# Wikipedia network: An analysis on society-related pages

**Mario Bianchi**
m.bianchi66@studenti.unipi.it
Student ID: 616658

## ABSTRACT

Wikipedia is the largest encyclopedia ever written in the history of mankind and it has become one of the most regarded reference works for many people. Its online nature has allowed a system of open collaboration, where each user who can access the website can become a contributor (with a few restrictions). The structure of the website is based on categories and subcategories, and each page is usually linked to many others. The aim of this study is to analyze the network of pages in the category of *Society* through the use of network science tools, and then to use community detection techniques to see if there is a match between subcategories and the communities discovered.
Moreover, the general consensus expects Wikipedia pages to be written in a neutral language: this paper aims at determining through the use of text analysis techniques if there is a difference in language among categories.
[1]

## KEYWORDS

Wikipedia, Social Network Analysis, Community Detection, Sentiment Analysis, NLP

[1]**Project Repositories**
Data Collection: https://github.com/sna-unipi/sna-2023-2023_bianchi/tree/main/data_collection
Network analysis: https://github.com/sna-unipi/sna-2023-2023_bianchi/tree/main/network_analysis
Network manipulation: https://github.com/sna-unipi/sna-2023-2023_bianchi/tree/main/network_manipulation_and_analysis
Report: https://github.com/sna-unipi/sna-2023-2023_bianchi/tree/main/report


*SNA '23, 2022/23, Università di Pisa, Italy*


**ACM Reference Format:**
Mario Bianchi. 2023. Wikipedia network: An analysis on society-related pages. In *Social Network Analysis '23.* ACM, New York, NY, USA, 10 pages.

## 1 INTRODUCTION

Wikipedia is an online encyclopedia written in more than 300 languages. Its peculiarity is its open writing system, which allows users to become writers and therefore to create new pages or to edit existing ones. This system has been the key to its growth. Founded in 2001, today it counts over 58 million articles, 6.6 million of which are written in English. It is structured following a hierarchy of categories and subcategories[2]. In particular, this study focuses on the *Society and social sciences* category, and more specifically on 14 subcategories under the *Society* category: *Education*, *Ethnic groups*, *Globalization*, *Government*, *Politics*, *War*, *Peace*, *Military*, *Activism*, *Rights*, *Finance*, *Mass media*, *Crime* and *Family*. While the topic of Society has been chosen because of its heterogeneity, the 14 subcategories appear to span across the spectrum of positive and negative connotation, which is interesting with regards to the analysis performed in Task 3.
The first intent of the study is to compare the categories set by Wikipedia with the communities of pages found by several community detection algorithms. The idea behind this experiment is both to check the performances of said algorithms and to test Wikipedia's partitioning into categories. After this first assessment, the analysis aims at comparing the type of language (more precisely, the use of positive or negative words) used in each page. While for some pages the result is easy to predict, many others appear more uncertain. After having studied the word frequency inside each category, the study ends with an experiment on Topic Modeling. In particular, after having performed said task on the corpus of texts from the pages and after having assigned each page to one of the topics found, the final partitioning is compared to the one set by Wikipedia.

## 2 DATA COLLECTION

No available dataset was suitable for the proposed analysis, therefore the data has been obtained through the use of

[2]List of Wikipedia's categories https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories

an API. Firstly the data has been organized as a dataset, in order to facilitate the tasks of data exploration and cleaning. Ultimately, the dataset has been used to create a directed graph, representing the network of pages.

### Selected Data Sources

As mentioned in the introduction, the 14 Wikipedia categories chosen for the analysis are *Education*, *Ethnic groups*, *Globalization*, *Government*, *Politics*, *War*, *Peace*, *Military*, *Activism*, *Rights*, *Finance*, *Mass media*, *Crime* and *Family*. These are part of the *Society* category, which is part of the *Society and Social sciences* macro category. The following data has been saved for each downloaded page:

- Title, which has been used as an *id* since it is distinct for each page;
- Category;
- Links to other pages;
- Text of the page.

The analysis focuses on pages written in English.

### Crawling Methodology and Assumptions

The data has been obtained through *Wikipedia API*[3], a Python package developed directly by Wikipedia. The first dataset obtained required some data cleaning, since the results included types like *Template*, *Module* and *Category*. Each page can in fact link to subcategories, too. For the aim of this analysis it has been considered preferable to include pages from heterogeneous categories, instead of focusing on a few categories and exploring all their subcategories. For this reason, for each one of the 14 categories, only the pages in the first two levels of subcategories were explored and added to the result. After performing the task of Sentiment Analysis, which will be described and analyzed in Task 3, the resulting dataset consisted of 14928 records and 5 attributes: *Page*, *Category Links*, *Text* and *Score* (i.e. sentiment analysis score ranging from -1 to 1).
Figure 1 shows the distribution of categories in the dataset. *Education* and *Government* appear to be the most frequent categories, while *Ethnic groups* is the less represented, with 214 records.

