2019 IEEE International Conference on Big Data (Big Data)

Wikipedia Polarization and Its Effects on Navigation Paths

Cristina Menghini
DIAG
La Sapienza, University of Rome
Rome, Italy
menghini@diag.uniroma1.it

Aris Anagnostopoulos
DIAG
La Sapienza, University of Rome
Rome, Italy
aris@diag.uniroma1.it

Eli Upfal
Computer Science
Brown University
Providence, Rhode Island, United States
eli_upfal@brown.edu

***Abstract*—Bias and polarization are not just about placing misinformation on the Web but also involve concerted efforts to change how we navigate it. One of the strongest points of Wikipedia is to allows readers to easily navigate a topic, through its hyperlinks structure. Thus, it is crucial to ensure a user to have the same probability of being exposed to knowledge that expresses different viewpoints concerning the given topic. In this work, we investigate whether the topology and polarization of a topic-induced-graph (e.g. *U.S. Politics induced network*) has an impact on users' navigation paths making them biased toward one of the possible topic perspectives. Modeling users behaviour and exploiting Wikipedia *clickstreams*, we analyze users exposure to different leaning during their sessions, thus the chance of being trapped within a *knowledge bubble* presenting a unique viewpoint about the topic, and differences among users that start their navigation from articles representing different perspectives.**

***Index Terms*—Wikipedia, Bias, Polarization, Knowledge bubble, Learning, Data Science.**

I. Introduction

Wikipedia is the main global, free, and most accessible source of knowledge and information on the Web. It is the result of human cooperation, thus it is likely to inherit human bias that, on our perspective, arises as:

Free to air bias when article's content is skewed towards a perspective.

Structural bias presents in the hyperlink network of articles and deriving from editors that add links according to their knowledge map, which can reflect a specific standpoint.

Free-to-air bias is monitored through peer-review process. In contrast, *structural bias* is hidden and difficult to detect without a comprehensive view of the article in the context of the induced network. Generally, readers access Wikipedia to broaden their knowledge about a specific topic [1]. It can appear as self-contained-topic page or set of articles. The former goes through a review process that ensure exposure to a neutral topic perspective. Differently, topics represented by many articles are **not** subject to any process that monitors whether their induced network is biased driving readers towards a partial topic overview. Given a topic of interest, that throughout this work is the *U.S. Politics*, we address the following research questions.

RQ1: Is the topic-induced-network polarized? E.g. We quantify the strength of links between *Democratic* and *Republican* articles.

RQ2: Does the network polarity have an impact on navigation behavior? E.g. What is the average number of consecutively visited Republican/Democratic pages before hitting the opposite *knowledge bubble*? On average, how many opposite leaning articles are visited before returning to the initial *knowledge bubble*?

RQ3: Are there graph topological characteristics that explain what we observe?

The rest of the paper focuses on the analysis of the research questions. In section III, we explain the data used and the graph experiments are run on. Then, in section IV, we investigate the level of polarity of the network relying on metrics previously defined in the literature. Once the network polarity is assessed, in section V and VI, we address respectively RQ2 and RQ3. Eventually, in section VII, we report limitations and further work.

II. Related Work

Although the topic of bias and polarization has been investigated extensively in recent years [2], no significant advances have been made in exploring and measuring these phenomena in Wikipedia. Wikimedia research is continuously working to make Wikipedia more reliable and uniform [3]. With this project, we want to shed light on the importance to monitor the impact that structural bias could have on Wikipedians. Previous work on Wikipedia's user behavior paves the way for assumptions made in our analysis [4].

III. Topic Induced Network

The topic of interest is identified by a set of Wikipedia categories referring to the topic. The collection of articles belonging to these categories, and respective subcategories [1], composes the set of nodes that induces the new network. So, we obtain the topic induced network, a directed graph whose edges represent the possibility, for a reader, to go from one page to the other.

[1]Only subcategories whose name include some specific keywords, related to the topic, are included.

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

A. Graph partitioning

The set of categories used to obtain the graph is exploited to label each node of the graph as expression of one viewpoint, e.g. *Democratic* or *Republican*. In this work we restrict to only two partitions, but it can be adapted for more leanings. Categories, in Wikipedia, are assigned by editors and validated through a strict review process. For this reason, even if the category structure can be messy, redundant and not well descriptive, we find them **reliable** for our purpose. This strategy, to partition the graph requires the categories to be already associated with a specific side. The strength of that procedure lies on the fact that labeling is not derived from articles hyperlinks structure. In this way, analysis on *topic induced network* should not be affected by graph building procedure itself.

