# Discovering Topical Contexts from Links in Wikipedia

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

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

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

## ABSTRACT

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 which are more relevant to a given concept. In this context, a well-known algorithm is Personalized PageRank; its performance, however, is hindered by pages with high in-degree that function as hubs and appear with high scores regardless of the starting point. In this work, we present how a novel algorithm based on cyclic paths can be used to find the most relevant nodes in the Wikipedia link network related to a topic. We present a case study showing how the most relevant concepts associated with the topic of "Fake news" vary over time and across language editions.

**ACM Reference Format:**
Cristian Consonni, David Laniado, and Alberto Montresor. 2019. Discovering Topical Contexts from Links in Wikipedia. In *Wiki Workshop 2019, May 14, 2019, San Francisco, CA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Wikipedia[1] 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 [1]. Wikipedia exists in more than 290 active different language editions [12], and its pages have been edited over 2.5 billion times. Wikipedia has a strong impact on the formation of public opinion on potentially any topic. It is also widely used in education in almost all the fields of knowledge as an increasingly va uablelresource by both students and teachers [7].

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 [3]. Surfing the links between Wikipedia articles has been described as fascinating[2] and serendipitous[3] [13]. Unlike in other hyperlink networks, in an encyclopedia each page corresponds to a concept, so the links between articles constitute a vast concept network, emerging from the collaborative process that involves thousands of users. Previous research in controversy mapping has shown how this network can be leveraged to analyze the dominating definition of a topic, such as *"Geoengineering"* [9], as it emerges in Wikipedia, shedding light on its boundary, context and internal structure. Furthermore, each linguistic community in Wikipedia produces a different network, which allows for comparing public debate about a topic across different language editions [11]. In this work, we present a novel approach to make sense of the Wikipedia link network, and to detect the concepts that are most relevant to a given topic.

The work presented in this paper has been realized under the Engineroom project, part of the *"Next Generation Internet"* (NGI) initiative by the European Commission aimed at defining a vision for developing the future Internet as to reflect a set of core values such as openness, inclusiveness, transparency, privacy, and cooperation. In this framework, the Engineroom project is focused on developing a data-driven methodology to identify and evaluate the key enabling technologies and topics that will underpin the Next Generation Internet. In order to study the framing of topics such as "Online privacy", "Online identity" or "Right to be forgotten" over several years and across languages, we aim to look at their context as it emerges from the Wikipedia link network. The connections between Wikipedia articles are valuable, but they are also very abundant. The English version has more than 160 million links [4] between its 5.7 million articles. How can one find guidance within this wealth of data? 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. The topics we are interested in are identified by single nodes or sets of nodes in the Wikipedia link graphs. Such nodes are *queries* we are interested in answering. We refer to a node of interest as a *reference node*, or a *seed*, i.e., a starting point for our search, and we need an algorithm to identify the topics which are more relevant with respect to it.

Given a reference node $r$ we are interested in answering the following questions:

- Which are the most relevant concepts related to topic $r$? Can we assign a score to all the nodes in the graph that captures how relevant to topic $r$ they are?
- How does such relevance vary across different language editions?
- How does the context of an article change over time? Since the entire edition history of Wikipedia is available, we can study how the context of a concepts change over time.

One established algorithm to answer these questions is Personalized PageRank: a variant of PageRank where the user can specify one or more nodes as queries and obtain a score for all the other nodes in the graph that encapsulates the relatedness between the

[1] https://www.wikipedia.org
[2] https://xkcd.com/214/
[3] https://www.youtube.com/watch?v=8Z9IcBmrmeY


*Wiki Workshop 2019, May 14, 2019, San Francisco, CA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

two topics. However, applied in the context of Wikipedia we have found that this algorithm does not produce satisfactory results since it usually includes very general articles in top position.

To overcome these limitations, we have developed a novel algorithm for finding the most relevant nodes in the Wikipedia link network related to a topic. The technique, called LoopRank, takes advantage of the loops 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.

After briefly introducing the dataset, in the next section we describe the algorithm and its implementation. Then in Section 3 we show the results obtained for a set of topics related to the Next Generation Internet, such as "Internet privacy" or "Right to be forgotten", and compare them with the results obtained with the Personalized PageRank algorithm. We further present a case study centered on the article "Fake news", including a visualization of the network around the article, and a longitudinal and a cross-language analysis. Finally, in Section 4 we present some directions for future work, including possible extensions of the algorithm, and we draw our conclusions in Section 5.

