CS 224W Project Final Report
Culture Dependent Dynamics of the WikiLinkGraph

Ştefania L. Moroianu
slmoro@stanford.edu

Sarah Najmark
snajmark@stanford.edu

Alexandre Gomes
asimoes@stanford.edu

Abstract

Wikipedia, the online encyclopedia, has become one of the most accessible sources of knowledge for a variety of topics. In this project we analyze the community structure of Wikipedia and use it to identify the most prevalent topics in the information network. Our project includes a full time-dependent analysis of the evolution of the graph structure from 2001-2018, tracking the number of articles, links and communities they form. We study how the most important topics vary across four languages (English, Spanish, French, German) and comment on the aspect of cultural-dependent bias in the online encyclopedia. We also investigate how the largest clusters in these networks evolve through time, and whether the topics associated with them change. Lastly, we use the BERT language model to embed nodes based on their article titles and ask if these similarity between embeddings correlates in any way to community structure.

1. Introduction

Among the 287 different language editions of the Wikipedia website, it is natural that some topics will receive different levels of attention, according to cultural aspects. For example, it is intuitive to expect that there will be more articles in Portuguese related to soccer (players, clubs, etc.) than in Chinese. In this context, the present project draws a comparison across languages, here associated with cultures, aiming to find similarities and differences in the group of most important topics for each edition and study those similarities and differences across time.

There is a growing literature that focuses on cross-culture comparisons using the Wikipedia dataset of articles and their connections. Our work contributes to this literature by including a different level of comparison and adapting techniques for this context.

In this project, we focus on identifying important cultural-dependent topics of the Wikipedia network by using community detections for several versions of WikiLinkGraph - a dataset released this year - in different languages: French, English, Spanish and German. WikiLinkGraph is the graph representing Wikipedia where the nodes represent pages and the label of each node is the title of the page. The nodes are related if there is a link between the corresponding pages. We also analyze the evolution of those communities through time to better understand and compare the dynamics of the knowledge domains of interest for different cultures. We determine the topics of the biggest communities of the WikiLinkGraph for different years. We study the evolution of those communities using their sizes, the number of edges going out of the biggest communities because that gives an idea of their importance. As a last step in our analysis, we use a language model to generate embeddings for all the nodes in a graph based on their titles. We investigate whether nodes being close in embedding space correlates with them belonging to the same community.

2. Related Work

We found one study that focused on using the Wikipedia network to analyze cultural characteristics. This paper [1] compares important historical figures across different cultures using the 24 different languages editions of the Wikipedia dataset. They use PageRank and 2D PageRank to evaluate the importance of pages in the network. The study suggests an interesting approach to measure similarities and differences across cultures but it still presents some limitations. The simple comparison of most important historical figures may not be sufficient to represent the complex interactions among different cultures, which reflects in the opinions, values, beliefs and traditions of the people in each of them.

Therefore, in our work, we analyze the WikiLinkGraph at a mesoscopic level, focusing on sets of nodes grouped together - rather than single nodes - to give better insights about the trends of cultures. We group pages based on common topics to better understand which are the most important topics of interests within a culture.

While network analysis methods like PageRank and HITS algorithms, have been extensively applied to the Wikipedia graph [2], few studies have analyzed its community structure [3]. We talk about previous work on com-

munity detection algorithms and how we chose which one to use in the Methods section. Recently, researchers used PageRank on 24 Wikipedia language editions to gain insight into interactions of cultures [1]. We draw inspiration from this work to build our study of the WikiLinkGraph at a mesoscopic level. We aim at understanding how Wikipedia clustering structure varies through time and with language edition to get better insights about interests of people depending on their cultures.

Our cultural and time dependent mesoscopic study of the Wikipedia network enables us to give a more comprehensive view of the similarities and differences between cultures compared to previous work.

3. Data

The dataset we use in this project is "WikiLinkGraphs: A complete, longitudinal and multilanguage dataset of the Wikipedia link networks" [4]. The data can be found online at: https://zenodo.org/record/2539424#.XenHUOhKjb1.

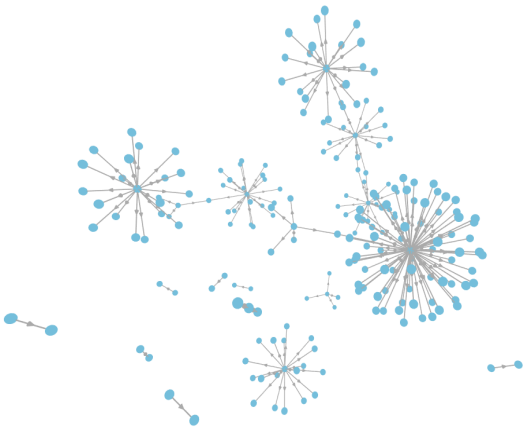

Figure 1: 3D visualization of a subgraph structure of the 2003 English edition from WikiLinkGraphs.