B. Clickstream as weights

To study the behavior of current Wikipedia users, we weight the graph according to *Clickstream* data [5], that record the monthly number of transitions among two pages. Since *Clickstream* data do not report streams smaller than 10, we set the passages among not-clicked hyperlinks to 1. This is because, these *wikilinks* could still be clicked in the future.

C. PoliNet: U.S. Politics Induced Network

We construct the topic induced network for politics in the United States (*PoliNet*). It is derived from the following list of seed categories: *Democrats*, *Democratic Party (United States)*, *Republicans*, *Republican Party (United States)*. The first two tagged as *Republican* and the last as *Democratic*. In Table I, we report information about *PoliNet*. The number of edges includes those added to sink pages, to which we add the possibility of back clicking to the article they are pointed by.

TABLE I
POLINET INFORMATION

# pages	# Republicans	# Democrats	# Wikilinks
20726	10403	10323	147967

Investigating about the hyperlinks within and between pages of different leaning, we find out that only the 25% of *wikilinks* join Republican and Democratic pages. The number of links among Democratic pages is slightly higher than that among Republican ones.

D. RandPoliNet: Baseline

As baseline to compare the analysis run on *PoliNet*, we use a uniform random graph, based on the *configuration* model [6], whose degree distribution is given and it is the same of *PoliNet*. This is called *RandPoliNet*. To have an idea of the differences among *PoliNet* and *RandPoliNet* we refer to Table II. Overall, in the random graph the edges that connect the two factions are 50% more than links within the same color.

TABLE II
RATIO OF EDGES BETWEEN POLINET AND RANDPOLINET

PoliNet / RandPoliNet	
Republican to Republican	1.50
Democratic to Democratic	1.51
Republican to Democratic	0.50
Democratic to Republican	0.49

IV. TOPIC POLARITY

To verify the presence of polarization on *PoliNet*, and more specifically that of *knowledge bubbles*, we rely on metrics that have been defined in past literature. First, the *modularity* [7] that measures the strength of division of a network into communities.

Then, we consider the *boundary polarization* [8], that quantifies the polarization of the graph by comparing the degree of preference, of the nodes on the boundary, to connect nodes of the same color and the boundary node.

The last metrics, *random walk controversy* [9], given two random walks, measures the difference of the probabilities that both random walks start and end in the same and different graph partition In Table III, we see that for all the metrics adopted, *PoliNet* turns out to be polarized, thus we assess the presence of two *knowlwdge bubbles*. The level of polarization of the Republican and Democratic knowledge chamber are almost the same. Looking at the RW Controversy scores, Democratics register a slightly higher probability of finishing the walks on the opposite side, then Republicans.

Result 1: Only considering the topological structure of the network, thus analysing the unweighted *PoliNet*, according to three polarization metrics, the *topic-induced-network* turns out to be polarized. In particular, the strength of polarization for Democratic and Republican pages is the same.

V. NAVIGATION BEHAVIOUR

To study the impact of the verified *topic-induced-network* polarity on users navigation habits, we simulate 40000 users that randomly move throughout the network according to *clickstreams*. Half of users start their session from the Democratic *knowledge bubble* and half from the Republican one. Results observed on *RandPoliNet* show the behaviour of random users on a network without *knowledge bubbles* and offer a baseline to measure the effect of polarity on user exploring *PoliNet*. Result 2 summarises what we observe in later subsections, V-A and V-B.

Result 2: The average number of pages a user needs to explore, before reaching one article belonging to the opposite faction, is 5 for Republican seeds and 7 for Democratic seeds. Thus, users that starts to read a Democratic article on average stays in the Democratic bubble longer than those that start to navigate in the Republican bubble. Further exploration shows that the Democratic knowledge chamber has a *retention-user-rate* higher than the Republican bubble for users that read more than 8 Democratic consecutive pages.

TABLE III
MODULARITY, BOUNDARY POLARIZATION AND RANDOM WALK CONTROVERSY MEASURED ON POLINET AND RANDPOLINET

	PoliNet			RandPoliNet		
	Overall	Republican	Democrats	Overall	Republican	Democrats
Modularity	0.26	-	-	-	-	-
Boundary polarization	0.215	0.212	0.218	-0.026	-0.023	-0.028
RW controversy	0.441	0.440	0.430	0.0024	0.027	-0.022

A. Opposite Color Hitting Time

First we address RQ2 measuring the average time a user needs to reach the opposite bubble. In Figure 1, we observe that, on RndPoliNet, the entire sample of users does not have a walk longer than 3 to reach the opposite Knowledge Chamber. Instead, referring to the boxes related to PoliNet, 75% of users that starts reading a Republican page arrives in the Democratic Knowledge Chamber in 2 clicks less than a user, starting from Democratic article, requires to arrive on a Republican page.

