Biographical Social Networks on Wikipedia

A cross-cultural study of links that made history

Pablo Aragon Andreas Kaltenbrunner
David Laniado Yana Volkovich
Barcelona Media Foundation, Barcelona, Spain
{name.surname}@barcelonamedia.org

ABSTRACT

It is arguable whether history is made by great men and women or vice versa, but undoubtably social connections shape history. Analysing Wikipedia, a global collective memory place, we aim to understand how social links are recorded across cultures. Starting with the set of biographies in the English Wikipedia we focus on the networks of links between these biographical articles on the 15 largest language Wikipedias. We detect the most central characters in these networks and point out culture-related peculiarities. Furthermore, we reveal remarkable similarities between distinct groups of language Wikipedias and highlight the shared knowledge about connections between persons across cultures.

Categories and Subject Descriptors

J.4 [Computer Applications]: Social and Behavioural Sciences—Sociology; G.2.2 [Mathematics of Computing]: Graph Theory—Network problems

Keywords

Wikipedia, social network analysis, cross language studies

1. INTRODUCTION

Social network analysis, one of the most studied subjects in the last decade, has been applied in very different contexts ranging from online social networks [11], over networks of fictitious comic characters [1] to animal social networks [5]. Here we present a study about connections of a different form. We use neither self-reported nor observed relations nor interactions inferred from activity logs. We focus on the links between notable humans as they are recorded in collective memory. To extract these connections and build the corresponding networks we use different language versions of Wikipedia, which can be seen as global memory place [12].

We exploit direct links between biographic articles as evidence of relations between the corresponding persons, and build biographical social networks for the 15 largest language versions of Wikipedia. We investigate these networks separately and analyse their similarity. We furthermore extract the most important persons according to several centrality metrics in these networks. This allows us to analyse and compare the different language communities over their perception and reporting of connections between notable persons. The visualisation of the shared links present in most of the different language networks highlights the connections commonly known across language and culture barriers.

2. RELATED WORK

Social networks analysis on Wikipedia has mainly exploited editor interactions, either via generating co-authorship networks [8] or analysing social interactions on article and user talk-pages [9]. Additionally, co-authorship has been used to create networks of similar articles [4].

The link structure of Wikipedia articles has been studied extensively: common features have been found in the network topology of several language versions of Wikipedia [14], and rankings of the most central entries in the English Wikipedia have been presented [3]. The idea of restricting the network to articles representing entities of a given type has been followed in [2], introducing a framework for visualising links between philosophers in the English Wikipedia.

Different language versions of Wikipedia have been compared to study cultural differences among their communities [13]. While this has been done mostly by analysing the behaviour of the editors, here we propose to study differences and similarities as they emerge from the link structure of the artifacts created by different communities.

3. DATA EXTRACTION

To obtain a list of articles about persons on the English Wikipedia, we relied on a dataset from DBpedia1. The dataset contains links to 296,511 existing English articles, which we parsed to identify the names of the corresponding articles in the other 14 language versions of Wikipedia with the largest number of articles at the moment we extracted the data (September 8 to 13, 2011).

For each language version we generated a directed network where nodes represent persons with a biographical Wikipedia article and a node i links to node j if the article of the person i is linking to the article about person j.

As many Wikipedia articles have alternative names which redirect to the same article, we had to track these redirects

1http://downloads.dbpedia.org/3.7/en/persondata_en.nt.bz2
for every person and every language with a script provided by the Wikimedia Toolserver. The number of redirects per page follows a heavy tailed distribution as can be observed in Figure 1. The table embedded in the figure lists the 13 persons with the largest number of redirects in the English Wikipedia. The article about Muammar al-Gaddafi leads the ranking with 251 different ways of linking towards it, more than doubling the redirects of Osama bin Laden, the second ranked person.

4. RESULTS

In this section we study global metrics calculated for the biographical networks of different languages. We also discuss rankings based on various definitions of centrality. In particular, we present the most central (linked) persons in the different language Wikipedia. Finally, we compare the similarities between the different language networks.

4.1 Global network statistics

A brief overview of the principal social network measures for the different language networks is given in Table 1. The largest network corresponds to the English Wikipedia with nearly 200,000 nodes. The second largest, extracted from the German Wikipedia, is only about one-third as large. All language networks show very low clustering. The only outstanding network is the Chinese with a clustering coefficient of 0.17, which indicates an important structural difference of the link structure in this language version.