This dataset contains yearly snapshots of Wikipedia's internal link network for the 9 largest language editions (de-German, en-English, es-Spanish, fr-French, it-Italian, nl-Dutch, pl-Polish, ru-Russian, sv-Swedish). The dataset spans 17 years, from the creation of Wikipedia in 2001 to 2018. The snapshots are taken on March 1st of every year. What makes this a very clean network is that the authors parsed each revision of each Wikipedia article to track links appearing in the main text only, thus discarding links automatically generated by templates. Hence, WikiLinkGraphs represents all and only the links intentionally added by editors. The data is saved in csv format, as a list of records with the following fields:

- page_id_from, an integer, the identifier of the source;
- page_title_from, a string, the title of the source article;
- page_id_to: integer, the identifier of the target article;
- page_title_to: string, the title of the target article.

The result is a directed graph where each node corresponds to a Wikipedia article, and there is an edge from node A to node B if there is a link to article B in the text of article A.

Fig.1 shows an example of how the WikiLinkGraphs are structured. We used the interactive Python library jgraph to visualize a small subgraph in 3D.

To represent the graph along with its node labels (the titles of the articles), we use the networkx Python library, example in Fig.2. However, this type of figures quickly become overcrowded, so we only use them to illustrate small subgraphs. For example, in the figure below we can distinguish between two communities, one related to topics about Greek gods and another related to US historical events. One of the questions we're asking is if community detection algorithms will place nodes belonging to these topics in different clusters.

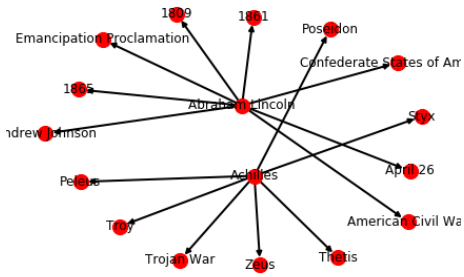

Figure 2: A subgraph of the 2003 English edition of WikiLinkGraphs. Nodes represent Wikipedia pages and the labels are the page titles.

4. Method

4.1. Community Detection

Central to our analysis of the WikiLink graph is the task of community detection. Network communities (also called clusters or modules) are sets of nodes with lots of internal connections and few external ones (to the rest of the network). Community detection is often used to understand the structure of large and complex networks, since it might reveal functional modules such as topics in information networks.

Identifying communities within an arbitrary graph is a computationally difficult task, and an exact solution to this

optimization is intractable [5]. Different methods exist for finding reasonably good partitions, and one of the most widely used is modularity maximization. Modularity is a benefit function that measures how well a network is partitioned into communities. Modularity values Q are in the range [–1, 1]. Q is positive if the number of edges within groups exceeds the expected number in a graph with uniformly random connections; Q greater than 0.3-0.7 means significant community structure.

The modularity maximization method searches over possible divisions of the graph for one or more that have particularly high Q. A popular such heuristic is the Louvain algorithm [5], which iteratively optimizes local communities until global modularity can no longer be improved with perturbations to the current community state. In our project we use the Leiden algorithm, a modified more complex version that outperforms Louvain both in terms of speed and quality of detected communities [6]. The next subsection briefly describes how this works.

4.1.1 From Louvain to Leiden

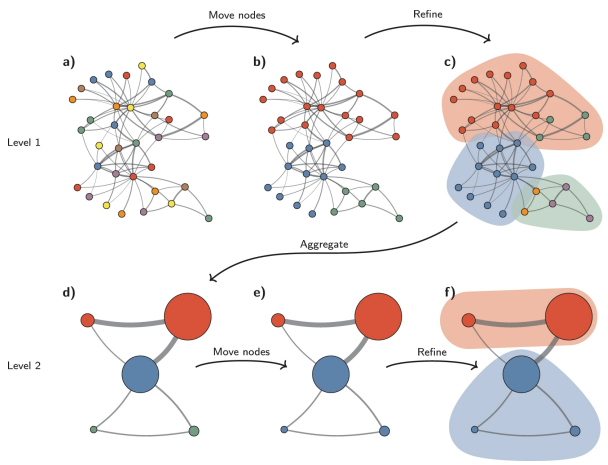

Figure 3: Leiden algorithm (figure reproduced from original paper [6]). In this example, the red community in (b) is refined into two subcommunities in (c), which after aggregation become two separate nodes in (d), both belonging to the same community. The algorithm then moves individual nodes in the aggregate network (e). In this case, refinement does not change the partition (f).