Through the use of the Python library *NetworkX*, the dataset has been used to build the directed graph. After removing from the column *Links* all the links to pages not present in the column *Page*, it was possible to build the first instance of the graph. Initially it counted 14928 nodes (Wikipedia pages) and 165882 edges (directed links from page to page). This graph was not weakly connected, thus it was considered useful to focus on the giant connected component, i.e. the largest weakly connected component. The resulting network counts 12762 nodes and 165568 edges.

[3]Wikipedia API https://www.mediawiki.org/wiki/API:Main_page

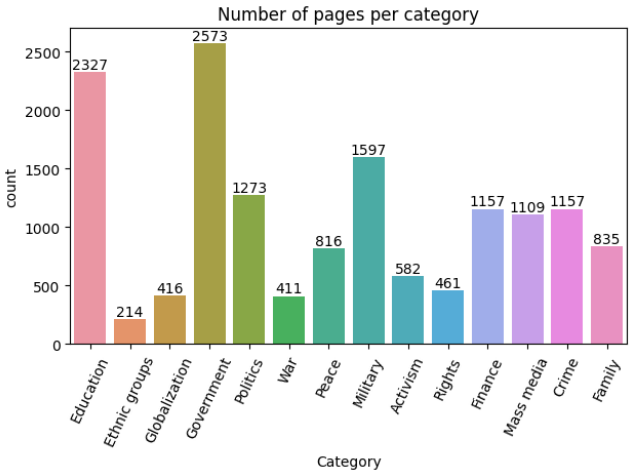


**Figure 1: Distribution of Categories in the dataset**

## 3 NETWORK CHARACTERIZATION

For the task of network characterization the network has been converted into an undirected graph with the same elements mentioned before: 12762 nodes and 165568 edges, with a network density of 0.00154, which makes it a sparse network. Figure 2, obtained with *Gephi*, shows a possible representation of the network, where categories are visualized in different colours. The average degree in the network is equal to 19.67387. Figure 3 shows the degree distribution among the nodes. As it will be further analyzed in the next sections, the distribution clearly follows a power law distribution, making it a **scale-free network**. Table 1 shows the pages with the highest degree.
The network, fully connected after removing the nodes not connected to the giant component, has a diameter of 10. The average path length is 3.76, while the average clustering coefficient is 0.36. In this regard, Table 2 represents the pages that form the highest number of triangles.
With regard to homophily, which is the tendency of similar nodes to be connected to each other with a higher probability, two different types of assortativity have been measured: degree assortativity, which measures the tendency of nodes to be connected to nodes with a similar degree, and attribute assortativity, where the attribute is the page category and which measures the tendency of nodes to be connected to nodes with the same label. Both range from -1 to 1. While the degree assortativity is 0.05, the attribute assortativity is 0.54, which denotes a high level of homophily with regard to the category.
The pages with the highest betweenness centrality are *World War II* (0.12), *Military* (0.07) and *World War I* (0.04). *World War II* is also the page with the highest closeness centrality (0.42), followed by *Democracy* (0.40) and *Human rights* (0.39).

**Table 1: Pages with the highest degree**

| Page | Degree |
|---|---|
| World War II | 960 |
| Military | 853 |
| World War I | 590 |
| Human rights | 523 |
| Democracy | 496 |
| War | 402 |

Figure 4 shows the degree of centrality over the different nodes.
Although the category *War* counts 411 pages, it is quite evident that some of its nodes are particularly relevant in the network, both in terms of centrality and in terms of triangles formed. In fact, pages in the categories *War* and *Military* appear to be the ones forming the highest number of triangles (*World War* 16539, *Military* 15792, *Military education and training* 14916, *War* 14562, *Lawfare* 14081, *War crime* 13830, etc.), and the same applies to pages with the highest closeness and betweenness centrality. A possible explanation is that the topic of war is often linked to a wide variety of situations, causes and effects. For instance it is easy to see how an article on the topic of war may be connected to pages in the categories *Ethnic groups*, *Government*, *Politics*, *Peace*, *Mass media* and *Crime*.

**Figure 2: Network representation on Gephi**

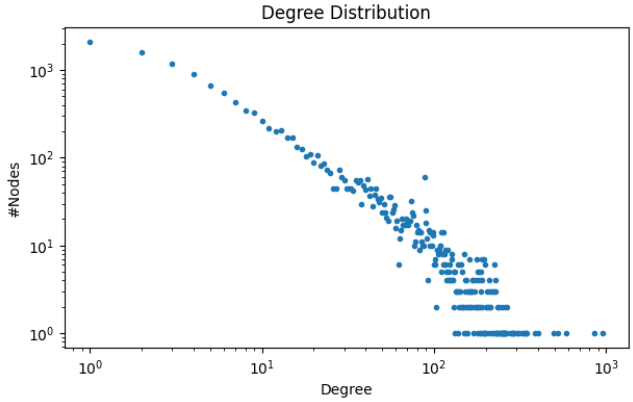


**Figure 3: Degree distribution of the Wikipedia pages network**

**Table 2: Pages that form the highest number of triangles**

| Page | Num. of triangles |
|---|---|
| World War | 16539 |
| Military | 15792 |
| Military education and training | 14916 |
| War | 14562 |
| Lawfare | 14081 |
| War crime | 13830 |

