



ARTICLE



<https://doi.org/10.1057/s41599-021-00801-1>

OPEN

# Complex networks reveal emergent interdisciplinary knowledge in Wikipedia

Gustavo A. Schwartz <sup>1,2</sup>✉

In the last 2 decades, a great amount of work has been done on data mining and knowledge discovery using complex networks. These works have provided insightful information about the structure and evolution of scientific activity, as well as important biomedical discoveries. However, interdisciplinary knowledge discovery, including disciplines other than science, is more complicated to implement because most of the available knowledge is not indexed. Here, a new method is presented for mining Wikipedia to unveil implicit interdisciplinary knowledge to map and understand how different disciplines (art, science, literature) are related to and interact with each other. Furthermore, the formalism of complex networks allows us to characterise both individual and collective behaviour of the different *elements* (people, ideas, works) within each discipline and among them. The results obtained agree with well-established interdisciplinary knowledge and show the ability of this method to boost quantitative studies. Note that relevant *elements* in different disciplines that rarely directly refer to each other may nonetheless have many implicit connections that impart them and their relationship with new meaning. Owing to the large number of available works and to the absence of cross-references among different disciplines, tracking these connections can be challenging. This approach aims to bridge this gap between the large amount of reported knowledge and the limited human capacity to find subtle connections and make sense of them.

<sup>1</sup>Centro de Física de Materiales (CSIC-UPV/EHU)—Materials Physics Center (MPC), San Sebastian, Gipuzkoa, Spain. <sup>2</sup>Donostia International Physics Center, San Sebastian, Gipuzkoa, Spain. ✉email: [gustavo.schwartz@csic.es](mailto:gustavo.schwartz@csic.es)

## Introduction

Specialisation has necessarily led to the fragmentation of knowledge, creating loosely connected disciplines in which discoveries in one area are hardly known in others. This implies that the flow of knowledge is severely restricted among disciplines or even among different areas within the same discipline. In recent decades, different approaches (Börner et al., 2003; Henry and McInnes 2017; Small et al., 2014) have been proposed to overcome this gap by means of, for example, co-occurrence or semantic models (Cory, 1997; Gabrilovich and Markovitch, 2007; Ibrahim et al., 2017) or bibliometric-based systems (Gates et al., 2019) that use citation information to find related items. However, notwithstanding the significant advances in these methods, mainly through the network of citing articles, interdisciplinary research lacks efficient tools for establishing quantitative connections among different disciplines (such as science, art and literature). This problem becomes even more important if we consider the amount of available knowledge, which is so large as to make it impossible for a human being to read or even access in its entirety. Suitable mining of public knowledge databases (Wikipedia, DBpedia, ConceptNet) can take us beyond these limitations, revealing surprising relationships among elements belonging to apparently distant disciplines.

Wikipedia is one of the most impressive collective creations: millions of anonymous editors work, in a non-coordinated way, to build the greatest source of knowledge that humanity has ever seen. Interestingly, in addition to the explicit knowledge contained in its articles, there is a vast amount of implicit learning that emerges from the underlying dense network of internal links that represent connections among people, ideas and works and constitutes a large conceptual network. Internal links refer to those links present in the main text of an article that connect relevant elements with other articles within Wikipedia. This giant network (~163M connections in the English version) can be converted into a directed graph and has actually been used in many studies ranging from computing semantic relatedness (Gabrilovich and Markovitch, 2007) to natural language processing (Gabrilovich and Markovitch, 2009). Inspired by these successful approaches and to overcome the lack of quantitative methods in interdisciplinary research, I propose a non-supervised method to reveal emergent knowledge in Wikipedia using network science.

As a proof of concept of the proposed method, I focus in this work on the relationship between the works of Albert Einstein and Pablo Picasso at the beginning of the twentieth century. Was it a coincidence that Picasso developed Cubism at approximately the same time that Einstein published his theory of relativity? Were they answering the same questions? Were they influenced by the same people/works? I show here that the use of appropriate network science tools may actually lead to self-emerging knowledge buried in the dense network of links of Wikipedia that not only shows the flow of knowledge between disciplines but also quantifies both individual and collective behaviour of the elements connecting different areas of knowledge.