## 2 METHODS AND MATERIALS

This section describes the dataset and methodology of this study: the data sources that we have used; Section 2.2 describes in detail the algorithms: Personalized PageRank 2.2.1 and LoopRank 2.2.2 .

### 2.1 Dataset Description

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*. For our analysis, we used the WikiLinkGraphs dataset consisting of the network of internal Wikipedia links for the 9 largest language editions, [4]. The dataset has been developed by us and it is publicly available on Zenodo[4].

The dataset contains yearly snapshots of the network and spans 17 years, from the creation of Wikipedia in 2001 to March 1st, 2018. 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 using templates. The dataset also handles special pages such as *redirects*, i.e., alternative article titles. 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 [12].

### 2.2 Methods

In the following section, we compare two techniques for assigning a score to all the nodes of a graph, with respect to a given reference node: *Personalized PageRank*, and a novel method *LoopRank*.

*2.2.1 Personalized PageRank.* PageRank [10] is a metric 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. PageRank can be computed in an iterative process, as the score of a node depends on the score of the nodes that link to it, however more efficient algorithms are available.

The idea at the basis of PageRank is that of simulating a stochastic process in which a user follows random paths in a hyperlink graph. From each node, the algorithm assumes equal probabilities of following any hyperlink included in the page and a certain probability of "teleporting" to another random page in the graph. The damping factor $\alpha$, generally assumed to be 0.85, defines the probability of continuing surfing the graph versus teleporting.

Personalized PageRank [10] is a variant of the original PageRank algorithm, in which teleporting is not directed to all nodes randomly, but to a specific node or set of nodes. In this way, the algorithm models the relevance of nodes around the selected set of reference nodes, as the probability of reaching each of them, when following random walks starting from this chosen set.

**Limitations of personalized PageRank**. The Personalized PageRank algorithm seems at first look to be suitable for our use case, as it can be used to represent a measure of relevance of Wikipedia articles strongly linked to (directly or indirectly) from the seed. However, applying this algorithm we found unsatisfactory results. Even starting from different seed articles, at the first place of the ranking we tend to find articles that are very central in the overall network. Such central articles act as hubs in the graph and have such strong relevance in the overall graph that, even starting from a seed article which is not specially related to them, one is very likely to end up reaching them in the exploration of the graph.

We argue that this is due to different factors. First, the fact that paths of any length can be followed, so - in a densely connected graph - many paths will tend to converge at a certain point towards the most relevant nodes. This aspect can be limited only partially by lowering the value of the damping factor. Even peripheral nodes can contribute to the PageRank score of more central nodes; we have found experimentally that excluding the furthest nodes worsens the results of the algorithm. Furthermore, the fact that the PageRank only accounts for in-links, and not for out-links. This is reasonable for web searches, where in-links are a good proxy for relevance, as they represent somehow the value attributed to a node by the other nodes of the graph; on the contrary, out-link have basically no value in this sense: it is very easy to add into one's web page many out-links to other web pages, and this is radically different from obtaining in-links from other web pages, for which one needs many other people to consider that web page relevant. In the context of Wikipedia, instead, links from an article to other articles may be subject to being deleted as much as incoming links from other articles, and so both outgoing and incoming links can be considered as indicators of relevance. In particular, out-links from other pages to an article can be a very valuable indicator that these pages are actually related to the topic; e.g., if an article contains links to *"Internet privacy,"* or to many other articles linking in turn to "Internet privacy", then we can assume that the content of the article is related to *"Internet privacy."* We can expect the article *"United States"* to have, instead, only a few links to articles related to *"Internet privacy,"* as it is not the main subject of the article.

[4] https://zenodo.org/record/2539424, DOI: 10.5281/zenodo.2539424

*2.2.2 The LoopRank algorithm.* In this section, we propose a more general approach to the problem, defining a new metric of the relevance of a node in a directed network, that accounts for both incoming and outgoing links. We call this metric LoopRank, as it is based on the idea of circular walks. We start from the observation that the personalized PageRank algorithm is not suitable for our context because random walks may easily lead to paths that are not related to the topic under consideration; th suwe only consider 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 article and come back. We defined a new algorithm to suit our purpose of identifying the nodes that are more relevant and related to a given reference node, accounting for both directions in the network. We first provide a definition of the LoopRank, then we present the outline of the implementation of the algorithm. Finally, we discuss how its results can be intuitively interpreted.

