BRILL

RESEARCH DATA JOURNAL FOR THE HUMANITIES
AND SOCIAL SCIENCES 5 (2020) 50-65

BRILL OPEN
brill.com/rdj

Networked Pantheon: a Relational Database of Globally Famous People

Social and Behavioural Sciences

Pablo Beytía
Department of Social Sciences, Humboldt University of Berlin, Berlin, Germany
beytiapa@hu-berlin.de

Janosch Schobin
Department of Sociology, University of Kassel, Kassel, Germany
jschobin@uni-kassel.de

Abstract

This article presents the *Networked Pantheon*, a relational database of biographies of globally famous people spanning the last 5,500 years of human history. This information source is intended to complement Pantheon 1.0 (Yu et al., 2016), a dataset that includes temporal, spatial, gender, and occupational information on 11,341 world-renowned people – defined as those who have biographies available in more than 25 languages on Wikipedia. The *Networked Pantheon* adds information about the biographical links between these historical figures, compiled from hyperlinks between the biographies in the English Wikipedia. This digital method enables techniques from network analysis to be used in studying the biographical relationships between globally famous people. Thus, distinct measures of historical centrality can be calculated for individuals, cities, countries, genders, and occupations. The *Networked Pantheon* includes indicators of figure centrality in the network of biographical references and provides an approximation of the information flows between various territories, genders, and occupations of famous people over time.

Keywords

human history – Wikipedia – biographies – collective memory – computational social science – digital humanities – digital sociology – network analysis

– Related data set "Networked Pantheon: a Relational Database of Globally Famous People" with DOI www.doi.org/10.17605/OSF.IO/QTU2J in repository "Open Science Framework"

1. Introduction

In recent years, there has been growing scientific interest in the organization of the knowledge stored in Wikipedia. On the one hand, the information structured in this digital encyclopedia has been used as input to study a large number of phenomena, such as historical trends (Jara-Figueroa et al., 2016; Menini et al., 2017; Reznik & Shatalov, 2016; Schich et al., 2014), links between languages (Aragon et al., 2012; Ban et al., 2017; Eom et al., 2015; Mehler et al., 2011; Ronen et al., 2014), geopolitical instabilities (Apic et al., 2011), global prestige of universities (Lages et al., 2016), underlying connections between proteins (Zinovyev et al., 2020) and the influence of infectious diseases (Rollin et al., 2019). On the other hand, it has also been used to analyze the "systemic biases" of Wikipedia's collective repository, emphasizing differences in the degree of information on various territories (Beytía, 2020; Graham et al., 2015; Roll et al., 2016), cultures (Eom et al., 2015; Nemoto & Gloor, 2011; Overell & Rüger, 2011) and genders (Gruwell, 2015; Shane-Simpson & Gillespie-Lynch, 2017).

Following a series of pioneer studies in the area (Michel et al., 2011; Murray, 2003; Popescu & Grefenstette, 2010; Schich et al., 2014; Skiena & Ward, 2013), a remarkable database for conducting these types of research was published in 2016. This is Pantheon 1.0, a dataset of globally famous people that includes information about the 11,341 biographies present in more than 25 language versions of Wikipedia (Yu et al., 2016). Using the number of languages in which each biography is available as a proxy for its global cultural relevance, this dataset gathers indispensable information that historically locates recognized personalities (such as the year, city, and geographic coordinates of birth), data on the personal characteristics of each individual (e.g., gender and main occupation), and indicators of their historical popularity. In this way, Pantheon 1.0 enables the multidimensional study of the organization of the world's biographical knowledge in Wikipedia, facilitating the exploration of the temporal, spatial, gender, and occupational aspects over an enormous timeframe (3,500 BC – to date).

This article presents the *Networked Pantheon*, a database designed to complement Pantheon 1.0 with relational observations. It provides essentially three

new types of information. First of all, data on the biographical links between the notable people included in Pantheon dataset, approaching these relationships from the hyperlinks between their biographies in English Wikipedia.[1] Second, network metrics for each biography, which can be used to better understand the structures of the information about prominent people in Wikipedia. Finally, the year of death of each historical figure, which enables each registered individual to be associated with a clearly delimited life period.