## Methods

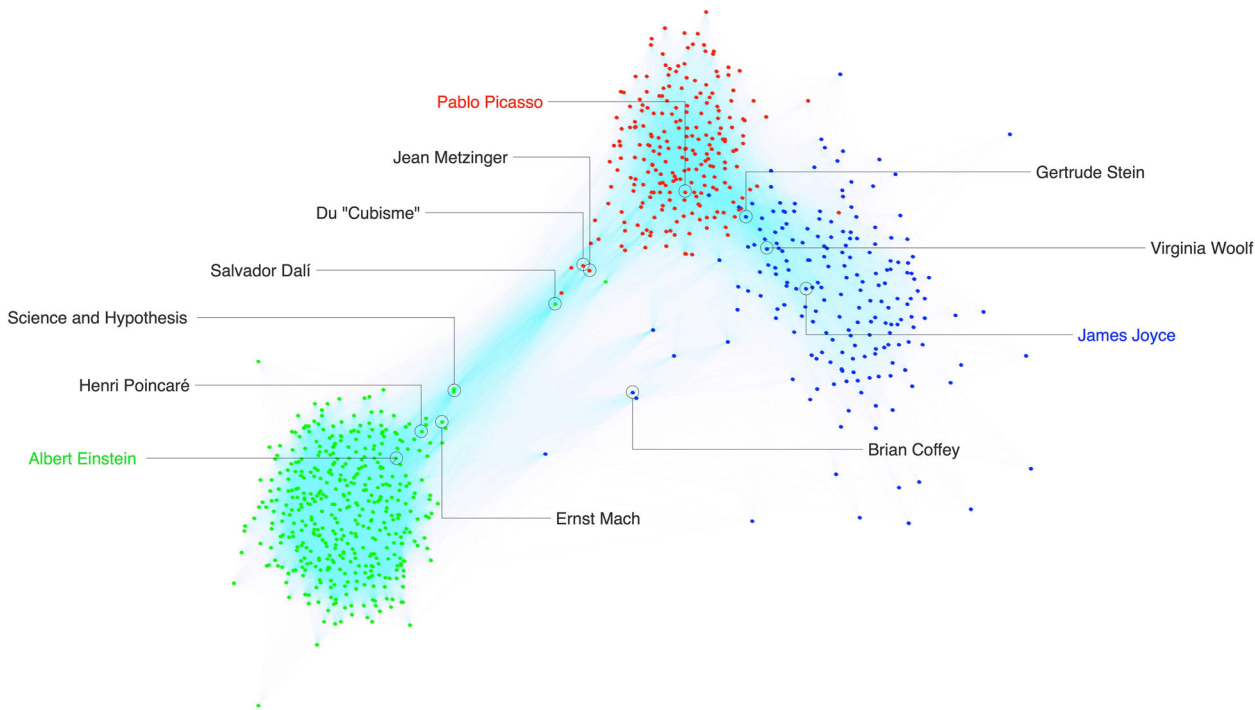
**Data set.** For this work, I used the publicly available WikiLinksGraphs datasets (Consonni et al., 2019) that contain the network of internal links (only those intentionally added by editors in the main text of the articles) for different dumps of Wikipedia. Specifically, this work was performed with a snapshot of Wikipedia from March 1, 2018. The original data set consists of a spreadsheet file with a size of 9.56 Gb (wikilink\_graph.

2018-03-01.csv) containing a table with four columns (page\_id\_from, page\_title\_from, page\_id\_to, page\_title\_to) and 163,380,007 rows corresponding to each of the internal links (from-to) of the complete Wikipedia content (see original article (Consonni et al., 2019) for more details). From this file, I took the second and fourth columns (page titles) and imported them into MATLAB in a directed unweighted graph (digraph) with 13,680,532 nodes and 163,380,007 edges. Then, it was necessary to iteratively remove poorly linked Wikipedia pages, those with zero in or outdegree that correspond to some redirect, disambiguation and other pages with no relevant information for the purpose of this work. The obtained graph, representing the network of internal links of Wikipedia, was called *wikiLinksClean* and contains 7,879,531 nodes and 150,995,780 edges (4.17 Gb). The number of nodes was still larger than the number of articles in Wikipedia (5.6M for 2018) due to the presence of many non-resolved redirects (pages without ‘real content’ that automatically send visitors to another page). This issue could affect the analysis and therefore was addressed in a later step as described below.

