Wikipedia Cultural Network

Applied Data Science Network Project

NYU CUSP

Jeremy Neiman

Avigail Vantu

**Introduction:**

Network Analysis is a methodology used to identify the relationship between different components and is used in fields such as transportation, human flow among others. Wikipedia is an open source multilingual encyclopedia and is biggest resource of worldwide written collaboration. We want to explore the structure of the Wikipedia network across different languages and see if language and culture influences the structure of the Network.

Wikipedia is structured into distinct sites for each language. For each language it is made up of Articles, as well as other types of pages, and intrasite links that go from an article to other articles of the same language. The same page in two different languages can and are likely to be completely different from each other, with a different set of other articles they link to. Therefore Wikipedia can be modeled as a directed, unweighted network of articles as nodes and links as edges, and each language’s site is unique and disconnected from each another language.

Wikipedia is a vastly used resource worldwide for all fields. Articles from one language can be translated to another language but more often they are written independently for each language (Ulrike et Al., 2006). By analyzing different languages we expect to get an understanding of the patterns and the importance of different subjects over others. In the attempt of unifying its articles Wikipedia governed a policy of neutral point of view (NPOV), which means the information provided should lack a bias and be given without a personal engagement (Ulrike et Al., 2006). Nonetheless, the cultural and personal biases of the authors almost certainly influence the content. Early research implies that an article quality and reliability improves as a dependency of the relevance of a certain article (Emigh et Al., 2005).

The cultural differences coming from geographical and languages variation derives from a complex arena. However, it does seem the roots for differences among nations and languages go way back in history (Ulrike et Al., 2006). Those differences can be better understood with the investigation of a spatial term coined by Kevin Lynch. *Mental Maps* describe the way people grasp their own environment in a way that differs from the objective reality. The mind enables us to collect and manipulate information about our spatial environment (Schenk, 2008). Therefore, people’s understanding and memories of places contain mainly qualitative characteristics, which create representation of a place. A good example of the mental map is the way most people refer to Eastern Europe nowadays. Although this term has a determinative definition of countries that geographically belongs to it Eastern Europe is primarily used to refer to the “other” and in particular to a geographical location that is placed eastern than the one referring to it (Schenk, 2008). The reason for this is the roots of this term invention and the development of it over time as a negative and less developed association.

We think the way different languages’ Wikipedia sites represent the same content can give insight into how that content is seen differently across the cultures. As the articles were mostly written in the origin languages and not translated and as Wikipedia allows open editing of each article the links chosen to be part of an article are not randomly chosen and indicates the editors’ choice to either include or exclude it. Therefore the links included in an article is an indicator of a personal and cultural perspective on a certain topic. In this way the structure of the Wikipedia network as a whole may be influenced by the culture. We hope the results of our network analysis could serve as tool of social and cultural analysis.

**Data and Methodology:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Language** | **Nodes** | **Edges** | **Articles** | **N / A Ratio** |
| **Turkish** | 87,500 | 351,000 | 275,000 | 0.32 |
| **Chinese** | 285,000 | 880,000 | 850,000 | 0.335 |
| **Japanese** | 250,000 | 812,000 | 1,000,000 | 0.25 |
| **English** | 550,000 | 2,000,000 | 5,000,000 | 0.11 |
| **Russian** | 208,000 | 645,000 | 1,273,000 | 0.164 |

Wikipedia provides an API that can, among other things, return the set of links on a page and also the corresponding page in each other language. Using this, we built a network of articles as nodes and links as edges for five languages – English, Turkish, Russian, Japanese and Chinese. For each language, we started from the same five points on Wikipedia – New York City, Istanbul, Moscow, Tokyo and Shanghai – and then followed links up to depth two away from those cities’ pages. So we ended up with networks for each of those languages’ Wikipedia sites which are subsets of their whole sites, centered around those five cities. Table 1 summarizes the resulting networks and provides the whole of each Wikipedia site as a comparison. We also used the Wikipedia API as a translation tool, using the list of interlanguage links to get the equivalent page from the English Wikipedia, when an English page was available.