The *Networked Pantheon* has been used to study how hyperlinks between biographies enhance the dissemination of content about people born in some countries, what increases the geographical bias of Wikipedia's biographical record (Beytía, 2020). But it could be used to answer several questions, such as:

- What links can be identified in the lives of world-renowned people?
- How are these links structured into networks of biographical references across space and time?
- Which historical figures are most central in Wikipedia's global network of biographical references?
- Which groups with high biographical interconnection can be identified?
- How independent or closed are the occupational networks in different periods and territories?
- To what extent have women been excluded from certain professions over time?
- How much have famous people from different occupations or territories tended to relate biographically to similar people?
- Which cities historically have stronger links in terms of the biographical connections of their scientists, artists or politicians?

The Networked Pantheon database is freely available on the Open Science Framework (OSF) server. It can be downloaded from the project's home page (www.osf.io/qtu2j/) or directly from the section that stores the files (www.osf.io/qtu2j/files/). It is registered under a Creative Commons "Attribution 4.0 International" license,[2] which implies that it can be freely shared and adapted, giving credit to its authors and indicating whether changes to the original version were made.

1 This method has been used in previous research, such as Aragon et al. (2012), Beytía & Müller (2019), and Skiena & Ward (2013).

2 The detailed explanation of this license is available at www.creativecommons.org/licenses/by/4.0/legalcode.

2. Methods

2.1. *Link Data*

The hyperlinks between the Wikipedia biographies of 11,340 historically famous individuals contained in the original Pantheon Database (Yu et al., 2016)[3] were extracted using the R software packages *rvest* (Wickham, 2016) and *stringi* (Gagolewski, 2020). These links were obtained from the English Wikipedia articles (extraction date 16.04.2018).[4] To this effect, first, we converted the html document of the Wikipedia article into an xml-tree. From this tree, all nodes of the class "p a" were selected. The class selector indicates that a node corresponds to a hyperlink, thus including all hyperlinks present in the Wikipedia article. In a second step, we checked which links linked to the Wikipedia article of another famous person in the Pantheon Database, excluding all links that did not.[5]

The English language version was chosen because it is the most complete version – i.e., that with the most articles, biographies, editions, and editors (Aragon et al., 2012; Nemoto & Gloor, 2011) – and the one that registers the largest number of historical figures with biographies in 25 or more different languages. English Wikipedia includes biographies for all but one of the famous individuals recorded in Pantheon 1.0 (11,340 people in total),[6] followed by the French (11,334), German (11,319), Russian (11,314), and Spanish (11,287) versions. While language selection might imply a better record of biographical links of English-speaking people, it has been documented that many of the hyperlinks between Wikipedia biographies overcome language barriers (Aragon et al., 2012) and this phenomenon should be more common among

3 Available at https://dataverse.harvard.edu/dataverse/pantheon. A new version, including people in more than 15 languages, is currently available at https://pantheon.world.

4 It has to be noted that updates of the Wikipedia pages might add or remove links from the Database in subsequent updates.

5 Initially, we compared this approach to (1) parsing the content of the Wikipedia-page into plain text and searching directly for the names of the famous individuals in the Pantheon Database using fuzzy string matching, and (2) to using more selective selectors (such as first filtering for "#content" and then selecting all nodes classed "p a"). Both yielded poorer results in heuristic validity checks that we performed on a small set of biographical Wikipedia articles from different historical domains and with different components and styles (for instance, names might be written in very different ways, e.g., in their Latin form; articles could include important biographical connections in biographical cards, picture descriptions or navigation boxes, and so on).

6 The only biography included in Pantheon, but not available in the English Wikipedia, is that of the Italian photographer Augusto de Luca.

biographies of globally known characters, such as those recorded in many different languages.

2.2. *Year of Death*

We extracted the year of death of the famous personalities from the html code of their Wikipedia pages by looking for the string "died" in relation to various date formats. Here, we selected the biography boxes from the xml-tree (selector ".vcard") to search for the year of death (extraction date 13.03.2018). The extracted dates were checked for plausibility against the birthdates already present in the original Pantheon database. Subsequently, we checked 436 cases manually that were implausible and corrected them (extraction between 13.03.2018 and 09.04.2018). Some biographies did not have a date of death, mostly because they describe very old historical figures who lack precise historical records. In a total of 136 cases, the dates of death had to be imputed. These imputations were calculated from the median lifespan associated with the historical period of each figure, which was approximated from the lifespan of the 10 closest cases according to the year of birth.