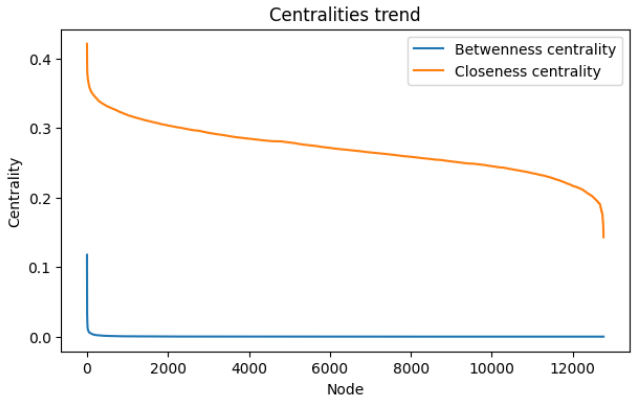


**Figure 4: Degree of Betweenness centrality and Closeness centrality**

## Comparison with Erdős–Rényi model

In the *Erdős–Rényi model* (ER), also called *random network* model, each other node can be connected to any other node with a probability $p$, that can be expressed as the average degree $\langle k \rangle$ divided by the number of nodes $n$ -1.
The value of $p$ for the built ER model is 0.0015, with 12762 nodes and 125656 edges. The difference in number of edges is attributed to the fact that the value of $p$ had to be computed

before building the network, which has been in fact built by specifying the number of nodes and $p$.
The average degree is 19.68 and Figure 5 shows the Poissonian degree distribution of the network compared to the scale-free degree distribution of the *Barabási–Albert model*, presented in the following section. This distribution is due to the fact that each node has the same probability of connecting to any other node, therefore the majority of the nodes have a degree close to the average degree in the network. This is a notable difference compared to the Wikipedia network, where each page has a higher probability of connecting to pages with a higher degree. The network density, expressed as the number of edges divided by the number of possible edges in the graph, is (almost) equal to the density of the Wikipedia network, with a value of 0.0015.
The diameter of the graph is equal to 5, with an average path length 3.51. As expected, the clustering coefficient is low and there are no nodes with a particularly high degree of centrality: the average clustering coefficient is 0.0015, and the highest number of triangles formed from one node is 5; with regard to centrality, the highest level of degree centrality is 0.0008, while the highest value of closeness centrality is 0.3112. The degree assortativity is 0.0012, which means that nodes are not homophilic in terms of degree. Table3 shows a comparison with the Wikipedia network and the *Barabási–Albert model*, which will be described in the following section.

### Comparison with Barabási–Albert model

The fundamental aspect of the *Barabási–Albert model* (BA) is the *power-law* distribution, typical of scale-free networks. This type of networks present a small number of nodes with an above average degree. These are called *hubs*, and they represent the tail of the degree distribution. The idea behind the model is that networks tend to grow with time, and while in random networks a new node would have the same probability of connecting to any other node, in the BA model a new node has a higher probability of connecting to a node with a higher degree. More precisely, this probability is equal to the degree of the node to be connected with, divided by the sum of all degrees.
In order to build a BA network, it is necessary to specify the value of $m$, i.e. the number of links each node establishes when joining the network. By dividing the number of nodes in the Wikipedia network by the number of edges, the resulting value of $m$ is equal to 9. The resulting network is made of 12762 nodes and 114777 edges. The average degree is 17.99: Figure 5 shows the power-law distribution of the network, which is the same distribution of the Wikipedia network. The network density, equal to 0.0014, is close to the values analyzed before. The network diameter is 5, with an average path length of 3.2. The average clustering coefficient is low,

**Table 3: Comparison among Wikipedia, ER and BA networks**

| | Wiki | ER | BA |
|---|---|---|---|
| N. of nodes | 12762 | 12762 | 12762 |
| N. of edges | 165568 | 125656 | 114777 |
| Density | 0.0015 | 0.0015 | 0.0014 |
| Regime | Conn. | Conn. | Conn. |
| Highest degree | 960 | 40 | 589 |
| Avg. degree | 19.67 | 19.68 | 17.99 |
| Diameter | 10 | 5 | 5 |
| Avg. path length | 3.76 | 3.51 | 3.20 |
| Avg. clust. coeff. | 0.36 | 0.00 | 0.01 |
| Degree assortativity | 0.05 | 0.00 | -0.01 |
| Highest betw. centr. | 0.12 | 0.00 | 0.04 |
| Highest clos. centr. | 0.42 | 0.31 | 0.46 |

with a value of 0.0075, but there is a noticeable difference with the ER model in terms of triangles: there are several nodes that form hundreds of triangles, with the highest values over 1000. There is no noticeable degree assortativity. With regard to centrality, the highest value of betweenness centrality is 0.0398, while the highest closeness centrality is equal to 0.4628.
Table3 shows a comparison among all the networks presented so far.

