everyday specialization

the coherence of editorial communities on wikipedia

Nathan Drezner

September 2020

McGill

.txtLAB COLLABORATIONS

How is the specialization of knowledge reflected in the behaviour of lay readers? While we know that expert systems of knowledge, such as universities, are strongly defined by practices of disciplinary specialization, does this remain true for more general readers and producers of knowledge? Using Wikipedia as a resource, our aim in this collaboration is to better understand the degree of segmentation that occurs with respect to editorial behaviour across different knowledge domains. Do we see similar levels of domain coherence when it comes to editorial activity on Wikipedia?

Recent work on editorial behaviour on Wikipedia has focused on features such as linking structures, leading to discoveries regarding normalizations in governance and disparities in gender coverage. Other work has focused on the time series of edits and revisions as a way to model editorial conflict and resolution. These studies have illustrated some of the ways editorship on Wikipedia shapes the kind of knowledge available through one of today's most important and widely-used information repositories.

Our work aims to complement this on-going line of research by studying **the ways editors on Wikipedia move across different fields of knowledge and investigate which fields inspire generalized editorial behaviour and which fields inspire more specialized behaviour.** Using social networks to model user behavior on Wikipedia, we measure the degree of community coherence and intradomain crossover across four different knowledge domains: culture, science, sports, and politics.

We find that editors generally tend to reproduce a model of editorial specialization, with the notable exception of politics. We hypothesize that given the higher levels of conflict surrounding political debate today that editors make intentional efforts to signal non-partisanship in their behaviour.

To study and quantify the specialization and segmentation of editorship on Wikipedia, we focus on two distinct questions:

1. *Editorial Engagement*

How do editorial levels of engagement differ between different domains with respect to the number of editors, edits, length of edits and time between edits?

2. *Editorial Segmentation*

What is the extent to which users focus their commentary on single domains or fluidly move between them?