2.3. *Network Measures*

For each biography registered in the *Networked Pantheon* database, a series of structural measures were calculated from the network of biographical connections. These indicators are as follows:

- *Degree*: number of connections or edges that one node (or biography) has to other nodes (Freeman, 1978–1979).
- *Indegree*: number of edges (hyperlinks) going into a node.
- *Outdegree*: number of edges coming out of a node.
- *Betweenness*: the frequency with which a node appears in the shortest path between the nodes of the network (Brandes, 2001; Freeman, 1978–1979).
- *Eigen Centrality* (*eigenvector*): the centrality of an actor in proportion to the sum of the centralities of its neighbors in the graph (Bonacich, 1987).
- *PageRank*: the measure of the global importance of nodes, computed recursively by placing greater weight on incoming connections from central nodes (Brin & Page, 1998; Page et al., 1999).
- *Eccentricity*: the distance between a node and that furthest away from it in the network (Hage & Harary, 1995).
- *Closeness Centrality*: the distance between a node and all other nodes in the network, based on the arithmetic mean of the minimum path between the nodes (Freeman, 1978–1979).
- *Harmonic Closeness Centrality*: the distance between a node and all other nodes in the network, based on the harmonic mean of the minimum path between the nodes (Rochat, 2009).

– *Authority*: a good authority is a webpage (biography) that is pointed to by many good hubs (Kleinberg, 1998).
– *Hub*: a good hub is a webpage (biography) that points to many good authorities (Kleinberg, 1998).
– *Clustering*: the degree to which the nodes tend to cluster together (Saramäki et al., 2007).

2.4. *Biographical Centrality Index* (*BCI*)

As a complement to the Historical Popularity Index (Yu et al., 2016), the *Networked Pantheon* includes a *Biographical Centrality Index* (*BCI*) for each historical figure that denotes their cultural ubiquity (approximated by the number of languages in which a biography is available) weighted by its biographical connectivity (approximated by the PageRank algorithm). Considering the number of language versions of a biography (NL) and its PageRank (PR), the *non-normalized BCI* is the multiplication of both values ($NL \times PR$). This indicator, however, was later normalized through Feature Scaling method. Once we identified, in the complete group of biographies, the minimum – $\min(NL \times PR)$ – and maximum – $max(NL \times PR)$ – values of the non-normalized BCI, normalization was carried out using this formula:

$$BCI = \frac{(NL \times PR) - min(NL \times PR)}{max(NL \times PR) - min(NL \times PR)}$$

BCI can be understood as a normalized indicator of the probability that a historical character would appear linked to a random biographical search in a random language in Wikipedia. It refers to a figure's degree of multilingual exposure and connectivity. The indicator is relative to the distribution of the cultural ubiquity and biographical connectivity of the total number of individuals considered, and it can be interpreted as the centrality of a biography compared to the most central one of the sample (as a value between 0 and 1, where the latter represents the greatest possible centrality).

The BCI should be clearly distinguished from the Historical Popularity Index (HPI) included in Pantheon 1.0 (Yu et al., 2016) for at least three reasons:

1. The BCI considers *biographical connectivity* as a relevant indicator for ranking the influence of characters on the discursive structure of Wikipedia.
2. It is an indicator focused on the organization of the *content produced* (supply of information), without considering the request or demand for biographical information ("page views" variable).
3. The BCI aims to be a tool to understand how historical memory is currently being structured in Wikipedia, highlighting disparities in the distribution of historical information and concentrations of hyperlink

"flows" in certain periods, territories, genders and occupations. Therefore, it is not an adequate indicator for approaching the historical importance of each character – which should consider, for example, an adjustment for the excessive relative importance of some 20th-century characters.

3. Database Description

- **Networked Pantheon deposited at Open Science Framework** – DOI:www.doi.org/10.17605/OSF.IO/QTU2J
- **Video supplement to Beytía & Schobin 2020**
  - "Networked Pantheon database" deposited at figshare – DOI:www.doi.org/10.6084/m9.figshare.12871727
- **Temporal coverage:** 5500 BC-2018 AD