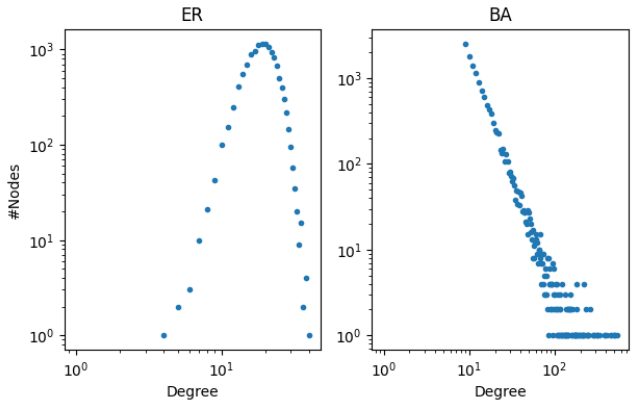


**Figure 5: Degree distribution in ER and BA models**

### Conclusions on Network characterization

As expected, the Wikipedia network is closer to a scale-free network than a random network. The power-law distribution is in fact expected to emerge in real-world networks. Some of the fundamental properties of a scale-free network are even more evident in the Wikipedia network. For instance the hubs, which in our case are the most referenced pages, are even more popular than the BA's hubs. The average

clustering coefficient is noticeably higher, too.
Another clear difference between the Wikipedia network and the synthetic graphs is the diameter: the real-world diameter is twice the diameter of ER and BA.

## 4 TASK 1: COMMUNITY DETECTION

Community Detection is an ill posed problem for several reasons. First, there is no unique definition of what a community is. The general consensus is that a community is a set of nodes more tightly connected to each other than to the other sets of nodes, or alternatively a community is a set of nodes where each node is closer to nodes within the community than to nodes outside of it. Apart from this, each algorithm implements its own definition of community. Another problem is that there are several possible approaches for the comparison of the communities found by the algorithms. In this study the evaluation measures considered are:

- Internal edge density: it is based on the notion that communities are sets of nodes densely connected to each other. In this regard, it is expected that a community presents a higher number of edges than what it would be expected in a random graph.
- Average node degree: it is based on the idea that since communities are densely connected areas of a network, the best partitioning is the one with the highest average degree.
- Modularity: it is a function ranging from -1 to 1. It measures the density of a community compared to the density expected in a random model. It is the foundation of many algorithms, included *Louvain's algorithm*, which maximizes the modularity score at each iteration.
- Conductance: since a community is a region of the network with a particularly high edge density, a good partitioning is a partitioning where an hypothetical flow in the network would have a hard time getting out of communities. In other words, the best partitioning is the less *conductive* partitioning.

The Python library used to perform community detection is *CDlib*, which implements both directed graph algorithms and undirected graph algorithms, therefore it has been decided to test both types. The algorithms chosen present different characteristics and belong to different algorithmic families, which means that they are based on a slightly different definition of community.
In the last section of this chapter the different results are compared with each other, while Task 4 will include a comparison between communities and the node label *category*, which could be considered the ground truth. In reality, as in many other cases from the real world, there is not a real ground truth in terms of communities, since pages of different categories may be strongly linked with each other and vice versa.

**Table 4: Community Detection - Directed Graph**

| | Infomap | EM |
|---|---|---|
| Number of communities | 894 | 15 |
| Largest community size | 600 | 8707 |
| Node coverage | 100% | 100% |
| Overlapping | False | False |
| Internal edge density | **0.90** | 0.01 |
| Average node degree | **4.80** | 2.69 |
| Modularity | **0.64** | 0.42 |
| Conductance | **0.36** | 0.78 |

### Community Detection on Directed Graph

The first algorithm tested on the directed version of the graph is **Infomap**, which belongs to the *Entity Closeness* family of algorithms. These types of algorithms consider communities to be sets of nodes that can easily reach members of their group, while it is more difficult for them to reach members of other communities. In other words the average path length inside a community is significantly shorter than the average path length in the graph. In particular, Infomap makes use of an optimization of the conductance measure while closely following Louvain's algorithm workflow, which will be described in the next section. Summarizing, since conductance is the measure of the *flow* among different sets of nodes, Infomap aims at minimizing this value for each community.
The second algorithm tested on the directed version of the graph is the **Expectation-Maximization (EM)** algorithm[1]. This iterative method, borrowed from the field of statistics, aims at finding structures in the network. It has been necessary to perform some parameter tuning. By setting Erdős–Rényi modularity as the partitioning quality score and by testing values of $k$ (i.e. number of communities) in the range between 1 and 19 with steps of 2 it has been possible to observe that $k$=15 gives the best overall results.
Table 4 summarizes the results obtained by the two algorithms.