The Louvain algorithm is elegant and fast. Each pass of the algorithm has two steps: (1) local moving of nodes, and (2) network aggregation. In the first phase, it calculates the change in modularity when a node is moved from its present community to another one, and then the node is assigned to the community that yields the maximum increase. This is repeated for each node (going over nodes multiple times if required) until there is no further maximization possible. Then the graph is contracted (2), such that each community becomes a single node in the aggregated network. The steps are repeated until convergence.

The Leiden algorithm, an improvement to the Louvain method, progresses through the following phases:

(a) Start with a singleton partition.

(b) Local moving of nodes from one community to another to find a partition.

(c) Refinement of the partition.

(d) Aggregation of the network based on this refined partition. The non-refined partition is used to create an initial partition for the aggregate network.

(e) The local moving, refinement and aggregation steps are repeated until no further improvements can be made.

Figure 3 illustrates an example of how the algorithm works.

One major problem with the Louvain algorithm was that sometimes it yields arbitrarily badly connected communities, or even disconnected ones. The use of a refined partition in Leiden addresses this issue, thus guaranteeing that communities are well connected. This, along with a more rapid convergence time make the Leiden algorithm preferable in practice. Refer to the original paper for more technical details.

4.1.2 Overlapping Communities

We note that the algorithms discussed above work for strictly non-overlapping community detection. This has obvious limitations, especially in huge information networks like Wikipedia where certain articles could fit under multiple topics or categories. The overlapping community detection problem poses a higher degree of complexity, the run times are much longer and there are few methods that scale to huge graphs [7]. In the reaction paper we surveyed several possible methods[8, 9, 10] and found the one presented in reference [11] to be most promising: “Overlapping Community Detection Using Seed Set Expansion” (2013). Running this on the WikiLink network is a direction for further work.

4.2. Node Embeddings: BERT Language Model

To validate our results from community detections, we analyze the similarity of the titles of the pages in the same community. We compute the similarity of the embeddings of those titles using BERT [12] which is a powerful language model to get sentences embeddings. BERT stands for Bidirectional Encoder Representations from Transformers. It is currently the most powerful language model in a series of benchmark datasets. Unlike previous language models,

BERT gives deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. BERT is pre-trained on a set of NLP tasks and then fine tuned on a different set of NLP tasks. To get our titles embeddings, we use a BERT model pre-trained on the entire dataset of Wikipedia English texts. The final embedding for each title consists of a vector with 768 dimensions. Based on these node representations, we also tested an alternative approach for community detection that used K-means, a traditional unsupervised learning method, to cluster embeddings.

4.3. Our Workflow

- We partition the graph into non-overlapping communities using the Leiden algorithm.
- Separately, we use BERT to generate node embeddings based on the article titles.
- We are looking to see if similarity between BERT embeddings of article titles translates to node similarity in the WikiLink graph. We use BERT on the English WikiLinkGraph only because BERT was pretrained on English sentences from Wikipedia. We use belonging to the same community as the primary measure of similarity between nodes in the original network. We use cosine similarity to compute similarity score between two BERT vectors.
- Identify the most important communities in the graph and see if they correspond to certain topics (e.g. history, politics, entertainment etc.). We use community size as a measure of importance, but recognize this is a coarse metric, especially since the Leiden method is a hierarchical clustering algorithm – hence the largest community will likely encompass several smaller clusters corresponding to different topics, making it hard to define a one-to-one correspondence between communities and topics. We considered possible ways to refine this in the future, such as having an average community PageRank.
- Repeat the process for each time snapshot of the WikiLinks English graph. There are 18 snapshots for each language edition, one for every year from 2001 (when Wikipedia was created) to 2018. Time evolution analysis.
- Perform the same investigation for other language editions in addition to English (see Fig.4). We chose French, Spanish and German, for several reasons: first, we understand the languages well enough to make a pertinent comparison between article titles; second, French and Spanish are some of the most widespread languages in the world [13].
- Identify the topics reflected by the most important communities in each culture and compare.

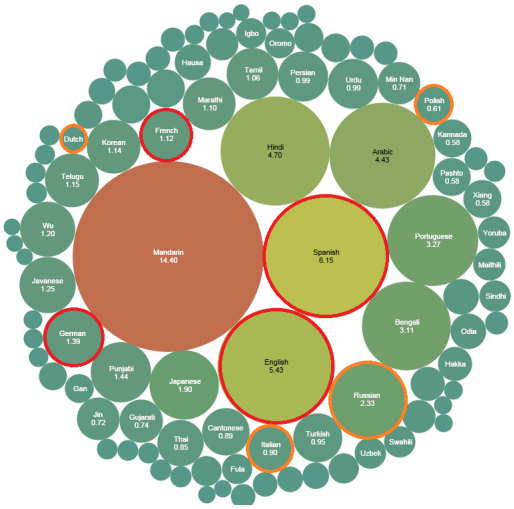

Figure 4: Bubble chart of languages by proportion of native speakers worldwide (source:[14]). Circled in red and orange are the languages available in the WikiLinkGraphs dataset – these correspond to the largest Wikipedia editions. The red circles indicate the four languages we look at in our study.