**Generate universe from seeds.** For the purpose of the present work, the idea is to unveil how two or more elements (concepts, people, works) are related and connected among them. Therefore, starting from these elements (entries of the Wikipedia hereafter called ‘seeds’), we defined a subgraph from the *wikiLinksClean* digraph by taking the nearest neighbours to each seed(s), i.e., all nodes in the digraph that are within distance  $d$  from each node  $s$ . Since the Wikipedia internal link network is a dense network (with an average shortest path length of ~4.1), a value of  $d \leq 2$  avoids having irrelevant links among seeds in the subgraph. For the case studied here, the seeds are ‘Pablo Picasso’, ‘Albert Einstein’ and ‘James Joyce’. Although we focus on the Einstein-Picasso relationship, including Joyce allows us to compare the relationships among art, science and literature and therefore to perform a deeper comparative analysis. Thus, based on these seeds, we obtained a subgraph (called the *universe*) containing 79,454 nodes and 3,166,325 edges. Once this subgraph was defined, the algorithm resolved the aforementioned issue with the non-resolved redirects by redirecting the inputs of nodes with outdegrees equal to one to the corresponding successor node and then removing the (redirect) nodes. Although this procedure can also remove a small fraction of weakly connected nodes, their influence on the final results is negligible. Additionally, after extracting the subgraph, nodes with zero in/outdegree were also removed. The obtained *universe* had 78,444 nodes and 3,159,866 edges.

**Measuring relatedness between nodes.** From the obtained *universe*, we wished to work only with the  $N$  most related nodes to each of the given seeds. Therefore, it was necessary to define a way of measuring the relatedness between each pair of nodes (Wikipedia articles). It is important to note that two articles can be strongly related even if there is no direct link between them. For instance, two articles can be co-linked by other articles, or they can co-link other articles. In these cases, we say that the two articles are structurally related. Among the many available metrics to measure distance (or relatedness) between elements in a complex network, the use of the *normalised Google distance* (Cilibrasi and Vitanyi, 2007) (NGD) provides excellent results for our purpose. It is defined as

$$d_{\text{in/out}}(a, b) = \frac{\log(\max(|A|, |B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|, |B|))}$$



**Fig. 1 Interdisciplinary knowledge map.** Each dot represents an entry of Wikipedia (related to people, concepts, works). Thin lines represent links between different elements according to the relatedness defined from the normalised Google distance (see Methods). Colours represent different clusters (disciplines) composed of the most related nodes to the given seeds: Picasso (red), Einstein (green) and Joyce (blue). Closeness among nodes/clusters is proportional to their relatedness.

where  $a$  and  $b$  refer to the two articles of interest,  $A$  and  $B$  represent the sets of nodes (articles) that link to/from ( $d_{in/out}$ )  $a$  and  $b$ , respectively, and  $W$  is the total number of nodes in the universe.  $\log$  refers to the base two logarithm whereas  $|A|$  represents the number of nodes in  $A$ . If  $|A \cap B| = 0$ , then the corresponding distance is infinite. There are two different distances between nodes  $a$  and  $b$ : one for nodes that link to  $a$  and  $b$  ( $d_{in}(a,b)$ ) and another for nodes that are linked from  $a$  and  $b$  ( $d_{out}(a,b)$ ). The total distance ( $d(a,b)$ ) was taken as the harmonic mean between the in/out distances. Finally, the relatedness between nodes  $a$  and  $b$  was defined as  $r(a,b) = \exp(-d(a,b))$ , which is always in the range  $[0,1]$ .

Based on this definition of relatedness between two nodes, we determined the  $N_j$  most related nodes to each of the given seeds, with  $N_j = k_j$ , where  $k_j$  is the outdegree of seed  $J$  in the *wikiLinksClean* digraph. Once the  $N_j$  'closest' nodes to each seed were known, the relatedness matrix ( $R$ ) was calculated for the nodes in this subset of the universe (hereafter called *near universe*). This matrix was then used to create an undirected weighted graph ( $g$ ) that represents the relationships among the different elements in the *near universe*. The weight of the link connecting nodes  $(i,j)$  was given by the corresponding element in the relatedness matrix ( $R_{i,j}$ ). The so obtained graph  $g$  contained (in our case) 856 nodes and 143,307 edges.

**Data clusterization and visualisation.** The way in which the  $N$  closest elements to each seed were chosen forces the formation of clusters and diminishes the inter-cluster connectivity. Therefore, the remaining links between elements of different clusters can be considered relevant enough for the purpose of this work. The nodes were assigned to different clusters according to what seed they were linked to in the original graph (*wikiLinksClean*). When a node was connected to more than a single seed, it was assigned

to the seed it was more related to. The obtained graph (with the identified clusters) was plotted using a force-directed layout that uses attractive forces between adjacent nodes and repulsive forces between distant nodes (Fruchterman and Reingold, 1991). The result is shown in Fig. 1.