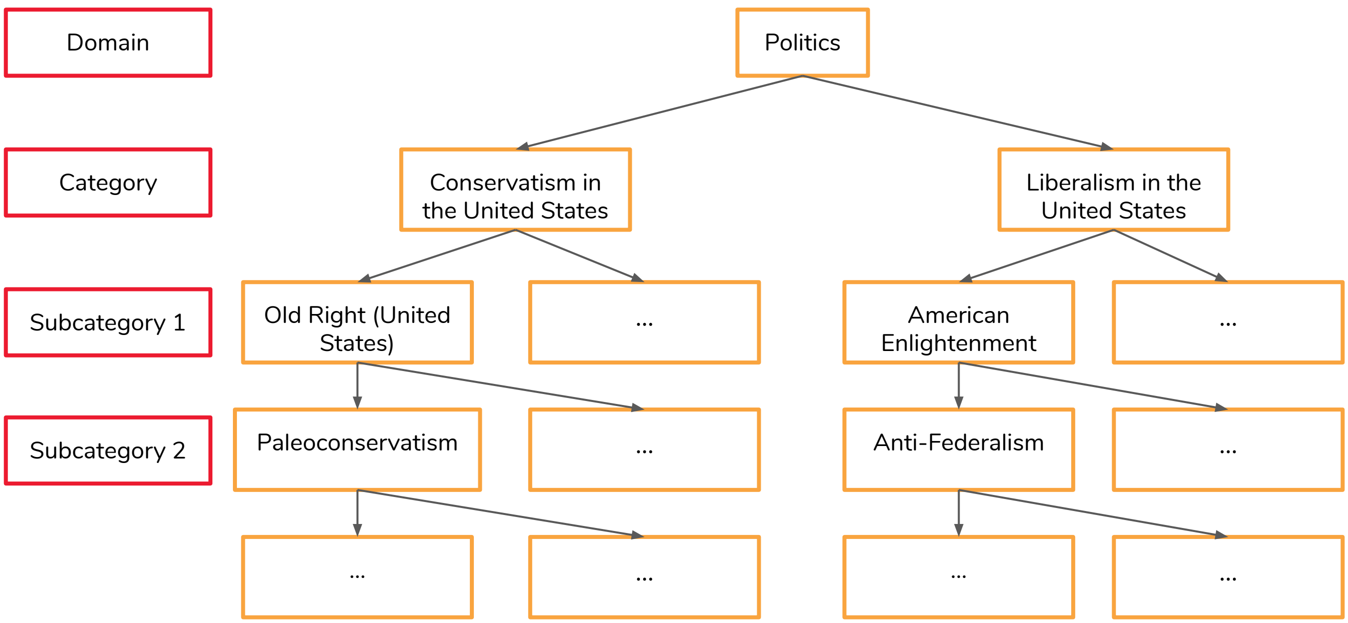
Data

The data in this study consists of a set of **30,054 Wikipedia articles**, including their metadata and edit history, whose subject matter represents four knowledge domains: culture, politics, science, and sports. These articles were sampled from Wikipedia category tags, which users add as an attribute to articles. Articles tagged as a subcategory of each category (to a depth of two sub-categories) were also treated as a member of that domain. Overall, **12 categories and 3,028 subcategories** of articles are represented in the dataset. This section outlines the selected domains for analysis, the reasoning behind that selection, and the methodology of data collection. All data and code are available here.

Capturing knowledge domains

To focus on several specific domains and the ways editors cross over between them, we use Wikipedia categories as a baseline of inclusion for each article into a domain. Because Wikipedia articles bridge a variety of styles—a given article can be anything from a biography, to a list, to a one- or two-sentence stub—selecting from Wikipedia categories allows us to sample across many different styles. Users on Wikipedia are able to add tags to articles marking them as a member of a certain category, and categories can be tagged as members of another category. This provides user-defined sets of articles whose subject matter is subsets of real-world social domains.

We focus on the general domains of **science**, **sports**, **culture**, and **politics**, and define each of these domains by user-defined Wikipedia categories. The graphic below illustrates the way categories nest and shows the categories and some of the subcategories for the domain of politics.



Sampling from categories

Every category also contains a set of articles tagged as belonging to that category. More than 370,000 articles are tagged by categories and subcategories up to a depth of two between each of the selected domains. From this, we randomly sample 2,500 articles from each category, distributed equally across levels. Because categories also contain some lists, templates, and user pages, the dataset is not a complete 32,500 articles. We cap the nesting at a depth of three based on a qualitative analysis of the relevance of articles at different depths.

Overview of selected categories

To control for national and linguistic barriers, we select categories for each domain with a focus on a North American framework, with the exception of science. For politics, categories focus on the primary category of US-American conservatism and liberalism. For culture, categories represent different mediums. For sports, categories represent the primary professional sports in North America, although they are not the categories for those leagues (those leagues do appear as subcategories). Four categories for popular fields of science represent the domain of science.

science	sports	politics	culture
Fields of mathematics	Ice hockey in the United States	Conservatism in the United States	American films
Branches of biology	American football in the United States	Liberalism in the United States	American novels
Chemistry	Basketball in the United States		Television in the United States
Concepts in physics	Baseball in the United States		

Table 1: Categories selected to represent each domain

It is worth noting the differences in syntax for each category title. Although they all belong to the same domain, there is no standardization in naming for each category by Wikipedia editors—for instance, names for each of the categories representing fields of science are different in syntax from one another.

Category behaviour

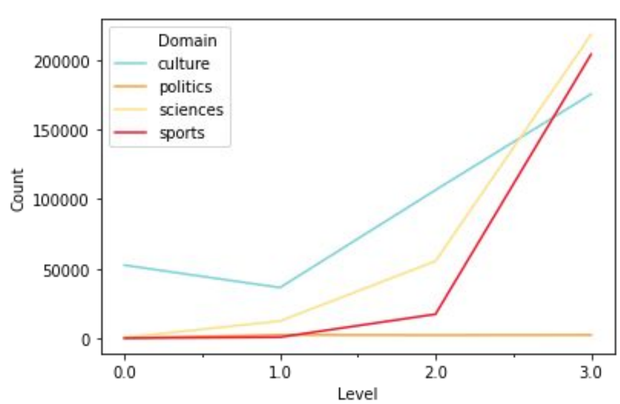


Figure 1: Overall count of articles represented by each category depth

Domain	Article count
Sciences	98,718
Sports	22,010
Politics	5,206
Culture	246,253

Table 2: Total number of articles encompassed by each domain

Each domain has a different count of articles at each level. In general, the number of articles represented by each category increases as a function of depth, as with each category a new set of subcategories is introduced. This is not the case for the politics domain, as it represents a more specific subset of articles.