5. Results

5.1. Community Detection

We ran the Leiden community detection algorithm for the chosen four languages, for all the years 2001-2018. We then were able to investigate the size and structure of the graphs at a macro- and mesoscopic level. As an example, Fig.5 shows a number of small communities, with various structures.

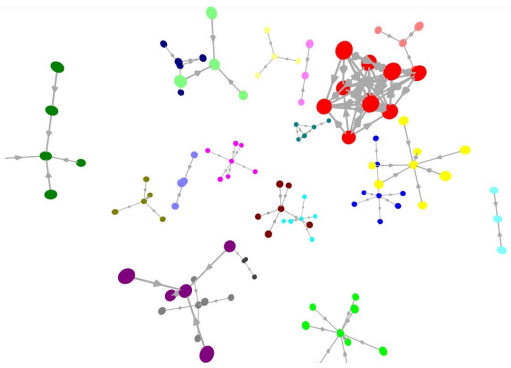

Figure 5: Visualization of a subset of detected communities. Different colors represent different communities as assigned by Leiden algorithm.

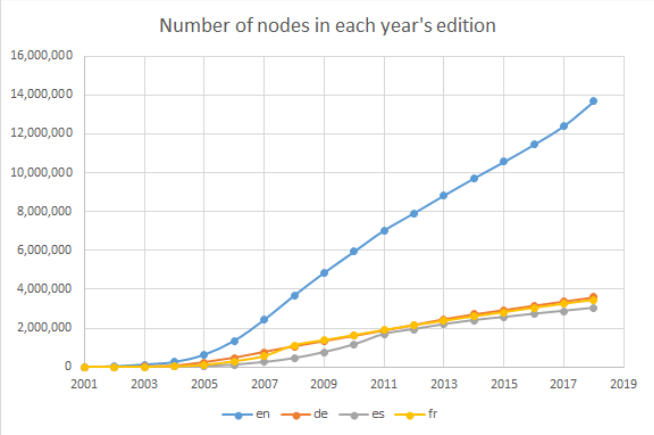

Figure 6: Number of nodes of the Wikipedia network from 2001 to 2018 for the english, german, spanish and french editions.

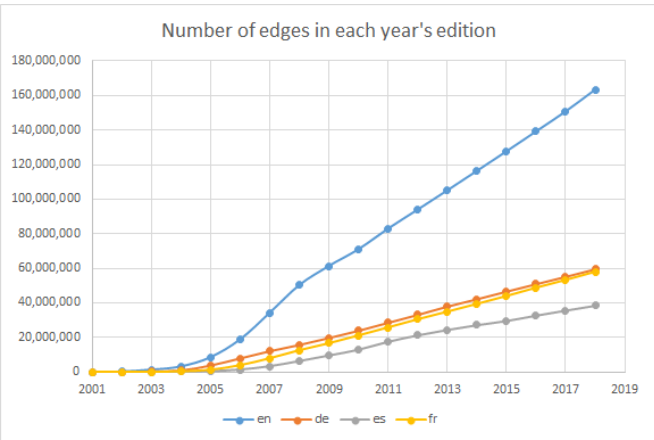

Figure 7: Number of edges in the WikiLinks graphs from 2001 to 2018 for the english, german, spanish and french editions.

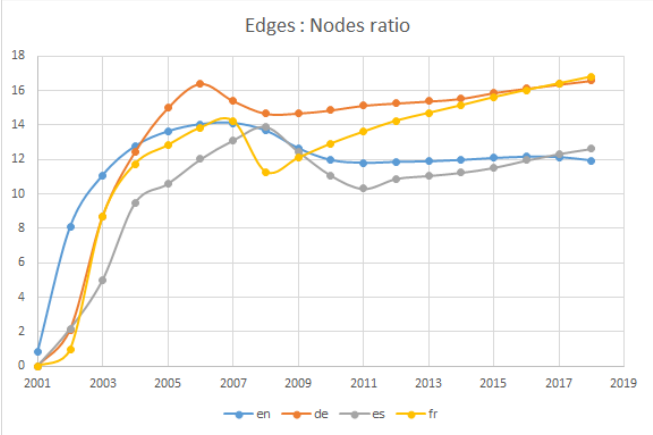

Figure 8: Ratio between the total number of edges and total number of nodes in the WikiLinkGraphs, for every year between 2001 and 2018.