3.1. *General Aspects*

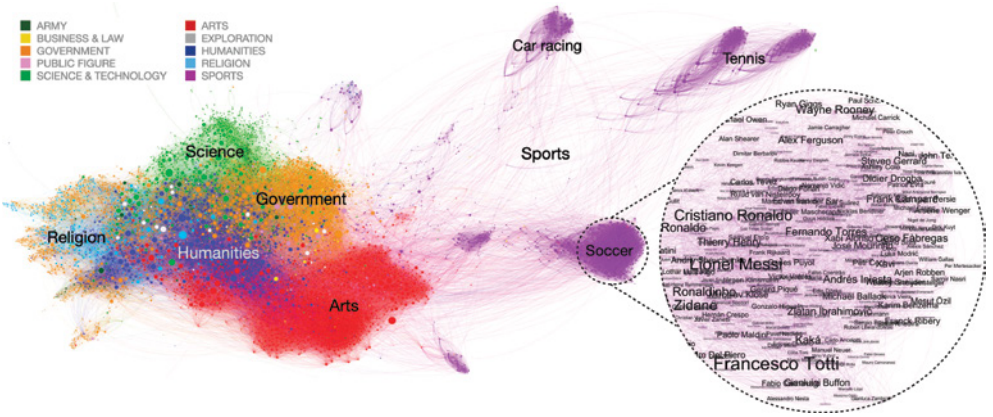
The *Networked Pantheon* records 126,279 direct relationships between 11,340 digital biographies of globally famous people, defined as those who have Wikipedia biographies in more than 25 languages. On average, each biography has 11.12 hyperlinks to others, ranging from 0 (e.g., Al Capone) to 105 (Meryl Streep). The number of incoming hyperlinks to these biographies varies between 0 (e.g., Josep Guardiola) and 414 (Barack Obama). The diameter of the network – that is, the shortest distance between the two most distant nodes in the network – is 15, while the average path length is 4.81.

Figure 1 illustrates the general topography of the network using colors to distinguish large occupational domains classified in the Pantheon dataset (army, business and law, government, public figure, science and technology, arts, exploration, humanities, religion, and sports). As an example of the composition of the network in specialized fields, a cluster of globally recognized soccer players is shown in greater detail.

3.2. *Dynamic Analysis*

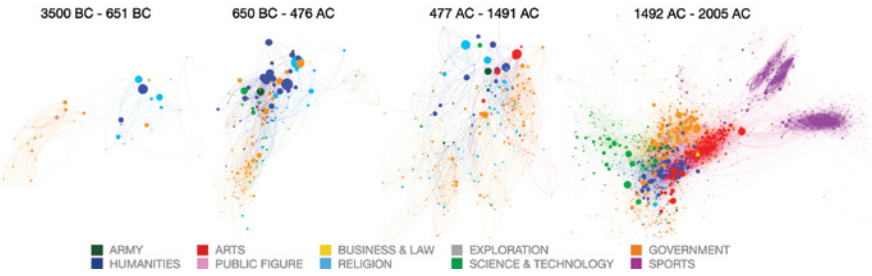
This database allows the analysis of networks of biographical references among contemporaries, and thus the comparison of linkage structures in specific historical periods.

Figure 2 shows an example of the structural variation of the network over time, distinguishing a period in which the archaic civilizations emerged (3500 BC – 651 BC), another associated with the rise of the Greco-Roman culture (651 BC – 476 AC), a third epoch that is commonly referred to as the Middle Ages



Note: The *Networked Pantheon* distributed using the ForceAtlas2 algorithm (Jacomy et al., 2014) and differentiating the occupational domains by color. The size of the nodes and soccer players' names is proportional to the number of languages in which each biography is available.

FIGURE 1 Network structure



Note: The colors represent general occupational domains. The size of the nodes is proportional to the number of languages in which each biography is available.

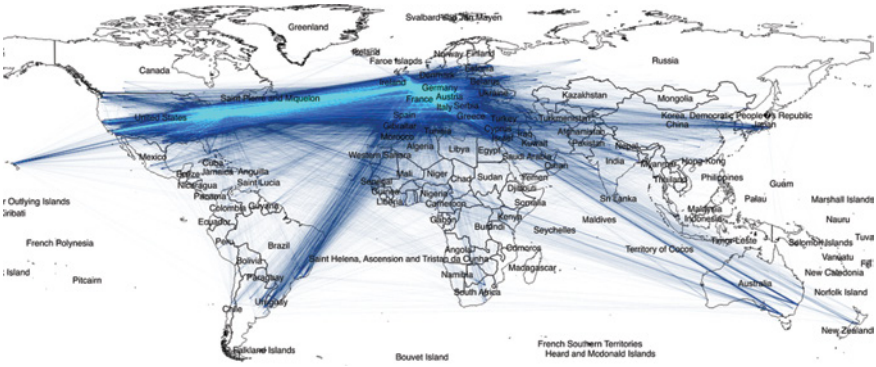
FIGURE 2 Example of temporal dynamics

(477 AC – 1491 AC) and the period after the spread of the printing press and the beginning of globalization (1492 AC – 2005 AC).