Community engagement

To better understand the differences in behaviour across domains, we first study how editorial engagement differs between each of the domains. Given the diversity in content for each domain, our first question is whether different topics invite different levels of editorial activity.

To do so, we measure the number of editors per article, edits per user, and the time between edits as indicators of editorial engagement. Understanding which domains are shaped by greater numbers of editors, more active edits, or more timely edits allows us to assess differences in community engagement.

Studying editor interactivity

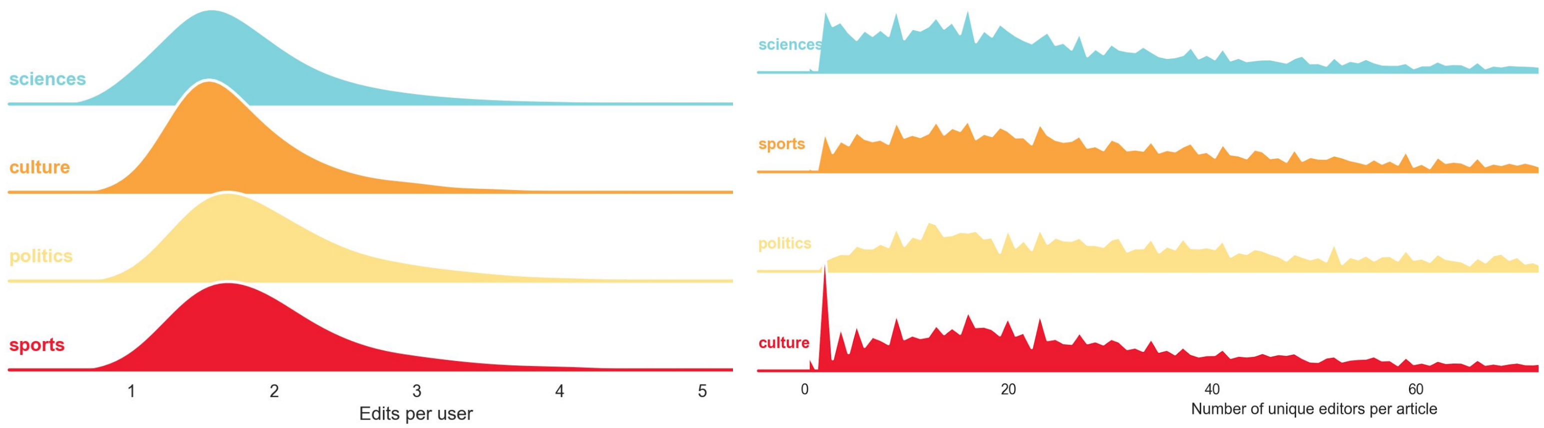


Figure 2: Comparison of user edits and the number of unique editors for the sampled articles

According the results of Table 3, we see how the domain of politics behaves somewhat differently than the other domains. **Articles within the political domain have significantly more unique editors, despite having a similar number of edits per user.** Articles from the political domain also tend to be edited more frequently, as indicated by the average number of days between edits. This suggests that the behaviour regarding political articles is distinct from the norms of other domains and favours more active engagement by a greater number of editors.

Domain	Unique editors per article	Average edits per user	Average days between edits	Average article age (years)
Culture	110	1.94	87.66	9.79
Politics	184	2.20	56.12	11.32
Science	105	2.01	140.25	10.81
Sports	93	2.15	73.54	10.06

Table 3: Comparison of user edits and the number of unique editors for the sampled articles

It is worth noting that, in general, editors tend to make an average two edits. This tendency is particularly interesting as it raises several questions: [1] whether those two edits are from the same article and [2] if they appear on different articles, whether those two articles are represented by the same domain or category.

This aspect of editorial behaviour signifies the ways users relate information back and forth, and we investigate this closely in our network analysis. The average of about two contributions per editor across all domains also illustrates how editorial patterns are characterized by one-off behaviour, where most editors contribute a low number of edits before "retiring."

Measuring editorial segmentation

In this next section, we study the extent to which editors concentrate their efforts within a particular domain or are equally likely to edit pages across multiple domains. To do so, **we model editorial behaviour as a social network**, where pages are represented as nodes and edges indicate when two pages are edited by the same editor. This allows us to observe the degree to which editors remain within a given domain or move across domains.