**Definition**. As for Personalized PageRank, the goal of LoopRank is to assign a score to all the nodes in a graph that measures their relevance to a given reference node provided as input. Intuitively, a node that is linked from the reference article, but does not link back to, is likely to be a concept that is not related to that subject, even if it is relevant to its definition. Specularly, a node that links to the reference article, but is not linked from it, is likely to be related, but not relevant. Nodes that are linked both from and to a reference node are the ones that we expect to be both relevant and related to it.

Extending this principle, we want then to be able to quantify the importance of a node with respect to a given reference node, accounting also for the indirect links, i.e., for the numbe rof paths that can be found linking it from and to the reference node. We do this by counting the number of loops of various lengths that contain the reference node and any other node. As short distances represent a stronger relationship, short loops rece viea higher weight.

*Definition 2.1 (LoopRank score of a node $i$ with respect to a reference node $r$ and maximum loop length $K$).* Given a directed graph $G(V, E)$ a reference node $r \in V$ and an integer $K > 1$ the LoopRank score of any node $i \in V$ is given by:

$$LR_{r,K}(i) = \sum_{n=2}^{K} \sigma(n) \cdot \ell_{r,n}(i) = \sum_{n=2}^{K} \frac{\ell_{r,n}(i)}{n} \tag{1}$$

where $\ell_{r,n}(i)$ is the number of loops of length $n$ that contain nodes $i$ and $r$, $K$ is a parameter representing the maximum length considered for loops, and $\sigma(n)$ is the general form of a scoring function that weights the score assigned for each loop. We set it to be $\sigma(n) = \frac{1}{n}$.

In this way, given a reference node, the LoopRank score of a node $i$ represents the number of loops including both the reference node and node $i$, normalized by the length of each loop. More precisely, each loop is considered to contribute with the same overall score of 1, that gets split equally among the various nodes involved in the loop. By definition, the reference node gets the maximum LoopRank score as it is included in all the loops considered.

The computation of LoopRank does not necessarily consider all the loops going through a node, which is the case only if the value of $K$ equals or exceeds the length of the longest loop appearing in the graph. This length is bounded by $D + 1$, where $D$ is the diameter of the directed network. Limiting $K$ allows to reduce the time needed to compute the score and it avoids the introduction of potential noise deriving from long loops that include popular nodes that are far from the reference node. This is not the case for Personalized PageRank, where we have found experimentally that limiting the network only to the nodes closest to the reference node worsen its results.

From our experimental evaluation, presented in Section 3, we have chosen to use $K = 4$, which produced the best results and at the same time requires a limited computational effort.

**Outline of the algorithm**. We present here an outline of the algorithm that calculate LoopRank scores on a graph $G(V, E)$ given a reference node $r$ and maximum length $K$. Our strategy is to reduce first the network dimension as much as possible. To this end, we employed known efficient algorithms to filter the network, discarding the nodes that for sure will not be used in the computation.

The algorithm performs the following steps:

(1) we perform a breadth-first search from the reference node $r$ on the graph $G(V, E)$ to calculate the distance of any other node from $r$, $d(r)$; unreachable nodes get $d(r) = +\infty$;
(2) we discard immediately all the nodes at a distance $d(r)$ greater than $K - 1$;
(3) we perform a breadth-first search from $r$ on the transposed network $G^T(V, E)$, so that we calculate the distance of $r$ from any other node in the original graph $G(V, E)$; we denote this distance with $d^T(r)$;
(4) we discard all the nodes for which $d(r) + d^T(r) > K$;
(5) we enumerate all the simple loops starting from $r$ using Johnson's algorithm [8];
(6) we calculate the LoopRank score of each node appearing in the found cycles using Equation 1; all other nodes get a LoopRank score of 0.

In steps 1-4, we discard all the nodes that are not reached by any loop of length lower than $K + 1$. All the surviving nodes by definition belong to a subset of the *strongly connected component* containing the reference node $r$; these are the nodes that will receive a LoopRank score greater than zero, while all the other nodes will get a score of zero since they do not belong to any loop of length at most $K$. In step 5, we execute Johnson's algorithm [8] from node $r$ to identify all the simple cycles going through $r$. Finally, in step 6, we compute the LoopRank score of every node according to Equation 1.

