile ad-hoc networks (De Rosa, Malizia, and Mecella 2005).

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 et al. (2016) 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.

Cross-cultural studies. More than 300 language editions of Wikipedia have being created since its inception in 2001 (Wikipedia contributors 2018b), of which 291 are actively maintained. Despite the strict neutral point of view policy which is a pillar of the project (Wikipedia contributors 2018d; 2018f), 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 (Pentzold et al. 2017). Likewise, the networks from different language versions give prominence to different influential historical characters (Aragon et al. 2012; Eom et al. 2015). 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.

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.

Acknowledgements

The authors 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 work has been supported by the European Union's Horizon 2020 research and innovation programme under the EU ENGINEROOM project, with Grant Agreement n^{o} 780643.

References

Alexa Internet, Inc. 2018. The top 500 sites on the web. https://www.alexa.com/topsites. [Online; accessed 28-December-2018].

Aragon, P.; Laniado, D.; Kaltenbrunner, A.; and Volkovich, Y. 2012. 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*, 19. ACM.

Armada de Oliveira, M.; Cerqueira Revoredo, K.; and Ochoa Luna, J. E. 2014. Semantic unlink prediction in evolving social networks through probabilistic description logic. In *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*, 372–377. IEEE.

Benczur, A. A.; Csalogany, K.; Sarlos, T.; and Uher, M. 2005. Spamrank–fully automatic link spam detection work in progress. In *Proceedings of the first international workshop on adversarial information retrieval on the web*.

Borra, E.; Weltevrede, E.; Ciuccarelli, P.; Kaltenbrunner, A.; Laniado, D.; Magni, G.; Mauri, M.; Rogers, R.; and Venturini, T. 2015. Societal controversies in Wikipedia articles. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 193–196. ACM.

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

Capocci, A.; Servedio, V. D.; Colaiori, F.; Buriol, L. S.; Donato, D.; Leonardi, S.; and Caldarelli, G. 2006. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical review E* 74(3):036116.

De Rosa, F.; Malizia, A.; and Mecella, M. 2005. Disconnection prediction in mobile ad hoc networks for supporting cooperative work. *Pervasive Computing, IEEE* 4(3):62–70.

Eom, Y.-H.; Aragón, P.; Laniado, D.; Kaltenbrunner, A.; Vigna, S.; and Shepelyansky, D. L. 2015. Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. *PloS one* 10(3):e0114825.

Franceschet, M. 2010. Pagerank: Standing on the shoulders of giants. *arXiv preprint arXiv:1002.2858*.

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

Hill, B. M., and Shaw, A. 2014. Consider the redirect: A missing dimension of Wikipedia research. In *Proceedings of The International Symposium on Open Collaboration*, 28. ACM.

Karimov, J.; Rabl, T.; Katsifodimos, A.; Samarev, R.; Heiskanen, H.; and Markl, V. 2018. Benchmarking distributed stream processing engines. *arXiv preprint arXiv:1802.08496*.

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

Lewandowski, D., and Spree, U. 2011. 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.

Lü, L., and Zhou, T. 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications* 390(6):1150–1170.

Markusson, N.; Venturini, T.; Laniado, D.; and Kaltenbrunner, A. 2016. Contrasting medium and genre on Wikipedia to open up the dominating definition and classification of geoengineering. *Big Data & Society* 3(2):2053951716666102.

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

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

Navigli, R., and Ponzetto, S. P. 2012. Babelnet: The automatic construction, evaluation and application of a wide-

coverage multilingual semantic network. *Artificial Intelligence* 193:217–250.

Pentzold, C.; Weltevrede, E.; Mauri, M.; Laniado, D.; Kaltenbrunner, A.; and Borra, E. 2017. Digging Wikipedia: The Online Encyclopedia as a Digital Cultural Heritage Gateway and Site. *Journal on Computing and Cultural Heritage (JOCCH)* 10(1):5.

Presutti, V.; Consoli, S.; Nuzzolese, A. G.; Recupero, D. R.; Gangemi, A.; Bannour, I.; and Zargayouna, H. 2014. Uncovering the semantics of wikipedia pagelinks. In *International Conference on Knowledge Engineering and Knowledge Management*, 413–428. Springer.

Preusse, J.; Kunegis, J.; Thimm, M.; Staab, S.; and Gottron, T. 2013. Structural dynamics of knowledge networks. In *ICWSM*.

Spoerri, A. 2007. What is popular on wikipedia and why? *First Monday* 12(4).

West, R., and Leskovec, J. 2012. Human wayfinding in information networks. In *Proceedings of the 21st international conference on World Wide Web*, 619–628. ACM.

West, R.; Paranjape, A.; and Leskovec, J. 2015. Mining missing hyperlinks from human navigation traces: A case study of wikipedia. In *Proceedings of the 24th international conference on World Wide Web*, 1242–1252.

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

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

Wikipedia contributors. 2018c. List of wikipedias — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=List_of_Wikipedias&oldid=873324169. [Online; accessed 28-December-2018].

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

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

Wikipedia contributors. 2018f. Wikipedia:neutral point of view — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Neutral_point_of_view&oldid=871557947. [Online; accessed 28-December-2018].

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

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

Wulczyn, E.; West, R.; Zia, L.; and Leskovec, J. 2016. Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, 975–985. International World Wide Web Conferences Steering Committee.

Yasseri, T.; Spoerri, A.; Graham, M.; and Kertész, J. 2014. The most controversial topics in wikipedia: A multilingual and geographical analysis. In P.Fichman, and Hara, N., eds., *Global Wikipedia: International and Cross-Cultural Issues in Online Collaboration*. Scarecrow Press.

Yeh, E.; Ramage, D.; Manning, C. D.; Agirre, E.; and Soroa, A. 2009. Wikiwalk: random walks on wikipedia for semantic relatedness. In *Proceedings of the 2009 workshop on graph-based methods for natural language processing*, 41–49. Association for Computational Linguistics.

Zachte, E. 2018. Wikimedia stats. [Online; accessed 30-December-2018].

Zeng, A., and Cimini, G. 2012. Removing spurious interactions in complex networks. *Physical Review E* 85(3):036101.