Moreover, this figure illustrates how the selected periods constitute different networks in terms of shape, size, density and occupational composition. Since data can be separated by year of birth, place of birth, gender and occupational domain, more specialized studies are also feasible.

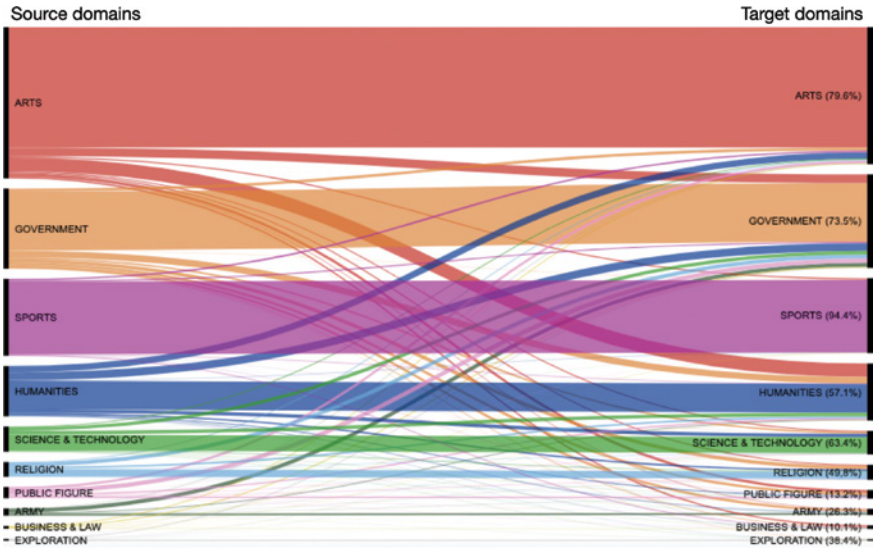
3.3. *Flow of Biographical References*

With this database, the biographical relationships between historical figures can be geographically located; thus, it is possible to approach the flow of



Note: The position of each node is defined by the place of birth.

FIGURE 3 Geographical distribution of the flow of biographical references



Note: The left side shows the occupational domains, ordered by the number of biographical references (hyperlinks) that they generate and the right side represents the same domains ordered according to the number of references that they receive. The percentage of links coming from the same domain is specified in parentheses on the right side.

FIGURE 4 Flows of biographical references between occupational fields

informational links between various regions of the world. Figure 3 displays the worldwide distribution of biographical references, showing a high concentration of links between Western Europe and the United States.

These reference flows can be also analyzed according to different classifications – such as the occupation or gender of the historical figures. For example,

Figure 4 represents the (origin and destination) flows of references between occupational domains. As can be seen, a large number of hyperlinks received in each domain come from the same domain. That degree of "self-referencing" is specified on the right side of the graph, which points out the percentage of received links that each occupational domain receives from itself. It could be inferred from this indicator, for example, that sports, arts and government are the most self-referential or autonomous occupational domains of the network.

3.4. *Structural Measures of Centrality and Influence*

Measures of network centrality observe various relational phenomena, so they can be used to order historical characters in different ways. Based on the hyperlinks between biographies in the English Wikipedia, Table 1 shows a comparison between the top 10 historical figures following selected coefficients. In the table, the Eigenvector highlights the biographical centrality of U.S. presidents, while the Authority coefficient prioritizes ATP tennis players; the PageRank highlights the role of 20th-century politicians and classical humanists, while the Betweenness widens the range of influential occupations to religious leaders, former politicians, scientists, sportsmen, and businessmen. The Closeness coefficient is the least biased towards Western culture, and it includes sultans, Nobel Laureates in Physics, athletes, and politicians.

3.5. *Biographical Centrality Index (BCI)*

The BCI is an indicator of the positioning of biographical information in Wikipedia, and it can be used for spatial and temporal analysis. Figure 5 shows the geographical distribution of the accumulated biographical centrality in the countries. As can be seen, the biographical centrality in Wikipedia is not evenly distributed across the territory but clearly concentrated in the United States and Western Europe.

Also, the biographical positioning is concentrated on certain historical periods (Figure 6). Given that the number of biographies has grown exponentially over the last five centuries, there is also a higher accumulation of biographical centrality (BCI sum) in that period. A less intuitive pattern emerges when observing changes in the biographical centrality mean over time. Figure 6 shows that there are several periods in ancient history when famous people average high levels of centrality within the current historical record. For example, there is a centrality peak around 400 BC – when Greece was a center of science and humanities – and other around the year 0 – when many religious figures linked to Christianity were born.