**Interpretation of the algorithm**. The LoopRank score can be seen as the time spent on a given node when following random loops from the reference node, assuming a fixed overall time for each loop, equally split among all the nodes encountered. The LoopRank algorithm is similar to the personalized PageRank in that it can be explained as random paths followed starting from the reference node, with the main difference being that only circular paths are allowed. No damping factor is needed in the case of LoopRank, as the random cyclic paths can be modeled as consecutive, always

starting from the reference node and coming back to it. The exact resulting probabilities can be calculated by enumerating all the possible loops including the reference node.

## 3 RESULTS

We first present in Section 3.1 the results of a comparison between Personalized PageRank and LoopRank over a variety of terms related to the *"Next Generation Internet"*. Then, in Section 3.2, we focus on the article "Fake news" and we explore the capabilities of LoopRank more in-depth by performing 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.

### 3.1 Comparison of LoopRank and PageRank

Tables 1 and 2 present a comparison between the top-10 results with the highest scores obtained with the Personalized PageRank and the LoopRank, 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 2.2.1: 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 [4].

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 LoopRank 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 LoopRank score of zero and do not appear in the rankings. In this way, LoopRank leaves space to articles whose content is more strongly associated with the reference topic to appear at higher positions in the ranking.

### 3.2 Case Study: *"Fake news"*

This section illustrates the results of longitudinal and cross-language analyses obtained with the LoopRank algorithm taking *"Fake news"* as a starting point. We have performed this analysis for all the topics pertaining to the scope of the ENGINEROOM project, but here we present just the results for *"Fake news"* for reasons of space.

*3.2.1 Network visualization.* Figure 1 shows a visualization of the network centered around the article "Fake news", where node size reflects the LoopRank 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 2.2.2, is used for this purpose: all the concepts which do not share any loop

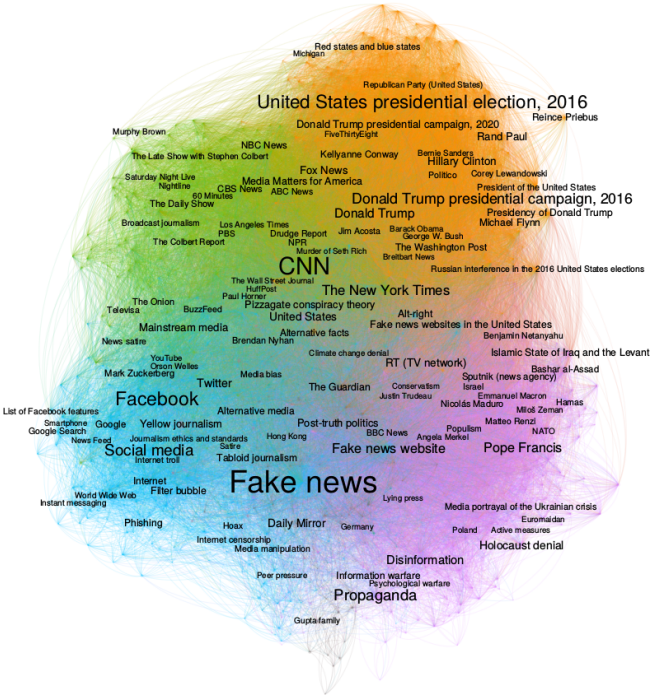


**Figure 1: Graph induced by the nodes with non-zero LoopRank score with reference node $r$ = "Fake news" and $K = 4$ on English Wikipedia. over the snapshot of March 1st, 2018. The network is visualized after applying the ForceAtlas2 algorithm. Colors represents clusters calculated with the Louvain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

shorter than $K = 4$ are removed so that only concepts having a LoopRank score greater than 0 are included in the visualization. For readability reasons, node label is shown only for articles having a LoopRank score 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] [2], 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 [6]. 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 [5].

*3.2.2 Longitudinal analysis.* Table 3 presents the LoopRank scores calculated over the snapshots of the wikilink graph taken on

[5] https://gephi.org/

| page | Fake news | | Right to be forgotten | | Online identity | |
|---|---|---|---|---|---|---|
| # | LoopRank | PageRank | LoopRank | PageRank | LoopRank | 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 1: Top-10 articles with the highest LoopRank 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 | |
|---|---|---|---|---|---|---|
| # | LoopRank | PageRank | LoopRank | PageRank | LoopRank | 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 2: Top-10 articles with the highest LoopRank 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 3: Top-10 articles with the highest LoopRank 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.**

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.