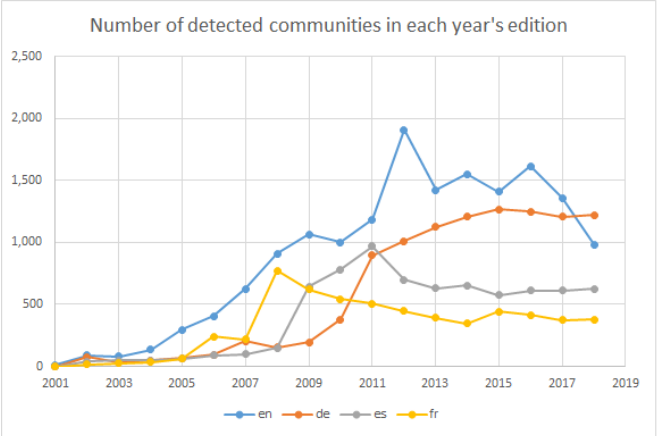

Figure 9: Number of detected communities in the WikiLinks network from 2001 to 2018 for the english, german, spanish and french editions.

First, we look at the total number of nodes (Fig.6) and edges (Fig.7) to see the macroscopic graph structure and how the network grows with time. We observe that while the Spanish, German and French editions have a similar growth rate, the English edition is expanding much more rapidly. Looking back at the chart showing number of native speakers around the world (Fig.4), we can see the bias for contributors writing English articles, given its status as the main international circulation language, especially on the internet. It's also interesting to note that since 2007 all the language editions have been growing at a linear rate, showing no signs of plateau in the number of nodes yet. The same linear growth rate is observed in the number of edges, or links between articles.

One interesting thing is that even though the number of nodes between English and the rest varies by an order of magnitude (14 million nodes in English 2018 compared to less than 4 million in the other 2018 editions), the ratio between the number of edges and the number of nodes is approximately the same for all the editions, see Fig.8.

Figure 9 shows the number of non-overlapping communities detected with the Leiden algorithm, for each year in each of the four language editions. Several things to note here. First, the number of communities plateaus and does not keep growing as more articles are added to Wikipedia. Second, while the number of nodes and edges in the network increases monotonically, there are sometimes rather large fluctuations in the number of communities.

Our interpretation: the Leiden algorithm starts with each

node in its own community and progressively merges them until modularity is maximized. Whenever new articles are added, they like start off as isolated, especially if there is little content on that topic on Wikipedia. As time goes by and the network grows, there appear more articles pertaining to the same category, and so it is more likely that they will introduce links to each other. When this happens, we see a drop in the number of communities, as a result of smaller clusters merging.

Lastly, we see the connection between the dynamics of the egdes-to-nodes ratio and the number of communities. When the ratio stabilizes, the number of communities also plateaus. Also, note that the number of communities grows very fast in the 2003-2009 period, and then does not grow anymore after 2013 – by that time the platform was 12 years old and had the time to acquire articles spanning a range of topics. In other words, it seems the set of topics present on Wikipedia is "saturated": the content continues to grow, but adding new articles doesn't really create new topics anymore (diminishing returns). After all, this makes sense from the perspective of everyday experience: 10 years ago, it was not uncommon to search for a topic on Wikipedia and to not get results, or to find the "This article about X is a stub. You can help Wikipedia by expanding it"; nowadays, on the other hand, it is much more rare to not find a Wikipedia article in the top results of a google search, whatever the subject.

5.2. Evaluating the importance of the largest detected communities in the Wikipedia network

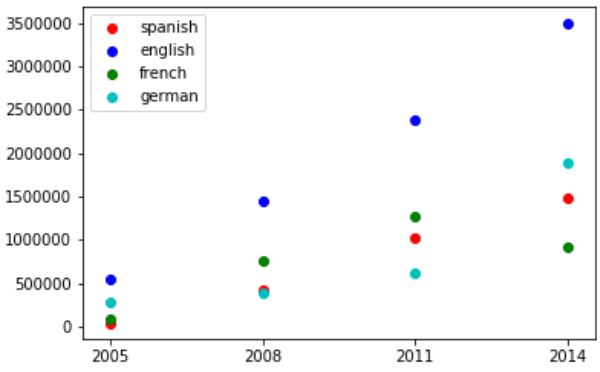

Figure 10: Evolution of the number edges connecting the largest community to the rest of a graph for the four language versions.