TABLE 1 Comparison of centrality measures (top 10 ranking)

Rank	Betweenness	Closeness	Authority	Eigenvector	PageRank	BCI
1	Pope John Paul II	Mohammed V of Morocco	Roger Federer	Barack Obama	Adolf Hitler	Barack Obama
2	Adolf Hitler	Hassan II of Morocco	Rafael Nadal	George W. Bush	Barack Obama	Adolf Hitler
3	Charlemagne	Eric Allin Cornell	Novak Djokovic	Ronald Reagan	George Bush	William Shakespeare
4	George Bush	Carl Wieman	Andrew Murray	William Shakespeare	William Shakespeare	George W. Bush
5	Benito Mussolini	Willis Lamb	Andy Roddick	Adolf Hitler	Joseph Stalin	Joseph Stalin
6	Albert Einstein	Oleg Blokhin	Tomáš Berdych	Bill Clinton	Ronald Reagan	Aristotle
7	Vladimir Putin	Valdis Zatlers	Jo-Wilfried Tsonga	Donald Trump	Plato	Plato
8	Pelé	Igor Belanov	Nikolay Davydenko	John F. Kennedy	Bill Clinton	Karl Marx
9	Ronald Reagan	Mathieu Kérékou	David Ferrer	Roger Ebert	Winston Churchill	Isaac Newton
10	Roman Abramovich	Raymond Davis Jr.	James Blake	Richard Nixon	Aristotle	Albert Einstein

Note: A brief explanation of these measures can be found in section 2.3.

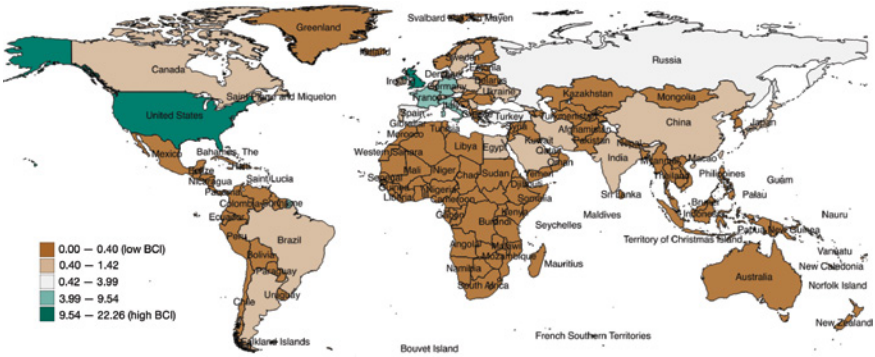
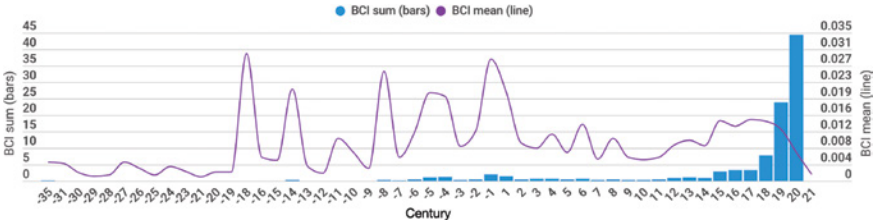


FIGURE 5 Geographical concentration of the BCI: cumulative sum by country



Note: The historical average of the BCI was calculated for the accumulation of historical characters over 50-year periods.

FIGURE 6 Historical concentration of the BCI: variation of the sum and average over centuries.

4. Concluding Remarks

The *Networked Pantheon* (www.osf.io/qtu2j/) aims to increase the huge analytical potential of Pantheon 1.0 by adding relational observations that can answer new questions about digitally constructed history, collaborative media, cultural influence, distribution of information, global collective memory, and biographical knowledge structuration, among other areas.

This database could be used for many purposes, such as:

– Studying the historical links between world-famous people, or more precisely, the collective memory of those links in Wikipedia.
– Modeling the networks of biographical references between historical figures across space and time.
– Calculating indicators of historical centrality for each famous individual, city, country, continent, or occupation.

– Identifying, in different historical periods, clusters of individuals highly interconnected by their biographical records.
– Calculating the independence or closure of occupational networks in different territories and historical periods.
– Investigating the degree of gender segregation in the biographical networks of particular occupations.
– Studying homophily – or tendency to relate between similar people – in occupations and geographical locations over time.
– Researching the scientific, artistic or political exchange (biographical flows) between cities, countries or continents.