### Community Detection on Undirected Graph

Three algorithms have been tested for the undirected version of the graph.
**Louvain's algorithm** (Louvain for short) belongs to the family of Internal Density algorithms. It is based on modularity: at first each node is assigned to a different community; at each iteration each node is moved to an adjacent community if the move increases the modularity score; communities are

**Table 5: Community Detection - Undirected Graph**

| | Louvain | Label Prop. | Infomap |
|---|---|---|---|
| Number of comm. | 27 | 391 | 787 |
| Largest comm. size | 2225 | 6543 | 451 |
| Node coverage | 100% | 100% | 100% |
| Overlapping | False | False | False |
| Internal edge density | 0.26 | **0.79** | 0.51 |
| Average node degree | **15.98** | 3.55 | 3.66 |
| Modularity | **0.69** | 0.55 | 0.64 |
| Conductance | **0.17** | 0.39 | 0.52 |

then collapsed in a single node, as it happens for Infomap.
The second algorithm tested is **Label Propagation**, which is a possible implementation of a *percolation* algorithm. In Label Propagation each node has a unique label, and at each iteration every node takes the label of the majority of its neighbours, until a stable state is reached.
The last algorithm tested is **Infomap**, implemented with the idea of comparing the result between the two graphs. Table 5 summarizes the results, which will be discussed in the next section.

### Conclusions on Community Detection

Overall, Louvain applied to the undirected graph provided the best results, ranking first by 3 of the 4 measures considered, while Infomap is the best algorithm by all measures between the 2 directed graph algorithms. Figure 6 provides an intuitive visualization of the results (*DG*: directed graph; *G*: undirected graph).
Since Louvain communities are the best possible partitioning, this has been compared to the ones obtained by the other algorithms through the use of *Normalized Mutual Information* (NMI), which is a measure of the overlapping between two partitionings. The following list ranks the algorithms by similarity to Louvain's result:

- Infomap (Undirected Graph): 0.54
- Label Propagation (Undirected Graph): 0.52
- Infomap (Directed Graph): 0.50
- EM (Directed Graph): 0.06

It is worth noting that every algorithm provided a widely different number of categories. At the same time, it is interesting that the algorithms that ranked first and last provided the closest number of communities (27 Louvain, 15 EM). EM has found the largest community, made of 8707 nodes, while Infomap is the algorithms that provides the smallest largest community size, well under 1000.
Figure 7 shows a possible visualization of the communities detected by Louvain's algorithm.

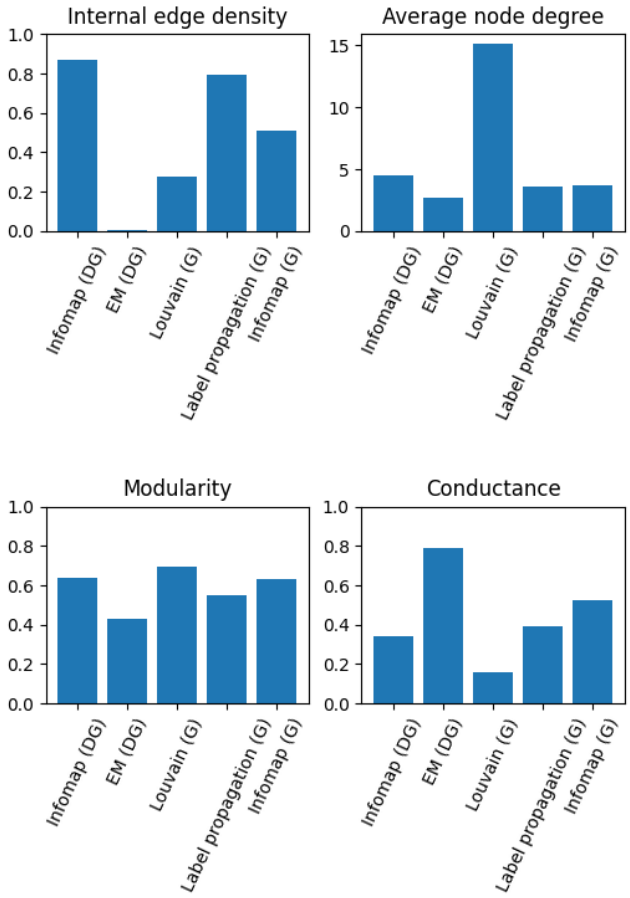


**Figure 6: Community Detection metrics**

## 5 TASK 2: FEATURE-RICH NETWORK ANALYSIS

Until this point the focus of this study has been the topology of the network. Although the analysis of the topology may reveal some interesting information about the network, it does not take into consideration a relevant aspect of the obtained data: the attributes. In fact, the nodes and edges from a real-world network contain many pieces of information useful to study the phenomena hidden behind the data, that would remain uncovered by a canonical graph analysis. This type of networks are called *Feature-reach network*.
The Network Characterization section reports the analysis on the assortativity based on the categories. Although such a measesure is useful to understand the general tendency of nodes to connect to nodes of the same category, the drawback of such a measure is its excessive simplicity and conciseness. It is in fact impossible to describe complex phenomena happening in a network with a single value. A novel approach in the field of feature-reach network analysis is the study of *conformity* [3]. Conformity is described as a