After detecting communities for different Wikipedia networks in different languages from different years, it is useful to assess the importance of the top detected communities to be able to answer to the question of the evolution

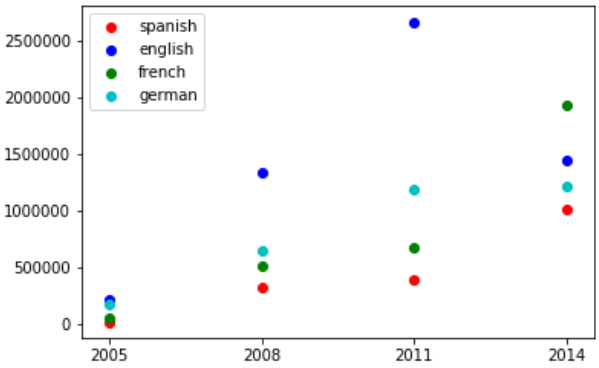

Figure 11: Evolution of the number edges connecting the third largest community to the rest of a graph for the four language versions.

of the important topics depending on the culture. We use the number of edges connecting the communities to the rest of the graph as a metric for the importance of communities but another idea would have been to transform communities into supernodes and use PageRank on them. Though probably being a more precise method, it would have been much more computationally expensive and our current method already gives a reasonable intuition about the evolution of the importance of the language-dependent communities through time. Using a simple PageRank algorithm to average the scores of the pages among a given community would have given erroneous results about the importance of a community within the graph because the PageRank scores of each page would have considered the scores of the pages within the community of that given page.

Using Fig. 10 and 11, we observe that throughout the years and for each of the language versions, the number of edges going out of communities usually increases as a result of the development of Wikipedia network which incorporates more and more pages as more and more poeple collaboratively built it. That also shows that the importance of top communities increases through time because they are more and more edges connecting communities with the rest of the graph.

We notice that for 2005 the top communities for German and English have a greater importance (around 10 times more edges connecting the communities to the rest of the graph) than French and Spanish. But throughout years, the importance of the top third community becomes much more similar for all four languages that may prove the Wikipedia networks for the different languages reach a closer level of development after some years.

But overall it is often the English version which always has more edges connecting its three top communities to the

rest of the graph at any year - especially the number of edges connecting to the rest of the graph is much higher for the top community.

Those observations make sense because the English Wikipedia network was the first to be developed and therefore reached its maturity earlier than the other networks which started being developed in 2001.

5.3. Qualitative comparison of topics between the languages

We analyzed the titles of the nodes in the top three biggest communities for the four languages and four years (2005, 2008, 2011 and 2014). That enabled us to assign topics to the groups of titles.

Language	Year	Topic 1	Topic 2	Topic 3
French	2005	Historical events	Places in France	Places around the world
French	2008	Historical events	Historical figures	Places in France
French	2011	Historical events	Historical figures	**Political figures**
French	2014	**Raw materials / energy**	Places in France	Political figures
English	2005	Historical events	Places in US	**Raw materials / energy**
English	2008	Political movements	Historical figures	History of religions
English	2011	US political culture	**European historical figures**	Artists
English	2014	Modern political movements	Historical political movements	Brain and psychology
Spanish	2005	Historical events	Worldwide geopolitics	Biology, medicine, botany, fauna
Spanish	2008	Historical events	**Sports (soccer)**	Historical figures
Spanish	2011	Worldwide historical events	Hispanic historical events	Nature and evolution
Spanish	2014	**Political movements**	Worldwide cultures	**Arts (music, cinema)**
German	2005	**Historical and cultural figures**	**Science, physics, engineering**	Cities in Germany
German	2008	Historical figures, artists	Cities around the world	**Science, biology, chemistry**
German	2011	Nature and geography	Important people, celebrities	Sightseeing, landmarks
German	2014	Cities and landmarks	Historical figures, famous people	Places and people outside Germany

Figure 12: Table highlighting some of the topics we identified as dominant in the largest 3 communities of the networks.

This table presents a qualitative analysis only, based on us scrolling through the title lists of each community and trying to identify the most prevalent types of articles. It is however not an exhaustive analysis. Even if we identify 'Historical events' as the main topic for the largest community in some editions, it does not mean that all the articles in that community thus proving our community detection method is not perfect. Nevertheless, some topics we identify are relevant to the culture and the year of the Wikipedia version. As Wikipedia encyclopedia is written collaboratively, the topics of the most important communities can also be interpreted as the topics of interests of the people who use it.

Overall historical events and historical figures seem to be two important topics for all of the four cultures.

For the French version of 2014, the pages of the largest community have titles related to raw materials and energy. Considering the environmental issues which happened in France at that time - such as the dramatic high levels of rainfall causing flooding in Var - it is intuitive that more and more pages around the topic of energy were written. We notice that the topic of raw materials and energy is also one of the main topics for the English but earlier in 2005 which is the year when the Energy Policy Act was signed.

For the French version, the topic of the third largest detected community is Political figures in 2011 that makes sense because the presidential elections in France happened in 2012. Therefore, more and more people at that time decided to write Wikipedia articles about French political figures. We observe that the same year, European political figures correspond to an important topic for the English version that proves the interest of the United States for the political events happening across the world.