These topics may be associated with widely established fields of research – such as digital humanities, media studies, computer science or computational linguistic – but also with the emerging computational and digital social sciences (Lazer et al., 2009), which in recent years have undergone an adequate level of methodological reflection (Rieder & Röhle, 2012; Rogers, 2013; Venturini et al., 2018) and growing institutionalization in new sub-disciplines, such as digital sociology (Lupton, 2014; Marres, 2017; Orton-Johnson & Prior, 2013), digital anthropology (Horst & Miller, 2013; Miller & Slater, 2000), and digital geography (Graham, 2014; Zook et al., 2004).

Acknowledgements

This research was supported by the German Academic Exchange Service (DAAD) and the National Agency for Research and Development (ANID) of the Chilean Government, through a doctoral scholarship awarded to Pablo Beytía.

References

Apic, G., Betts, M. J., & Russell, R. B. (2011). Content disputes in Wikipedia reflect geopolitical instability. *PLOS ONE*, *6*(6), Article e20902. www.doi.org/10.1371/journal.pone.0020902.

Aragon, P., Laniado, D., Kaltenbrunner, A., & Volkovich, Y. (2012). Biographical social networks on Wikipedia: a cross-cultural study of links that made history. *WikiSym: Proceedings of the eighth annual international symposium on Wikis and open collaboration* (Article 19, pp. 1–4). ACM. www.doi.org/10.1145/2462932.2462958.

Ban, K., Perc, M., & Levnajić, Z. (2017). Robust clustering of languages across Wikipedia growth. *Royal Society open science*, *4*(10), Article 171217. www.doi.org/10.1098/rsos.171217.

Beytía, P. (2020). The positioning matters: Estimating geographical bias in the multilingual record of biographies on Wikipedia. *WWW20: Companion proceedings of the web conference 2020* (pp. 806–810). www.doi.org/10.1145/3366424.3383569.

Beytía, P., & Müller, H. (2019). Towards a digital reflexive sociology: exploring the most globally disseminated sociologists on multilingual Wikipedia. www.doi.org/10.31235/osf.io/3pfrv.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, *25*(2), 163–177.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117.

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, *92*(5), 1170–1182.

Eom, Y. H., Aragón, P., Laniado, D., Kaltenbrunner, A., Vigna, S., & Shepelyansky, D. L. (2015). Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. *PLOS ONE*, *10*(3), Article e0114825. www.doi.org/10.1371/journal.pone.0114825.

Freeman, L. C. (1978–1979). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.

Gagolewski, M. (2020). *R package stringi: Character string processing facilities*. https://stringi.gagolewski.com.

Graham, M. (2014). Internet geographies: Data shadows and digital divisions of labour. In M. Graham & W. H. Dutton (Eds.), *Society and the Internet: How networks of information and communication are changing our lives* (pp. 99–116). Oxford University Press.

Graham, M., Straumann, R. K., & Hogan, B. (2015). Digital divisions of labor and informational magnetism: Mapping participation in Wikipedia. *Annals of the Association of American Geographers*, *105*(6), 1158–1178.

Gruwell, L. (2015). Wikipedia's politics of exclusion: Gender, epistemology, and feminist rhetorical (in) action. *Computers and Composition*, *37*, 117–131.

Hage, P., & Harary, F. (1995). Eccentricity and centrality in networks. *Social networks*, *17*(1), 57–63.

Horst, H. A., & Miller, D. (Eds.). (2013). *Digital anthropology*. A&C BlackAnthropology.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLOS ONE*, *9*(6), Article e98679. www.doi.org/10.1371/journal.pone.0098679.

Jara-Figueroa, C., Yu, A. Z., & Hidalgo, C. A. (2016). The medium is the memory: how communication technologies shape what we remember. *arXiv*:1512.05020v3. www.arxiv.org/abs/1512.05020v3.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*.

Lages, J., Patt, A., & Shepelyansky, D. L. (2016). Wikipedia ranking of world universities. *The European Physical Journal B*, *89*(3), Article 69.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. *Science*, *323*(5915), 721–723.

Lupton, D. (2014). *Digital sociology*. Routledge.

Marres, N. (2017). *Digital sociology: The reinvention of social research*. John Wiley & Sons.

Mehler, A., Pustylnikov, O., & Diewald, N. (2011). Geography of social ontologies: Testing a variant of the Sapir-Whorf Hypothesis in the context of Wikipedia. *Computer Speech & Language*, *25*(3), 716–740.