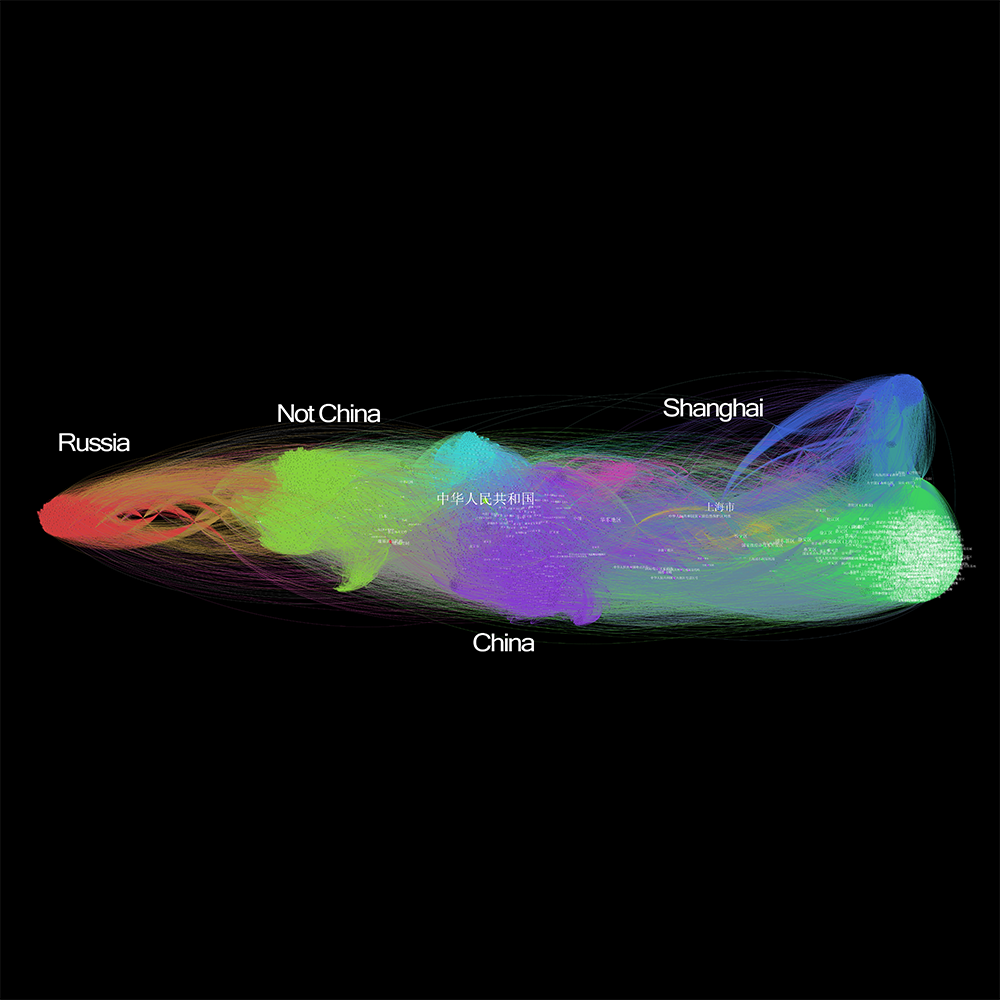
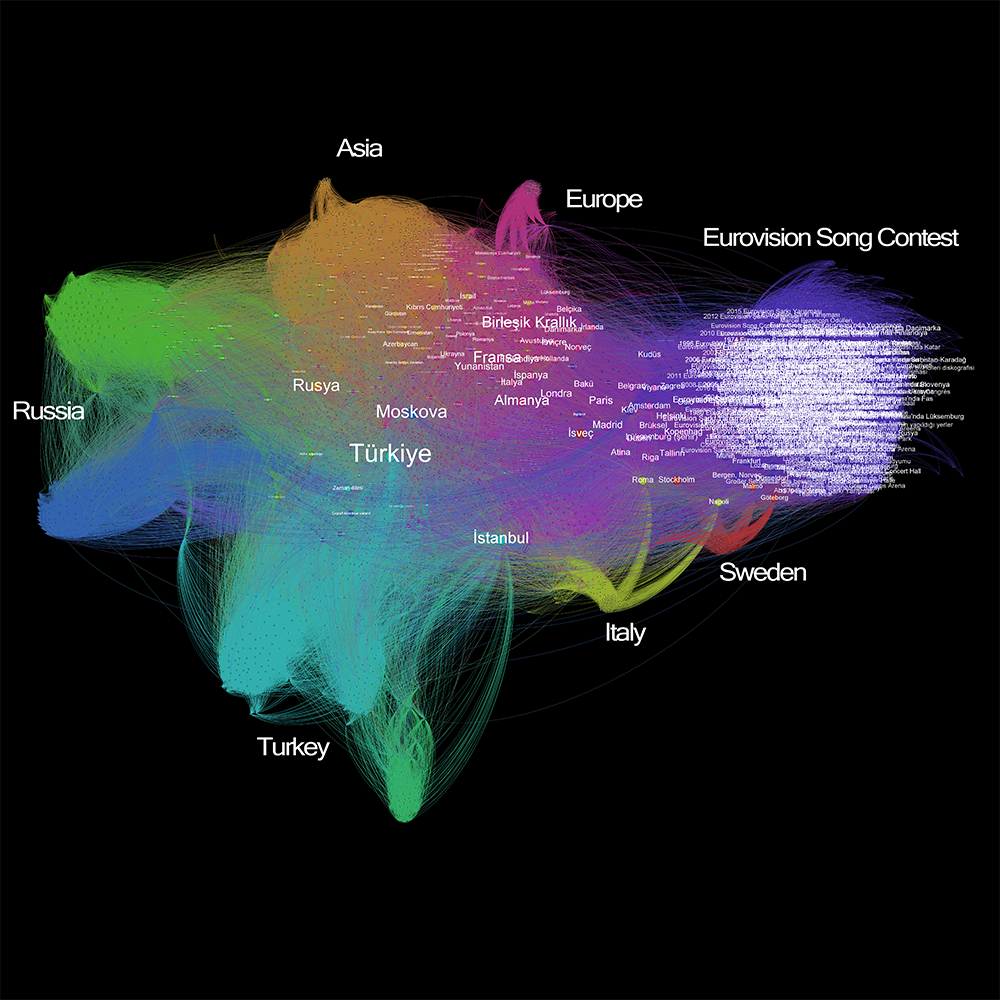
After building the networks using the Wikipedia API, we did the following analysis for each of the networks individually:

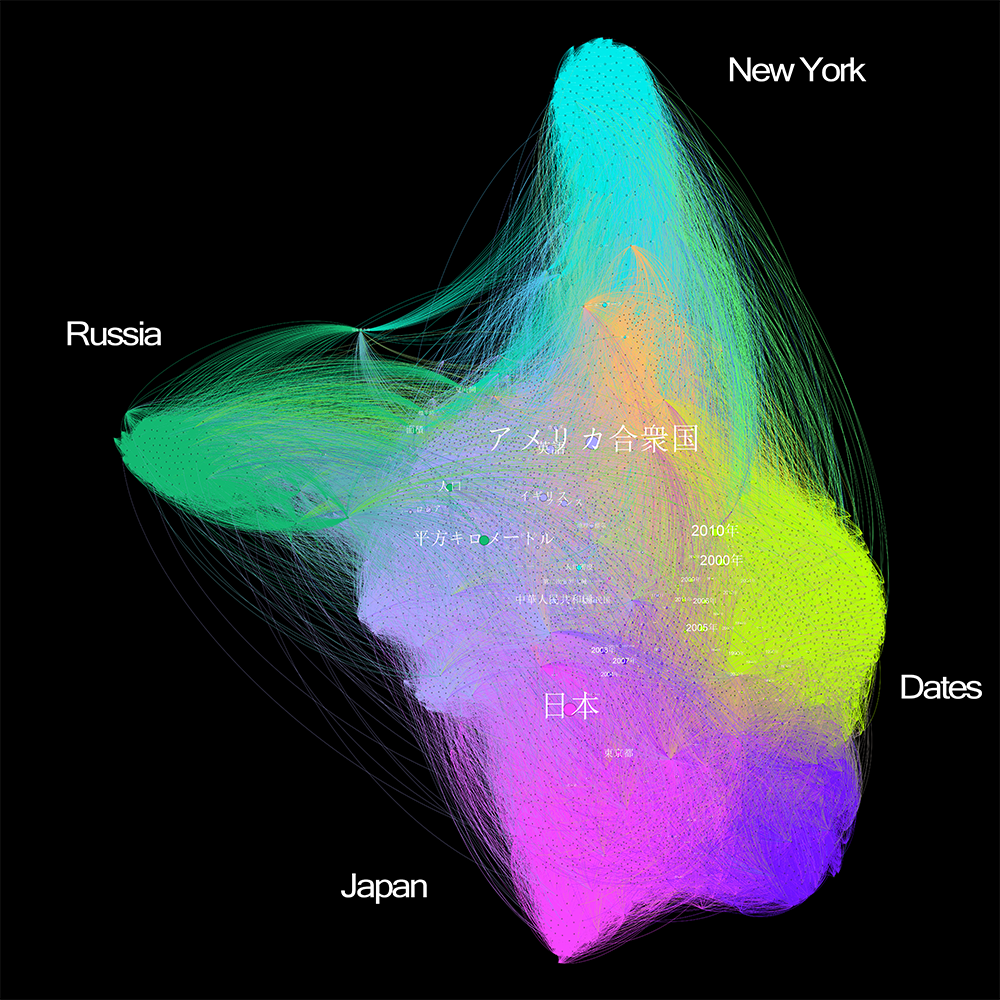
1. Used the eigenvector centrality algorithm to determine the most important pages in the network.
2. Filtered the set of nodes down to about the 4000 most central nodes to facilitate further analysis and visualization.
3. Used modularity to partition the networks into communities.
4. Visualized the networks.
   1. Nodes were laid out using the ForceAtlas2 algorithm. This is a gravity based model where each node exerts a repulsive force on every other node, and edges pull nodes back together. This results in nodes and communities that are central in the graph being visualized in the center, while nodes or communities that are not well connected to the rest of the network drifting away from the center.
   2. Nodes were colored based on the community that they were in and edges were colored with the same color as the node it came from.
   3. Node and label size is proportional to the eigenvector centrality of the node.