> **Do editors contributing to different domains of knowledge on Wikipedia tend to contribute as members of distinct sects of specialization or as generalists?**

We use two metrics to quantify this behaviour: 1) the Louvain algorithm for **community detection** to identify the domain purity of sub-communities within networks and 2) **nominal node assortativity** to study the likelihood of whether editors editing two different articles are editing articles from the same domain.

Why segmentation?

We see editorial segmentation as a useful way of understanding the degree of specialization on Wikipedia. Do editors tend to reproduce expert-driven systems that rely on high-degrees of specialization or, conversely, are editors on Wikipedia motivated by a desire to exhibit more generalist forms of knowledge that can be applied to multiple domains simultaneously. Interestingly, we find different levels of specialization corresponding to different types of knowledge domains. There is no one-size-fits all model of editorial behaviour.

Designing a network representation

To study the way that users tend to cluster around domains, we design social networks based on the set of articles curated across each of the domains. These networks are designed to represent the degree of connection between any two articles, where nodes represent articles and weighted edges represent the number of common editors between any two articles. Each network is composed of an equal number of articles (150) sampled from two separate categories **within the same domain**. Thus a sample network in the domain of "culture" will be composed, for example, of 150 articles randomly drawn from the category "Television in the United States" and 150 articles from "American Novels" and/or their related sub-categories.

We then generate a set of 5,000 random social networks using this sampling procedure where each network is composed of a single pair of categories from a given domain. The one exception is "All", where we build networks of categories drawn from different domains in order to establish a baseline of cross-domain behaviour relative to in-domain.

To generate a social network, we define edges as the number of common users between two articles. Nodes are tagged by title, domain, and category, attributes which are used to study class coherence relative to generated clusterings and for measuring the likelihood of nodes from the same category being connected.

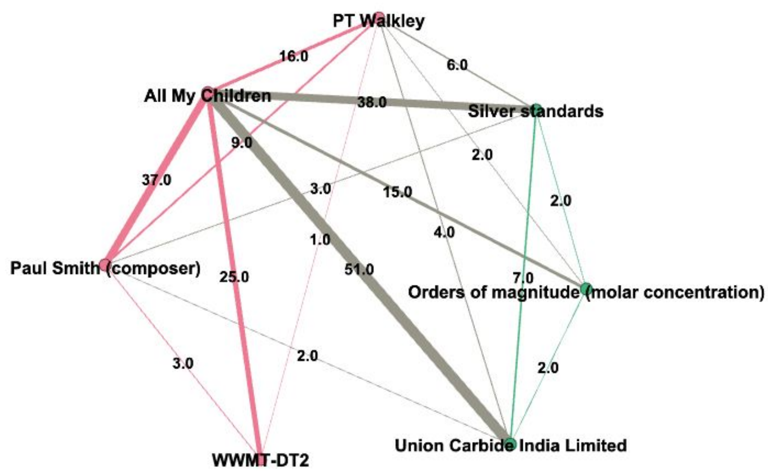


Figure 3: An example network with seven nodes from our All category. Nodes labelled in pink are from the category "Television in the United States" while nodes labelled in green are from the category "Chemistry." Edges represent the number of common editors between two articles.

The combined techniques of bootstrapping data to generate sample networks, weighted edges of common editors between articles, and metadata tagging of articles by domain and category allows us to study the behaviour of editors as a function of the categories of articles they tend to work on.

Community coherence

For each sample network, we then subset the network into communities using the Louvain method, including edge weights as a metric. Given that each network has nodes drawn from two categories, we condition on the two largest communities constructed by the Louvain algorithm. We then measure the "purity" of each community, i.e. the percentage of nodes in that community that belong to the more common category represented in that community.

A purity score of 1.0 would indicate that nodes that have clustered together all belong to a single category, while a purity score of 0.5 would indicate that categories are equally balanced within node clusters. Higher scores thus indicate that editorial behaviour tends to focus on editing articles drawn from the same category within a given domain.