For the Spanish version of the WikiLinkGraph we highlighted three topics which are quite representative of the Spanish culture. Sports and especially soccer was a crucial topic in 2008 during the European Football Championship of 2008. Spanish people are well known for their interest in that sport that is why there were that many pages about soccer in the community named Sports. In 2014, the rise of Podemos seems to have pushed people to write a lot of Wikipedia articles about political movements. The same year, the International Film Festival which happened in San Sebastian in Spain led to many new Wikipedia pages in the field of Arts.

The top community of the German version of Wikipedia from 2005 contains a lot of writers name, that is relevant to the focus of German people on Literature. We also notice that for several years, the topics of the main communities for the German edition are related to Science and Germany is historically well-known for the huge contribution of Science and Technology to its economy.

5.4. BERT Embeddings

We use BERT to compare the clustering of pages from similarity computation of embeddings of titles to the community detections we obtained using Leiden algorithm.

All the results presented in this section correspond only to the english edition of the WikiLinkGraph of 2005.

5.4.1 t-SNE

In order to make a first analysis of the similarity of the embeddings we used t-SNE to reduce their dimensionality. By representing them in two dimensions it was possible to visually evaluate the proximity of embeddings of nodes within and across leidenalg communities (Fig.13). We can observe that there is no clear correlation between the proximity of embeddings and the fact that the nodes belong to the same community. Only for the nodes in the green community there seems to be a significant correlation and this is composed mainly by names of localities in the United States, which is a type of semantic relation that is usually captured by sentence embeddings.

5.4.2 K-means

We evaluated an alternative method for community detection that does not take advantage of the network structure of the dataset. Our implementation used K-means with 20

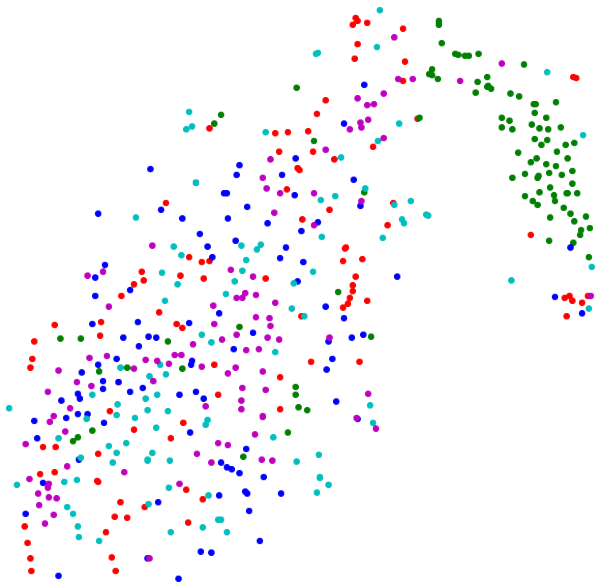

Figure 13: 2-dimensional plot of the title embeddings generated with t-SNE. Each leidenalg community is represented with a different color.

clusters since this was the number of communities with a significant number of nodes (> 50) generated with leidenalg. We compared the detected communities under both methods using the Rand index. This metric can assume values between 0 and 1, where 1 represents complete alignment between clusters. The algorithm that calculates the index makes a pairwise comparison between all items and counts how many pairs appear in the same community in both partitions, in only one of them, and in none of them. The value obtained in our case was 0.05, which indicates that there is almost no alignment between the communities detected with leidenalg and with the embeddings.

Additionally, we made a qualitative evaluation of the topics in the main k-means communities. Although this was not an extensive exploration (there are more than 600000 nodes in the network) we were not able to detect topics that are compatible with the context of the dataset. Some of the common features detected within clusters are aspects such as page titles that represent names of personalities, names of localities or dates. Broader aspects such as culture and history could not be detected in the embedding-based communities.

5.4.3 Cosine Distance

The last analysis with the Bert embeddings again explored the correlation between their similarity and the leidenalg communities to which the node they represent belong. We selected 5 of these communities and calculated the average cosine distance between nodes that belong to the same community and nodes that belong to different communities. The final result is represented in Fig.14.

Community	**0**	**1**	**2**	**3**	**4**
0	0.12	0.12	0.12	0.12	0.12
1	0.14	0.14	0.14	0.14	0.14
2	0.13	0.13	0.12	0.12	0.12
3	0.13	0.14	0.12	0.11	0.10
4	0.13	0.14	0.12	0.10	0.08

Figure 14: Cosine distance matrix.

Once again we can observe that there is no clearly larger distance between embeddings of different communities compared to the ones in the same community. Only for community 4, that is composed exclusively by names of Enochian angels, there seems to be a significant smaller distance for the nodes within community. This result can again be explained by the fact that word embeddings are usually capable of detecting names of people, especially in this case where the names are atypical.