*3.2.3 Cross-language analysis.* Tables 4 and 5 present the LoopRank 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 LoopRank is able to find results that are pertaining to the local Wikipedia edition, for example, results from 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:

(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.

## 4 FUTURE WORK

We have presented a version of the algorithm with a simple scoring function, however many variations and extensions could be explored. We have assumed that the starting point, or seed, for the algorithm is a single node, which we have called the reference article. However, as in the case of Personalized PageRank, it would be possible to take as seed a group of articles. Then, all loops around each of the seed nodes could be considered. Alternatively, instead of counting loops, 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 considering all paths from the source to the target within K steps. In this way, the metric 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. So, 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"?

Finally, in the definition of the score, we have chosen the denominator to be linear in the number of nodes; this is an intuitive solution which gives more weight to closer nodes belonging to shorter loops. This solution is also easily explainable as it assumes each loop to contribute with the same value, which is split into equal parts among the nodes belonging to it. We have empirically validated the good quality of the results obtained through this solution for the Wikipedia graph in many contexts. However, a more extensive evaluation could be developed to assess the results obtained with different solutions, such as a quadratic denominator, which would further penalize longer loops, and could be suitable for the purpose to give more importance to closer nodes compared to popular, but less close nodes. The suitability of different solutions

[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 LoopRank score greater than zero.

| # | de | | it | | 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 4: Top-10 articles with the highest LoopRank 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: (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 5: Top-10 articles with the highest LoopRank 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: (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.**

could also be studied focusing on the structural properties of the network, such as its link density or clustering coefficient.

We are conducting a quantitative analysis of LoopRank to better characterize its effectiveness in different contexts compared to Personalized PageRank. This analysis will also provide some insights into the choice of the most appropriate scoring function to use for LoopRank.

## 5 CONCLUSIONS

Links in Wikipedia enshrine valuable knowledge, however, their abundance poses a limit on how we can understand which links are most relevant. In this work, we have presented a novel algorithm, called LoopRank, that assigns a score to all nodes that are connected to a given reference node by one or more cyclic path in the link graph. We have described the difference between LoopRank and Personalized PageRank, a well-known existing algorithm that surfs over links in a graph. Both these algorithms answer the question of finding the most relevant context around a given topic of choice, however their results differ. We have analyzed the case of several keywords related to the Next Generation Internet in Wikipedia, and we have shown that rankings of relevant related pages produced by LoopRank are more specific to the topic of choice. Finally, we have presented a case study focused on the topic "Fake news", including a longitudinal analysis over time and a cross-language analysis in 8 languages to provide an insight into the capabilities of the algorithm. Looprank has proven to be a flexible algorithm that can return interesting results; finally, we are currently conducting quantitative analysis to better characterize and evaluate the performance of the algorithm.

## ACKNOWLEDGMENTS

The authors would like to thank the following Wikipedians for their help with their local Wikipedia: 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).

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

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

[2] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media.*

[3] Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. 2015. Societal controversies in Wikipedia articles. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems.* ACM, 193–196.

[4] Cristian Consonni, David Laniado, and Alberto Montresor. 2019. WikiLinkGraphs: A complete, longitudinal and multi-language dataset of the Wikipedia link networks. *arXiv preprint arXiv:1902.04298* (2019).

[5] Scott Emmons, Stephen Kobourov, Mike Gallant, and Katy BÃűrner. 2016. Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale. *PLOS ONE* 11 (07 2016), 1–18. https://doi.org/10.1371/journal.pone.0159161

[6] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* 9, 6 (2014), e98679.

[7] Dariusz Jemielniak and Eduard Aibar. 2016. Bridging the gap between wikipedia and academia. *Journal of the Association for Information Science and Technology* 67, 7 (2016), 1773–1776.

[8] Donald B Johnson. 1975. Finding all the elementary circuits of a directed graph. *SIAM J. Comput.* 4, 1 (1975), 77–84.

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

[10] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

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

[12] Wikipedia contributors. 2019. List of Wikipedias — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=List_of_Wikipedias&oldid=886713365. [Online; accessed 13-March-2019].

[13] Wikipedia contributors. 2019. Wikipedia:Getting to Philosophy — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Getting_to_Philosophy&oldid=880926083. [Online; accessed 13-March-2019].