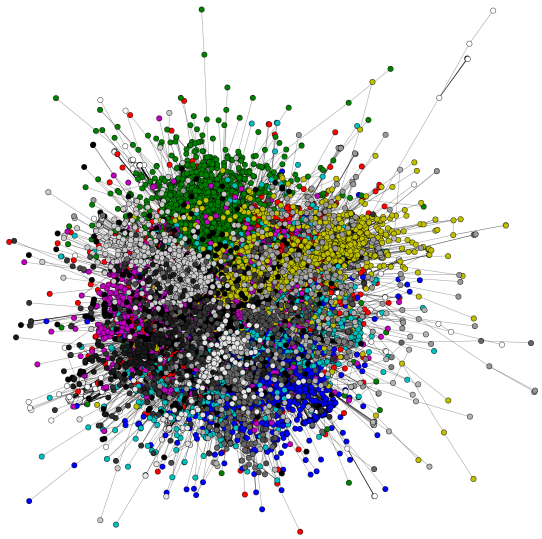

**Figure 7: Louvain's algorithm communities**

*node-centric path-aware measure*, in the sense that it gives results for single nodes and takes into account the distances to nodes of the same category. The idea is that in real-world networks nodes of the same category tend to stay close to each other. The aim of conformity is not to give a synthesis of the homophilic behaviour of the whole network, instead it aims at identifying the homophilic patterns at a local level. Conformity is measured as follows:

$$\psi(u,\alpha) = \frac{\sum_{d\in D} \frac{\sum_{v\in N_{u,d}} I_{u,v} f_{u,l_v}}{|N_{u,d}|d^{\alpha}}}{\sum_{d\in D} d^{-\alpha}} \quad (1)$$

where:

- $d$ is the distance between nodes $(u,v)$, and $D$ is the maximum distance in the network;
- $\alpha$ is a damping factor that controls the level of interaction between nodes, and which decreases exponentially when the distances between nodes increases;
- $N_{u,d}$ is the neighbourhood of node $u$ at distance $d$;
- $I_{u,v}$ is an indicator function that compares the values of nodes $(u,v)$that has value one if the attribute of $u$ $(l_u)$ is equal to the attribute of $v$ $(l_v)$ and zero otherwise;
- $f_{u,l_v}$ is a function that, if node $u$ shares the attribute $l_u$ with at least one node of its neighborhood, computes the share of the nodes in the neighborhood with the same attribute.

The result is that conformity computes for each node the similarity of its attribute to the others in the network, weighted for the distance. The score obtained ranges from -1 to 1.

Through the use of the Python library *Conformity*[4] it has been possible to calculate the conformity score for each node of the network. After testing with a few values of $\alpha$, $\alpha$=3 has been chosen because it gave more interpretable results. The furthest categories in terms of conformity (by looking at the average values) are *Education* and *War*, followed by *Education* and *Politics*. Figure 8 shows the conformity score of the nodes in these categories. while Table 6 ranks the categories by average conformity.

The following step has been to check if there is a correlation between conformity and any other topological attribute. In order to do so it has been necessary to build a dataset which included the following attributes:

- Page (node)
- Conformity
- Category
- Triangles formed from the node
- Betweenness centrality
- Closeness centrality
- Clustering coefficient
- Degree

After having calculated the correlation among these attributes, it has emerged that the highest correlation that involves conformity is a negative correlation with closeness centrality (-0.35). The second strongest correlation is a positive correlation with the clustering coefficient (0.23). The results are displayed in Figure 9.

In terms of single nodes, the pages with the highest and lowest values of conformity are different from the most relevant pages highlighted until now. They are marginal pages, with extremely low values of centrality, degree and clustering coefficient. On the other hand, pages like *World War I* and *World War II* have average values of conformity, respectively -0.25 and -0.18. This can be explained by the fact that the central nodes are connected to a greater number of nodes, both with the same label and with a different label, which results in an average score. Unlike central nodes, marginal nodes are more influenced by the few (and often by the single) nodes connected to them. The result is a more extreme score in conformity.

## 6 TASK 3: OPEN QUESTION

The communities detected by Louvain in Task 1 have been used as a reference to make a comparison with the labeling of the categories set by Wikipedia. After having conducted this experiment, the dataset is analyzed with text mining tools. Through the use of Sentiment Analysis it has been possible to conduct an analysis on the emotional tone used

[4]Conformity - Python library https://github.com/GiulioRossetti/conformity

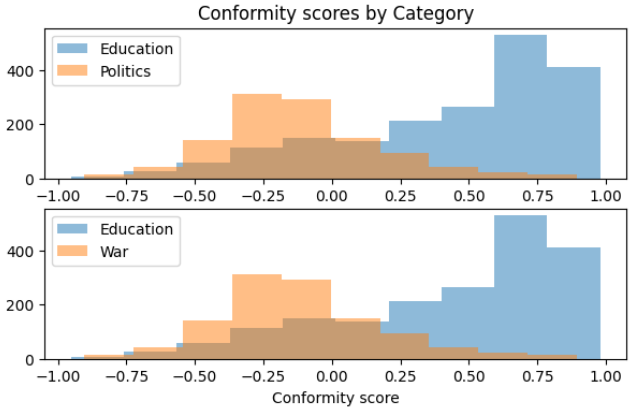


**Figure 8: Conformity for nodes in *Education*, *Politics* and *War***

**Table 6: Average conformity by category**

| Category | Conformity |
|---|---|
| Education | 0.44 |
| Finance | 0.37 |
| Mass media | 0.30 |
| Family | 0.29 |
| Ethnic groups | 0.25 |
| Government | 0.24 |
| Military | 0.21 |
| Crime | 0.20 |
| Peace | 0.06 |
| Globalization | 0.04 |
| Rights | -0.02 |
| Activism | -0.08 |
| Politics | -0.12 |
| War | -0-22 |

for writing pages in different categories. At last, after having applied a popular Topic Modeling technique, the dataset is split into topics, which have been compared to the categories partitioning.