5.5. Qualitative error analysis for detected communities

While some detected communities were quite consistent - for examples the community about Sports for the Spanish version of the WikiLinkGraph from 2008, it was tough to assign a topic to other communities since page titles were really different. That lead to using a really broad community name for the largest community of the English version of 2005 (see Fig. 12). For this version, historical events such as “Abraham Lincoln’s Burial and Exhumation” or “Black Hawk War” were mixed with titles of historical movies such as “Schindler’s List”. The focus is history for all of those pages so the community assignment makes sense but historical events names and historical movies titles are quite different entities. That proves that our community detection results make sense but are not perfect.

6. Conclusion and Further Work

As Wikipedia the online encyclopedia is written collaboratively by people who use it, we focused on leveraging the analysis of the temporal and cultural trends of the network to draw a parallelism with the evolution of the characteristics of four different cultures - English, Spanish, French and German culture.

We used the leidenalg algorithm to perform community detection on the pages of the network, hoping those communities would be based on the common topics of the pages grouped together. Though we were able to assign a global

topic to communities, we noticed that the community detections were not perfect.

However, in a comparison with a different topic detection approach that relied on embeddings of the page titles, we observed that leidenalg's communities are more expressive in this context. This is due to the fact that community detection algorithms for networks can detect semantic relations expressed through the connections of nodes. The embeddings-based approach is only capable of extracting meaning from the words and, since we are not using the entire text of the pages, there is considerable loss of information.

We found out that some topics of the largest communities for different versions of the WikiLinkGraph from different years were really representative of the cultural trends. For example, Sports and more especially soccer was a main topic for Spain and political figures was a fundamental topic for the French version during the presidential elections of 2012.

There are some limitations to this analysis of topics since some topics such as historical events and historical figures were mostly present for all languages and years so we are not able to discriminate cultures based on this criteria. This category is also really broad so it is not straightforward to deduce any specific interest for an historical period depending on the culture.

Overall, our cultural and time dependent analysis of the Wikipedia network enables us to validate some ideas about the characteristics of different cultures. But some improvement of the community detection method could be useful to refine the topics we determine from the communities and therefore get a better idea of the interests of people depending on their cultural origins.

We also noticed that the largest editions of the WikiLinkGraph do not reflect the largest speaking populations for example the Swedish or Polish editions are among the biggest, with a relatively small population of native speakers. Wikipedia primarily reflects western cultures knowledge. This introduced bias makes it an imperfect tool to analyze cultural characteristics.

Acknowlegdements

The three members of the team contributed equally to the project and report. Also, we would like to thank Michele Catasta for helpful discussions throughout the quarter.

References

[1] Y.-H. Eom, P. Aragón, D. Laniado, A. Kaltenbrunner, S. Vigna, and D. Shepelyansky, "Interactions of cultures and top people of wikipedia from ranking of 24 language editions," *PLoS ONE*, vol. 10(3): e0114825, 2015, https://doi.org/10.1371/journal.pone.0114825.

[2] F. Bellomi and R. Bonato, "Network analysis for wikipedia," *Proceedings of Wikimania 2005, The First International Wikimedia Conference*.

[3] D. Lizorkin, O. Medelyan, and M. Grineva, "Analysis of community structure in wikipedia (poster)," 18th International World Wide Web Conference, Madrid, 2009, http://www.ambuehler.ethz.ch/CDstore/www2009/proc/docs/p1221.pdf.

[4] C. Consonni, D. Laniado, and A. Montresor, "Wikilinkgraphs: A complete, longitudinal and multi-language dataset of the wikipedia link networks," 2019, https://arxiv.org/pdf/1902.04298.pdf.

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008. [Online]. Available: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008

[6] V. A. Traag, L. Waltman, and N. J. Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *Scientific Reports*, vol. 9, p. 5233, 2019, https://doi.org/10.1038/s41598-019-41695-z.

[7] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: the state of the art and comparative study," *ACM Computing Surveys*, 2013.

[8] R. Andersen, F. Chung, and K. Lang, "Local graph partitioning using pagerank vectors," *In FOCS*, 2006.

[9] M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "Demon: a local-first discovery method for overlapping communities," *In KDD*, 2012.

[10] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," *In WSDM*, p. 587–596, 2013.

[11] J. J. Whang, D. F. Gleich, and I. S. Dhillon, "Overlapping community detection using seed set expansion," *In Proceedings of the 22nd ACM international conference on information and knowledge management*, vol. CIKM'13, p. 2099–2108, 2013.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[13] "The most spoken languages," 2018, https://blog.esl-languages.com/blog/learn-languages/most-spoken-languages-world/.

[14] Wikipedia, "List of languages by number of native speakers," https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers#cite_note-Nationalencyklopedin-11.