**Results**:

**Image 2: Turkish**

**Image 1: Chinese**

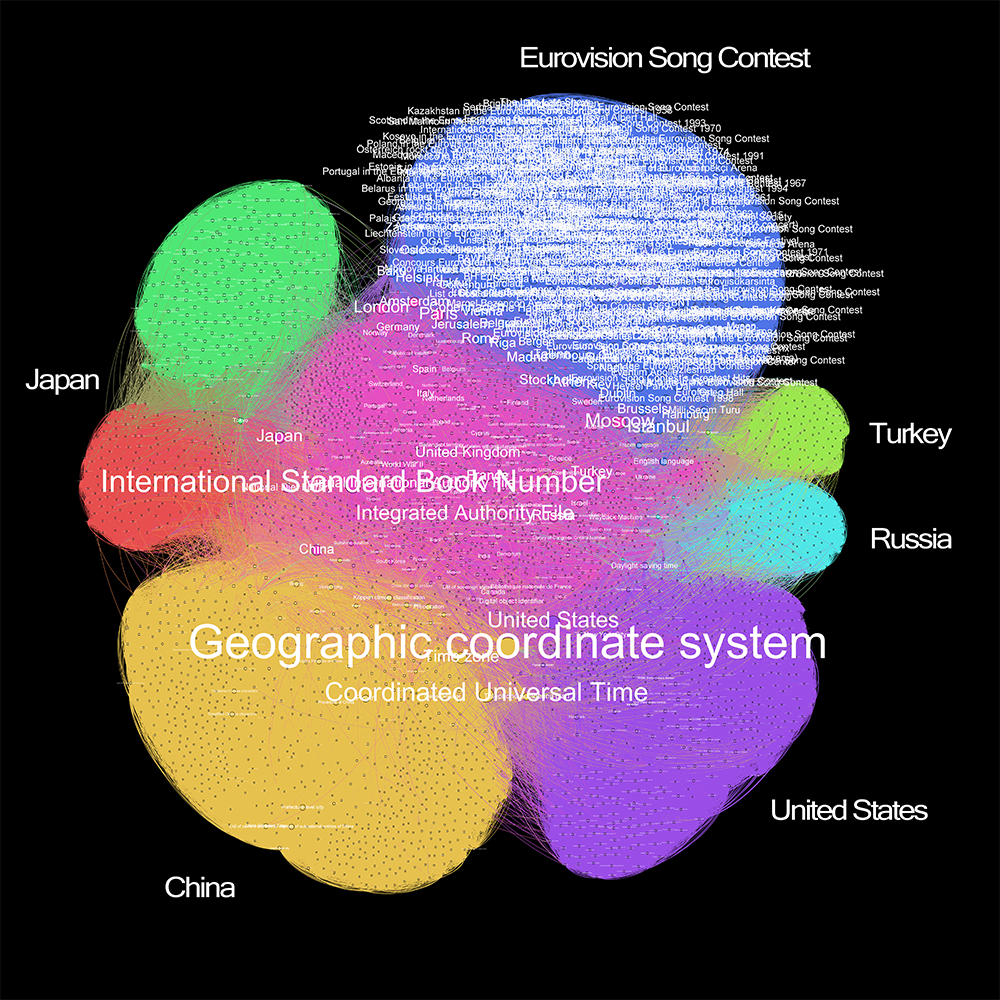
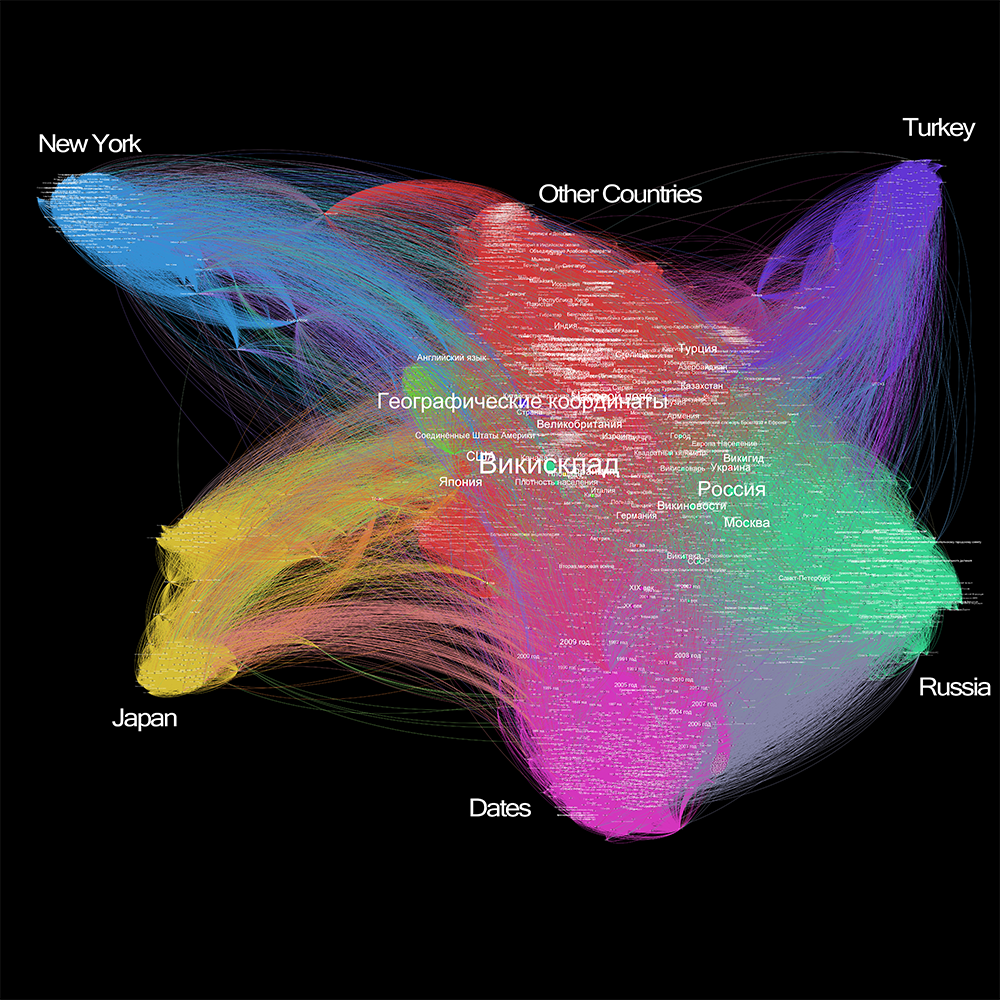