### Communities and Categories

After having tried the different community detection algorithms mentioned in Task 1, the communities discovered are now compared to what can be considered the *ground truth*, that is the categories set by Wikipedia. In reality it is not a given that the labels assigned by Wikipedia match with the possible definitions of what a community is, but still it is an interesting comparison to make. In fact the experiment can be considered as a test for both the algorithms and Wikipedia's labeling system. The communities discovered

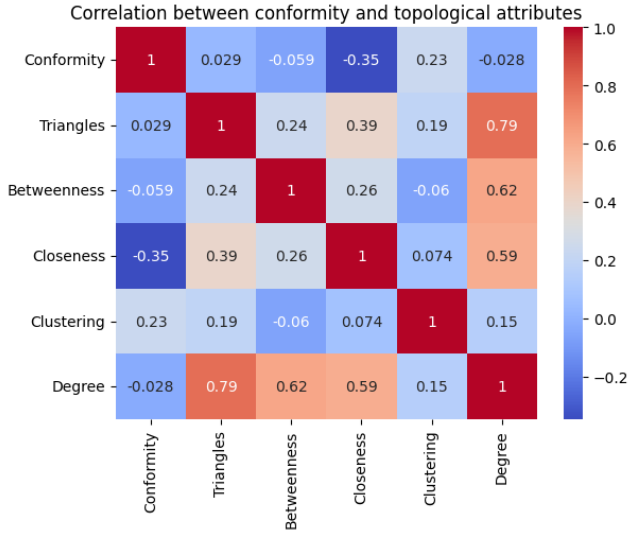


**Figure 9: Correlation between Conformity and topological attributes**

**Table 7: Similarity to Wikipedia's categories**

| Algorithm | NMI |
|---|---|
| Infomap (G) | 0.375 |
| Louvain (G) | 0.367 |
| Infomap (DG) | 0.366 |
| Label Propagation (G) | 0.365 |
| EM (DG) | 0.050 |

have been compared by calculating the Normalized Mutual Information (NMI). Infomap applied to the undirected graph is the algorithm that provides the partitioning which is closer to the one of Wikipedia, while EM had the worst performance. Infomap, Louvain and Label Propagation provided similar results, with an NMI around 0.37. The results are summarized in Table 7.

The communities based on the categories are non-overlapping and cover all the nodes in the graph. The biggest community contains 2263 element (close to Louvain). The ground truth partitioning got the following evaluation:

- Internal edge density: 0.025
- Average node degree: 15.797
- Modularity: 0.519
- Conductance: 0.284

All the results obtained are close to the ones obtained by Louvain, which scored above average results and has been considered the overall best community detection algorithm

for this study. The only exception is the internal edge density, which is significantly low. This indicates that the communities based on the category label have are not densely connected.

### Sentiment Analysis

Sentiment Analysis is the process of identifying the emotional tone behind some text. As many other fields of Natural Language Processing (NLP), nowadays it has become particularly relevant, since it is an efficient way to analyze and categorize the huge amounts of data produced everyday on the internet. The objective of the analysis proposed in this paper is to understand how different topics require a different emotional tone. While for some topics it is easy to predict if the words used have a positive or negative connotation, some are more difficult to predict *ex ante*.
Given the task proposed, it was important to choose the adequate Sentiment Analysis technique. After having considered the state-of-the-art, the favorite choice has been VADER[2], implemented by the Python library *nltk*. The algorithm belongs to the family of *weak labeling* (also called *weak supervision* or *semi-supervised learning*), which is based on the comparison to a set of pre-labeled data, which in the case of VADER is a lexicon. VADER compares the words found in the text and outputs four scores in the (-1,1) range: negative features, neutral features, positive features and a compound score, which is the overall score used for this analysis. The algorithm was designed originally for social media platforms, but in general it is particularly useful in situations when there is a large amount of data to catalogue. Proper supervised learning techniques would have been too computationally expensive.
By looking at the distribution of Sentiment scores (Figure 10) it is evident that VADER provides polarized results on this dataset. A large share of pages have a Sentiment score close to 1 or -1. Figure 11 shows the average sentiment score for each category. Predictably, the lowest scoring categories are *Crime*, *War* and *Military*. Some of the categories that should theoretically have a high score, like *Peace* and *Rights*, actually contain a lot of references to their opposites, thus the average score.