As we can see in Figure 4 (left), the “All” set establishes a baseline of what to expect for cross-domain behaviour. While the purity of cross-domain segmentation is higher than all other within-domain samples -- suggesting that editors are highly likely to stay within domains -- we can see how editors of culture and sports pages are also strongly identified with single categories. We observe highly predictable behaviour when it comes to editing pages within the domains of sports and culture, with editors strongly favouring a single category (i.e. single sport or cultural medium).

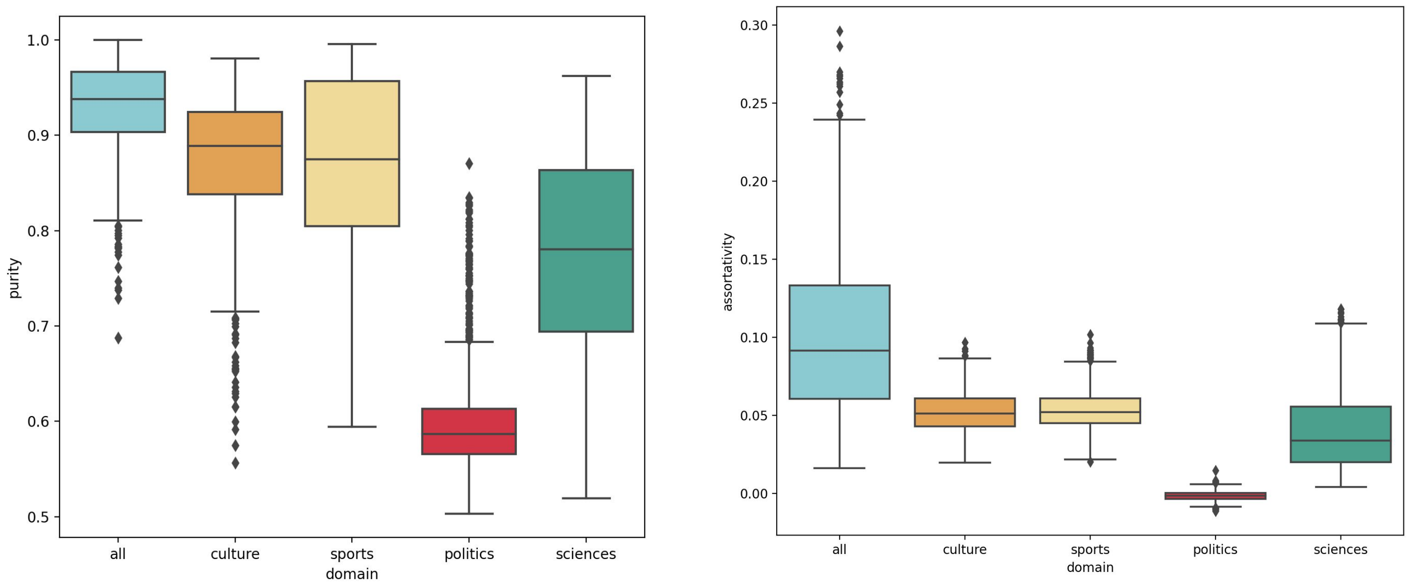


Figure 4: Left: Purity scores for the two largest clusters of each sampled network.
Right: Assortativity scores for sample networks by domain.

The domain with the lowest average purity score is politics, which approaches our 50% threshold with a mean of 0.559. This suggests that editors who engage with political pages do so in a much more non-partisan way when compared with editors of culture and sports. Although implicit political (and otherwise) bias still abounds on Wikipedia, this diversity of editor behaviour is an indicator that editors are attentive to issues of political bias and spend considerably more time editing across partisan divides. Finally, science indicates an interesting boundary case, where cross-category behaviour is considerably higher than in sports or culture, but not as extreme as politics.

Editors of conservative and liberal articles do not tend to group by political ideology.

Our purity scores indicate that editorial practices on Wikipedia differ strongly by domain. Although certain domains exhibit editing practices that are well-confined to a single category, other domains do not show this tendency. The results of this section indicate that “specialization” is a tendency that is domain-specific. We turn next to our assortativity measures.