**Upper right: Image 5: Japanese**

**Lower right: Image 4: Russian**

**Lower left: Image 3: English**



**Conclusions and discussion**:

We started this research in the search of a similarity between networks for different languages. Using five different languages that represent five cultures and countries. And as we acknowledge that all languages are spoken across the world it is still reasonable to assume that the majority of speakers for each language are located in the country of origin. The English language is probably an exception as there a few English speaking countries and many people speak and understand it within the entire world and therefore is reached to for missing or esoteric articles throughout.

Here are our main conclusions from this analysis:

Interestingly, despite starting from cities as the centers of the traversal, we ended up with communities of countries except for Shanghai in the Chinese network. This is quite surprising as we expected that with globalization the nation will be blurred and the center of gravity will result in cities as the important players.

Also the English network is an outlier among the networks and it appears more balanced than other networks. A possible explanation could be that it is much larger than the other networks and because English is a common second language across the world.

All the networks (besides English) are centered on their respective countries, but Chinese is very heavily focused on China, compared to the other networks. This may be because of government censorship, or because the Chinese culture is more regionally focused.

The reason we see the Eurovision community and Dates community in a couple of the networks is because they are very dense parts of the network, with many pages linking to each other. For the same reason we see a community for New York *State* show up in Russian and Japanese. However, it is surprising that the Eurovision, which is a European regional event, is more connected than the Wikipedia articles for Olympics or other sports worldwide events.

Future research may both narrow down the research and examine the deep relationship inside one language. Also, we would like to conduct a local research and examine a big non English network such as Spanish or Arabic to see if we could identify the political power balance through the centrality and community identification.

**Resources**:

Belloni, F , Bonato, R, *Network Analysis for Wikipedia*, University of Verona, 2011

Ulrike Pfeil, Panayiotis Zaphiris,and Chee Siang Ang , *Cultural Differences in Collaborative Authoring of Wikipedia*, Volume 12, Issue 1, pages 88–113, October 2006

Schenk,Frithjof Benjamin, *Mental Maps: The Cognitive Mapping of the Continent as an Object of Research of European History*, 2008

William Emigh, Susan C. Herring , *Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias* , Indiana University Bloomington, 2005

# 