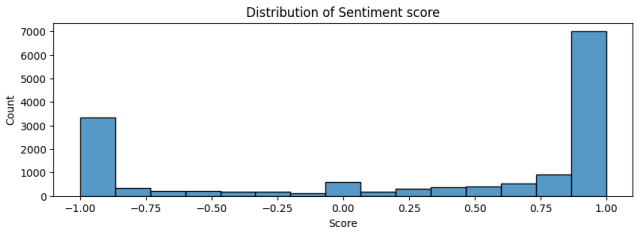


**Figure 10: Sentiment Analysis score distribution**

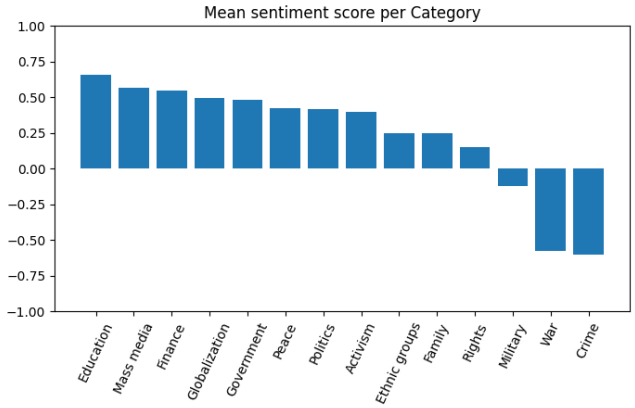


**Figure 11: Average Sentiment Analysis score by category**

Thanks to a quantile-based discretization function provided by the *Pandas* library, the Sentiment score has been discretized into 14 and then into 27 equally-sized buckets. This operation, called *binning*, has been performed to assign each page to one of the 14 bins and then to one of the 27, in order to compare the partitioning to the one provided by the 14 categories of Wikipedia and to the 27 communities obtained by Louvain. Such an approach by definition provides communities of (almost) the same size, which is a limitation. The 14-bins communities compared to the 14 categories obtain an NMI value equal to 0.056, while the 27-bins communities compared to the partitioning provided by Louvain have an NMI of 0.052. The experiment provided unsatisfactory results.

### Topic Modeling

Before performing the task of *Topic Modeling*, which is the task of extracting topics (or categories) from a corpus of texts, it was deemed interesting to visualize the most frequent words among the texts in the dataset. The most frequent **significant** word is *war*, with a frequency of 30516, followed by *government*, which appeared 29916 times. The same has been done for each distinct category, but the results were unsurprising.
After this preliminary analysis, the texts have been prepared for topic modeling, with the aim of extracting 14 distinct topics to compare to the 14 categories. The preparation involved the removal of stopwords, of punctuation and conversion to lowercase characters. The model applied is *LdaMulticore* provided by the *gensim* library.
The 14 communities (i.e. topics) obtained are non-overlapping and cover all the graph nodes. The largest community size is 2535, which is a result close to the one obtained by Louvain. The communities based on topic modeling obtained the following scores:

- Internal edge density: 0.021
- Average node degree: 12.467
- Modularity: 0.431
- Conductance: 0.385

While the evaluation scores for modularity and conductance would have been average among the other CD algorithms, the average node degree would have been second only to Louvain, while the internal edge density is extremely low. The comparison with the 14 categories obtained an NMI score of 0.269, which a score significantly higher than the one obtained by the approach based on the Sentiment score.

## 7 DISCUSSION

After having created an initial dataset through Wikipedia API, this data have been used to build a directed network of pages connected by the links amongst them. This network has been analyzed and compared to well-known synthetic models: from this comparison it emerged that the network obtained is a scale-free network and that the most popular pages (i.e. the hubs of the network) are mainly from the topics of war and military.
The following task focused on discovering communities in both the original graph and in the undirected version of the graph. Louvain applied to the undirected graph appeared to be the overall best algorithm for this analysis. Task 2 focused on the analysis of conformity, a measure of homophily from the field of feature-rich network analysis. This analysis showed a clear result: the pages with the highest absolute value of conformity are pages on the edge of the network.
The final task focused on comparing several partitionings with the labeling into categories set by Wikipedia. At first, the categories obtained by Louvain have been used to make such a comparison, and the results were not far off. Then, after having performed Sentiment Analysis on the texts of the pages in order to prove that pages from different categories are written with different emotional tones, the nodes have been divided on bins based on the Sentiment score, in order to make a comparison with Louvain and the original categories. The results were unsatisfactory. At last, 14 topics have been extracted from the texts thanks to a popular Topic Modeling library. Again, the resulting partitioning have been compared to the 14 categories, with better results than the previous experiment.

## REFERENCES

[1] Ahmed Ibrahem Hafez; Abaul ella Hassanien; Aly A. Fahmy; M.F.Talba. 2013. Community Detection in Social Networks by using Bayesian network and Expectation Maximization technique. (2013).
[2] C.J.Hutto; Eric Gilbert. 2014. VADER: A parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. (2014).
[3] Giulio Rossetti; Salvatore Citraro; Letizia Milli. 2020. Conformity: a path-aware homophily measure for node-attributed networks. (2020).