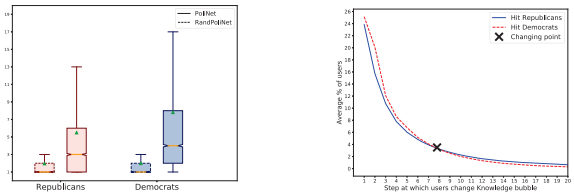

Fig. 1. The left plot shows the *hitting time distribution* for simulated users on *RandPoliNet* and *PoliNet*. On the right, the percentage of users that change bubble at a given step.

To investigate further the *opposite bubble hitting time*, we refer to the right plot on Figure 1. The plot shows that, on average, the percentage of users that reaches the opposite bubble in less than 8 step is higher for readers starting from a Republican page. This result is particularly interesting because shows that, for paths longer than 8, the *user-retention-rate* in the Democratic bubble is higher than those starting in the Republican one.

B. Opposite Bubble Exploration

Once a user arrives in the opposite-seed-color bubble, we are interested in knowing for how long it remains in it before moving back to the starting bubble. Analyzing the behaviour of simulated users, the distribution of the time spent in the Republican bubble starting the navigation from a Democratic page, is the same as the time spent in the Democratic bubble. To explain these results, one could dig more into the ratio of blue and red neighbors of the bridging nodes. The distributions only differ with respect to the median. Indeed, half of users coming from Republican bubble, go back to it in 1 click. Beside, half of the users coming from the Democratic bubble, requires 2 clicks. The distributions related to *RandPoliNet* are the same obtained computing the distribution of the average hitting time.

VI. GRAPH TOPOLOGY

We want to see if the topology of the bubbles has effects on the results we presented. To do so, we run a BFS from each node of the network and, for each depth of the exploration, we study the average exploration rate starting from Democratic and Republican pages. For this purpose, our interest is on the network underlying structure, thus we consider the unweighted *PoliNet*.

Result 3: At a given depth of exploration, travelers starting from Republican pages reach a percentage of Republican nodes greater than the percentage of Democratic nodes touched starting from a Democratic article. It seems that the Republican bubble is composed of pages that point to each other more than those inside the Democratic bubble.

The main effect of Result 3 on users' navigation paths is that for short paths, readers starting from Republican pages are more likely to stay in their seed node chamber, as shown in V-A.

VII. LIMITATION AND FURTHER WORK

The presented analysis have few limitations. The first drawback is that of using categories to label node. Indeed, since categories are not perfect, we can miss some articles that have not been inserted in any of the four categories (and corresponding subcategories). On the other hand, the readers analysis is based on aggregated data (*clickstreams*) thus the kind of analysis one can run is restricted and nor completely descriptive of the phenomenon. Despite these limits, that can be tackled, we consider this poster as a first step to shed light on the problem of bias in Wikipedia underlying network structure. Early results look promising for extending the analysis on other *topic induced networks* and defining metrics that help Wikipedia community to deal with the phenomenon.

REFERENCES

[1] F. Lemmerich, D. Sáez-Trumper, R. West, L. Zia. Why the World Reads Wikipedia: Beyond English Speakers. WSDM'19.
[2] K. Garimella, G. De Francisci Morales, A. Gionis, M. Mathioudakis. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. WWW'18.
[3] T. Piccardi, M. Catasta, L. Zia, R. West. Structuring Wikipedia Articles with Section Recommendations. SIGIR'18.
[4] D. Lamprecht, K. Lerman, D. Helic, M. Strohmaier. How the structure of Wikipedia articles influences user navigation. New Review of Hypermedia and Multimedia 2017.
[5] E. Wulczyn, D. Taraborelli. Wikipedia Clickstream. Website 2015.
[6] H. Gropp. Configuration and graphs. Discrete Mathematics 111 1993.
[7] M. E. J. Newman. Modularity and community structure in networks. National Academy of Sciences of the United States of America 2006.
[8] P.H.C. Guerra, W. Meira, C. Cardie, R. Kleinberg. A Measure of Polarization on Social Media Networks Based on Community Boundaries. ICWSM 2013.
[9] K. Garimella, G. De Francisci Morales, Aristides Gionis, Michael Mathioudakis. Quantify controversy on social media. WSDM 2016.