By looking at the link reciprocities we find that it is quite rare that two persons are mutually connected. One of the possible causes of this observation may be the presence of parasocial interactions [6], i.e. one-sided interpersonal relationships in which one part knows a great deal about the other, but the other does not. E.g., when a person is influenced by the works of somebody who died decades before.

We see that all networks are well connected, as the percentage of nodes in the giant component (GC) is nearly 200% for every person and every language with a script provided by the Wikimedia Toolserver.

Table 1: Properties of the language networks ordered by network size: number of (not isolated) nodes $N$ and edges $K$, average clustering coefficient $(C)$, percentage of nodes in the giant component GC, average path-length between nodes $(\langle d \rangle)$, reciprocity $r$ and maximal distance $d_{max}$ between two nodes in the network.

| lang | N    | K    | $(C)$ | % GC | $\langle d \rangle$ | r   | $d_{max}$ |
|------|------|------|-------|------|-------------------|-----|----------|
| en   | 198  | 928  | 0.03  | 95%  | 6.53              | 0.17| 43       |
| de   | 62   | 402  | 0.05  | 94%  | 6.83              | 0.14| 33       |
| fr   | 51   | 283  | 0.06  | 96%  | 6.11              | 0.15| 36       |
| it   | 35   | 256  | 0.06  | 95%  | 6.28              | 0.14| 42       |
| es   | 34   | 238  | 0.08  | 96%  | 6.29              | 0.16| 36       |
| ja   | 26   | 199  | 0.08  | 96%  | 6.47              | 0.20| 26       |
| nl   | 24   | 172  | 0.08  | 94%  | 7.91              | 0.18| 37       |
| pt   | 23   | 105  | 0.07  | 94%  | 6.98              | 0.18| 45       |
| sv   | 23   | 75   | 0.07  | 91%  | 8.27              | 0.20| 46       |
| pl   | 22   | 403  | 0.08  | 85%  | 8.94              | 0.16| 43       |
| fi   | 18   | 154  | 0.07  | 87%  | 7.80              | 0.17| 30       |
| no   | 18   | 144  | 0.09  | 89%  | 8.31              | 0.22| 48       |
| ru   | 16   | 348  | 0.06  | 86%  | 1.01              | 0.10| 35       |
| zh   | 11   | 471  | 0.17  | 91%  | 7.20              | 0.20| 32       |
| ca   | 11   | 432  | 0.09  | 93%  | 7.14              | 0.17| 32       |

Finally, we also analyse the in- and out-degree distributions and observe heavy-tails, as found in many real-life networks, for all language Wikipedias (data not shown).

4.2 Most central persons

In this section we focus on centrality metrics for the above defined biographical networks. In Table 2 we present the top-ranked persons according to the degree centrality for the English Wikipedia. We also provide results for other centrality measures together with the corresponding rankings. Betweenness measures the fraction of shortest paths between other pairs of nodes passing through a given node, while PageRank gives a measure of the global importance of nodes, computed recursively putting a larger weight on incoming connections from central nodes.

Table 2: The top 25 persons in the English Wikipedia ranked by in-degree. Ranks for out-degree, betweenness and PageRank in parenthesis.