Assortativity

In order to corroborate our first method, we also apply the method of assortativity, which examines the probability that nodes belonging to different categories will be joined by common editors. Assortativity is a wide-spread measure used to understand the idea of homophily, the degree to which social groups are defined by the homogeneity of individual identities (the "birds of a feather" phenomenon). The assortativity of each network is a measure of the probability that given an edge between two nodes, the nodes are either from the same category or different category. Higher scores indicate greater degrees of group homogeneity, while lower scores indicate more heterogeneity. As can be seen in Figure 4 on the right, our assortativity scores generally track our community detection approach, with politics once again indicating a strongly different type of behaviour overall. The generally low assortativity scores overall indicate that despite the strong clustering by category, there is still a fair amount of crossover among individual editorial behaviour. Specialization, while predictable, is not extreme according to our measures.

Domain	Mean assortativity	Mean purity
All	0.094	0.914
Culture	0.052	0.874
Politics	-0.001	0.599
Science	0.052	0.772
Sports	0.040	0.872

Table 4: Means across each domain for purity and assortativity. Assortativity varies between 1 and -1 with a higher score indicating nodes are more likely to connect to nodes of the same category, a score of 0 indicating no correlation, and a score of -1 indicating nodes are more likely to connect with nodes of the opposite category. Purity varies between 0.5 and 1 with 0.5 indicating a cluster with equal nodes from each category and 1 indicating only nodes from one category are in each cluster.

This is especially true when measuring assortativity across networks made up of two categories from different domains, where assortativity is far higher than in any other category. However, as with clustering, the domain of politics behaves differently with a slightly negative mean assortativity score, suggesting articles are just as likely to be connected to articles of the same ideology as they are to articles of the opposite ideology.

This segmentation between domains as a function of purity and assortativity shows how editors on Wikipedia tend to group, and provides an interesting contrast. Assortativity scores tend to be low, suggesting that although edges are likely to connect two nodes from the same category, there are many connections between nodes of different categories. Yet the distinct clustering—a mean of above 85% in two domains and above 90% in general—shows that despite many edges connecting nodes of different categories, in general groupings do form by category, further isolating the behavior of the political domain.

Conclusion

In this paper, we have investigated the ways editors on Wikipedia behave as coherent communities within knowledge domains. Using a set of 30,054 articles, representing twelve Wikipedia categories and four knowledge domains, we observe that some domains, such as culture and sports, indicate a high degree of specialization, while other domains, such as politics, indicate a very low degree. The sciences offer an interesting middle ground between these two extremes. We hope that this collaboration can prompt further studies of editorial behaviour on Wikipedia. How might these insights about the social network structures of specialization align with editorial language? Do we see similar levels of semantic coherence in editorial behaviour alongside domain coherence? One could also explore further knowledge domains to gain a better understanding of the distribution of specialization across domains.

Acknowledgements

This project would not have been possible without the generous support of many organizations and people. I would like to thank the Faculty of Arts Internship program whose funding allowed me to significantly expand my work on this project over the course of 2019. I would also like to thank Professor Richard So for kick-starting this project and helping guide it to fruition, in particular thanks to his love for Wikipedia. I would also like to thank the other members of the .txtLab who were always open to having new, odd ideas thrown at them as this project has evolved. Finally, I would like to thank Professor Andrew Piper for his support, guidance, and suggestions throughout the course of this project.

Works cited

DeDeo, Simon. "Conflict and Computation on Wikipedia: A Finite-State Machine Analysis of Editor Interactions." *Future Internet* 8, no. 3 (2016): 31.

Heaberlin, Bradi, and Simon DeDeo. "The Evolution of Wikipedia's Norm Network." *Future Internet* 8, no. 2 (2016): 14.

"Ideological Bias on Wikipedia." Wikipedia. Accessed June 8th, 2020. https://en.wikipedia.org/wiki/Ideological_bias_on_Wikipedia

Wagner, C, D. Garcia, M. Jadidi, & M. Strohmaier. "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia." *Ninth International AAAI Conference on Web and Social Media*. 2015.

Appendix

In order to collect data for this piece, I designed a modular tool to scrape information from Wikipedia in Python. This tool collects the complete history and metadata for an article on Wikipedia given its title. This tool is accessible via PyPi: https://pypi.org/project/wikipedia-histories/ and on GitHub: https://github.com/ndrezn/wikipedia-histories. The tool also includes functionality to recreate this study using the same or different domains and categories, including building and analyzing social networks. I hope others have a chance to use this tool for their own experimentation, curiosity, and academic research.