Menini, S., Sprugnoli, R., Moretti, G., Bignotti, E., Tonelli, S., & Lepri, B. (2017). RAMBLE ON: Tracing movements of popular historical figures. In A. Martins & A. Peñas (Eds.), *Proceedings of the software demonstrations of the 15th conference of the European chapter of the Association for Computational Linguistics* (pp. 77–80). Association for Computational Linguistics. www.aclweb.org/anthology/E17-3020.pdf.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331* (6014), 176–182.

Miller, D. & Slater, D. (2000). *The Internet: An ethnographic approach*. Berg.

Murray, C. (2003). *Human accomplishment: The pursuit of excellence in the arts and sciences, 800 B.C. to 1950*. Harper Collins.

Nemoto, K., & Gloor, P. A. (2011). Analyzing cultural differences in collaborative innovation networks by analyzing editing behavior in different-language Wikipedias. *Procedia – Social and Behavioral Sciences*, *26*, 180–190.

Orton-Johnson, K., & Prior, N. (Eds.). (2013). *Digital sociology: Critical perspectives*. Palgrave Macmillan.

Overell, S. E., & Rüger, S. (2011). View of the world according to Wikipedia: Are we all little Steinbergs? *Journal of Computational Science*, *2*(3), 193–197.

Page, L., Brin, S., Motwani, R. , & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the Web. *Technical Report. Stanford InfoLab*.

Popescu, A., & Grefenstette, G. (2010). Spatiotemporal mapping of Wikipedia concepts. *JDCL '10: Proceedings of the 10th annual joint conference on Digital libraries*, 129–138.

Reznik, I., & Shatalov, V. (2016). Hidden revolution of human priorities: An analysis of biographical data from Wikipedia. *Journal of informetrics*, *10*(1), 124–131.

Rieder, B., & Röhle, T. (2012). Digital methods: Five challenges. In D. M. Berry (Ed.), *Understanding Digital Humanities* (pp. 67–84). Palgrave Macmillan. www.doi.org/10.1057/9780230371934_4.

Rochat, Y. (2009). *Closeness centrality extended to unconnected graphs: The harmonic centrality index*. Lausanne, Institute of Applied Sciences.

Rogers, R. (2013). *Digital methods*. MIT Press.

Roll, U., Mittermeier, J., Diaz, G., Novosolov, M., Feldman, A., Itescu, Y., Meiri, S., & Grenyer, R. (2016). Using Wikipedia page views to explore the cultural importance of global reptiles. *Biological conservation*, 204, 42–50.

Rollin, G., Lages, J., & Shepelyansky, D. L. (2019). World influence of infectious diseases from Wikipedia network analysis. *IEEE* Access, *7*, 26073–26087.

Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences*, *111*(52), E5616–E5622. www.doi.org/10.1073/pnas.1410931111.

Saramäki, J., Kivelä, M., Onnela, J.-P., Kaski, K., & Kertész, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, *75*(2), Article 027105.

Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A.., Martino, M., Barabási, A.-L., & Helbing, D. (2014). A network framework of cultural history. *Science*, *345*(6196), 558–562.

Shane-Simpson, C., & Gillespie-Lynch, K. (2017). Examining potential mechanisms underlying the Wikipedia gender gap through a collaborative editing task. *Computers in Human Behavior*, *66*, 312–328.

Skiena, S., & Ward, C. B. (2013). *Who's bigger? Where historical figures really rank*. Cambridge University Press.

Venturini, T., Bounegru, L., Gray, J., & Rogers, R. (2018). A reality check(list) for digital methods. *New media & society*, *20*(11), 4195–4217.

Wickham, H. (2016). *rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2*. https://cran.r-project.org/package=rvest.

Yu A. Z., Ronen, S., Hu, K., Lu, T., & Hidalgo, C. A. (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific Data*, *3*, Article150075. www.doi.org/10.1038/sdata.2015.75.

Zinovyev, A., Czerwinska, U., Cantini, L., Barillot, E., Frahm, K. M., & Shepelyansky, D. L. (2020). Collective intelligence defines biological functions in Wikipedia as communities in the hidden protein connection network. *PLOS Computational Biology*,*16*(2), Article e1007652.

Zook, M., Dodge, M., Aoyama, Y., & Townsend, A. (2004). New digital geographies: Information, communication, and place. In S. D. Brunn, S. L. Cutter, & J. W. Harrington (Eds.), *Geography and Technology* (pp. 155–176). Springer.