| person               | in-degree | out-degree | btw.  | PageRank |
|----------------------|-----------|------------|------|----------|
| George W. Bush       | 2123      | 89 (107)   | (1)  | 0.00209  |
| Barack Obama          | 1677      | 51 (710)   | (8)  | 0.00162  |
| Bill Clinton          | 1660      | 70 (205)   | (4)  | 0.00156  |
| Ronald Reagan         | 1625      | 90 (103)   | (2)  | 0.00156  |
| Adolf Hitler          | 1497      | 119 (26)   | (5)  | 0.00149  |
| Richard Nixon         | 1299      | 86 (127)   | (7)  | 0.00136  |
| William Shakespeare   | 1229      | 25 (4203)  | (63) | 0.00113  |
| John F. Kennedy       | 1208      | 104 (53)   | (5)  | 0.00123  |
| Franklin D. Roosevelt | 1052      | 71 (237)   | (15) | 0.00131  |
| Lyndon B. Johnson     | 1000      | 106 (50)   | (12) | 0.00108  |
| Jimmy Carter          | 953       | 80 (158)   | (9)  | 0.00113  |
| Elvis Presley         | 948       | 82 (142)   | (27) | 0.00063  |
| Pope John Paul II     | 941       | 59 (444)   | (13) | 0.00083  |
| Dwight D. Eisenhower  | 891       | 55 (564)   | (22) | 0.00095  |
| Frank Sinatra         | 882       | 108 (47)   | (18) | 0.00056  |
| George H. W. Bush     | 878       | 87 (118)   | (19) | 0.00096  |
| Abraham Lincoln       | 846       | 54 (593)   | (40) | 0.00089  |
| Bob Dylan             | 835       | 151 (11)   | (14) | 0.00055  |
| Winston Churchill     | 748       | 84 (136)   | (10) | 0.00092  |
| Harry S. Truman       | 743       | 81 (145)   | (24) | 0.00099  |
| Joseph Stalin         | 725       | 69 (265)   | (43) | 0.00089  |
| Michael Jackson       | 663       | 71 (237)   | (34) | 0.00042  |
| Elizabeth II          | 653       | 52 (665)   | (6)  | 0.00074  |
| Jesus                 | 572       | 38 (1595)  | (51) | 0.00068  |
| Hillary Rodham Clinton| 554       | 87 (118)   | (32) | 0.00063  |
We find many American presidents, iconic American musicians, and European leaders during the WW2 period among the most linked. Interestingly, we observe that Pope John Paul II appears to be a more central figure than Jesus. Comparing the number of incoming and outgoing links, we observe that in-degrees are of an order of magnitude greater than out-degrees. We explain this phenomenon again by the presence of the parasocial relations. For betweenness and PageRank we do not find large differences in the rankings. The only exception is Shakespeare, whose low betweenness value can be caused by the low number of out-going links. Interestingly, Shakespeare’s page is one of the most central for several languages (see Table 3), but not for English.

In Table 3 we show the most central characters in Wikipedia for the 15 analysed languages ranked by the betweenness centrality. We observe that most of the presented persons are known to be (or have been) highly influential in many aspects. Thus, in these lists we find political leaders, revolutionaries, famous musicians, writers and actors. We note that political figures such as Adolf Hitler, George W. Bush or Barack Obama dominate in almost all top rankings. Interestingly, William Shakespeare and Michael Jackson are also among the central figures for several languages.

For many languages we find, however, that the top ranked persons reflect country specific issues. Thus, for example, Pope John Paul II is only present in the top five list of the Italian and Polish Wikipedia, two countries which have a special tie with this figure. In the English Wikipedia the most central figures are former US presidents, while the Spanish-speaking Wikipedia community marks out Latin American revolutionaries. The Russian version surprisingly highlights William Shakespeare and also Kenneth Branagh, known for several film adaptations of Shakespeare’s plays. Only the Japanese Wikipedia ranks the author Mishima prominently, while two Chinese leaders, Chiang Kai-Shek and Deng Xiaoping, are in the top-5 in the Chinese version.

When looking at these results, it should be taken into account that there is an Anglo-Saxon bias in the dataset, as we relied on a list of notable persons extracted from the English Wikipedia, and persons from other cultures not know internationally might be missing. In that sense the above list reflects centrality among persons with at least limited international notoriety.

### 4.3 Similarity between languages

In this section we focus on similarities between the networks emerging from the different language Wikipedias. We calculate the similarity for every pair of networks as their Jaccard coefficient, i.e. the ratio between the number of links present in both networks (their intersection) and the number of links existing in their union. Table 4 shows the obtained similarity results. For every language row we highlight the two languages with the largest similarities. The most similar language is also underlined.

Figure 2 further illustrates these similarities by drawing a language similarity network. In this network a language A is connected to another language B if language B is one of the two most similar to A. Applying the Louvain method, we divide this graph into three clusters. In agreement with [10], we observe that most of the links can be explained by language-family relations (e.g. Romance and Slavic languages) and geographic or historical ties (e.g. Scandinavian group, or Russia and Finland). We also find a number of less obvious connections, e.g. Japanese to Spanish and Finnish
to Dutch. In fact, Dutch seems to serve as a bridge between different language and culture groups.

4.4 Links present in most language networks

Finally, Figure 3 depicts the network of connections which are present in at least 13 of the 15 language Wikipedias. The network is relatively small, containing 1663 nodes and 1738 links, but allows to visualise the quintessence of knowledge about biographical connections present in most of the analysed language communities.

The largest connected component in Figure 3 corresponds to a cluster of US Presidents which connects over Ronald Reagan to a cluster of British Premier Ministers. This group is related through Winston Churchill to a cluster of persons from WW2’s axis powers. The second largest component is composed of several clusters related to the music and entertainment business, and the third one of two clusters of male and female tennis players connected through Dinara and Marat Safin. Other large isolated clusters can be found around such diverse groups as Russian and Chinese political figures, French presidents, Israeli and Palestinian politicians, Formula One pilots, World Chess Champions or actresses.

5. CONCLUSIONS

Our results show that biographical connections are recorded differently in the distinct language versions of Wikipedia. Although the global social network measures are largely similar for all these networks, the most central persons unveil interesting peculiarities about the language communities. A study of similarity reveals that networks are more similar for geographically or linguistically closer communities. Nevertheless, there also exists a great number of biographical connections which can be found in most of the analysed language Wikipedias. Knowledge about these social connections trespasses cultural barriers and represents part of the shared global collective memory of our civilisation.

Possible directions for future work include the application of the methodology to generate subnetworks of other kinds of article categories, and a specific study of links present in most networks but missing only in a few language Wikipedias.

Finally, the gender gap among Wikipedia editors is a serious concern for the community, and has been related to the topics covered in the encyclopedia [7]. Our results point out a very small presence of females also among the most central persons in the encyclopedic content, suggesting the link between these two phenomena as an intriguing subject for future investigation.

Acknowledgements

This work was partially supported by the Spanish Centre for the Development of Industrial Technology under the CENTI program, project CEN-20101037, “Social Media”. Yana Volkovich acknowledges support from the Torres Quevedo Program from the Spanish Ministry of Science and Innovation, co-funded by the European Social Fund.

6. REFERENCES

[1] R. Alberich, J. Miro-Julia, and F. Rossello. Marvel Universe looks almost like a real social network. cond-mat/0202174, 2002.
[2] S. Athenikos and X. Lin. The WikiPhil Portal: visualizing meaningful philosophical connections. J. of the Chicago Colloq. on Digital Humanities and Comp. Sci., 1(1), 2009.
[3] F. Bellomi and R. Bonato. Network analysis for Wikipedia. In Proc. of WikiSym, 2005.
[4] R. Biuk-Aghai. Visualizing co-authorship networks in online Wikipedia. In Proc. of ISCIT’06, 2006.
[5] P. I. Chiyo, C. J. Moss, and S. C. Alberts. The influence of life history milestones and association networks on crop-raiding behavior in male african elephants. PLoS ONE, 7(2):e31382, 2012.
[6] D. Horton and R. Wohl. Mass communication and para-social interaction: Observations on intimacy at a distance. Psychiatry, 19(3):215–229, 1956.
[7] D. Laniado, C. Castillo, A. Kaltenbrunner and M. Fuster-Morell. Emotions and dialogue in a large peer-production community: the case of Wikipedia. In Proc. of WikiSym, 2012.
[8] D. Laniado and R. Tasso. Co-authorship 2.0: Patterns of collaboration in Wikipedia. In Proc. of Hypertext, 2011.
[9] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In Proc. of ICWSM, 2011.
[10] H. Liao and T. Petzold. Analysing geo-linguistic dynamics of the world wide web: The use of cartograms and network analysis to understand linguistic development in Wikipedia. Cultural Science, 3(2), 2011.
[11] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In Proc. of IMC, 2007.
[12] C. Pentzold. Fixing the floating gap: The online encyclopedia Wikipedia as a global memory place. Memory Studies, 2(2):255–272, 2009.
[13] U. Pfeil, P. Zaphiris, and C. Ang. Cultural differences in collaborative authoring of Wikipedia. JCMC, 12(1):88–113, 2006.
[14] V. Zlatić, M. Božićević, H. Štefančić, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. Physical Review E, 74(1):016115, 2006.

Figure 3: Biographical network of the connections in at least 13 of the 15 analysed language versions. Larger version available at http://goo.gl/Lc2Xl.