**Data analysis**

Despite the qualitative insights that Fig. 1 can provide, complex network analysis allows quantitative characterisation of the structure and the interactions of the nodes in the network.

**Assortative mixing.** The preference for nodes of one type to link other nodes of the same type is well-known in social networks and is called *assortative mixing* or *homophily*. Following the ideas of Newman (Newman, 2003), an assortativity matrix  $A$  can be defined, where the elements  $a_{i,j}$  represent the sum of the weighted links in the network that connect nodes from clusters  $i$  and  $j$ . The normalised matrix is given by  $A_N = A/||A||$ , where  $||A||$  represents the sum of all the elements of  $A$ . Thus, the elements of  $A_N$  measure the fraction of links that fall between clusters of types  $i$  and  $j$ . In addition, Newman defines the assortativity coefficient as:

$$r = \frac{\text{Tr}(A_N) - ||A_N^2||}{1 - ||A_N^2||}$$

where  $\text{Tr}$  means the trace of the matrix. Note that  $r$  is zero for a randomly mixed network and one for a perfectly assortative network.

**Modularity.** Another way of measuring the 'quality' of a partition in a network is by means of the quantity  $Q$ , called modularity,

which for a weighted network is defined as (Newman, 2003):

$$Q = \frac{1}{W} \sum_c \left( W_c - \frac{S_c^2}{4W} \right)$$

where  $W$  is the total weight of all links in the network,  $W_c$  is the total weight of all internal links in cluster  $c$ ,  $S_c$  is the total weight (internal and external) of all nodes in  $c$ , and the sum runs over the clusters in the network. It is important to remind here that the maximum value of  $Q$  for a given network has a non-trivial value, which is given by:

$$Q_{\text{Max}} = 1 - \sum_c \frac{S_c^2}{4W^2}$$

**Weak and strong communities.** Given a network with different clusters (communities), we say that we have a *strong community* if the internal degree of each node (in such cluster) is greater than its external degree. On the other hand, if the *sum* of the internal degrees of all nodes exceeds the *sum* of their external degrees, it is a *weak community* (Fortunato, 2010).

**Openness.** For a given cluster, openness is defined as the percentage of nodes for which the external degree is greater than its internal degree.

**Average degree.** This parameter is defined for each cluster and is calculated as the diagonal elements of the normalised assortativity matrix (sum of the internal links for each cluster) divided by the number of nodes in the corresponding cluster.

**Stirling coefficient.** The Stirling coefficient is a well-known diversity index that measures variety, balance and disparity in an equilibrated way. For a given node, this coefficient is defined as (Stirling, 2007):

$$D = \sum_{ij(i \neq j)} d_{ij} p_i p_j$$

where  $d_{ij}$  is the distance between clusters  $i$  and  $j$  (defined as  $d_{ij} = \left(a_{ij}\right)^{-1}$ ) and where  $p_i$  and  $p_j$  are the fraction of the total degree (for the given node) that links the node to clusters  $i$  and  $j$ , respectively. The sum runs across all non-identical pairs of clusters ( $i \neq j$ ).

## Results and discussion

Figure 1 shows a visual representation of graph  $g$ , which constitutes a knowledge map for the relationships among Picasso-, Einstein- and Joyce-related elements. We can clearly observe three well-defined clusters corresponding to the elements most related to each seed. These clusters also account for the three domains to which each of the seeds belongs: art, science and literature. At a glance, we can observe differences in the size, density and relative distance among clusters. Artistic and literary domains are close and very well connected; much more than any of them with the cluster related to Einstein. On the other hand, science-related nodes show a stronger connection with those related to art than with those in the literary domain. We quantify these structural characteristics later.

From a qualitative point of view, the visual representation plotted in Fig. 1 using a force-directed layout correctly locates representative elements (people/works/concepts) on their precise relative position. For instance, Ernst Mach, Henri Poincaré and his book *Science and Hypothesis* are three of the most science-related elements that link scientific and artistic clusters. On the

other hand, Jean Metzinger and his oeuvre *Du 'Cubisme'* (written with Albert Gleizes) constitute important nodes that connect art-related elements with the scientific cluster. All these people and works constitute the so-called *knowledge dealers* (the most relevant elements that connect different clusters) between Picasso and Einstein, as was also stated by Arthur I. Miller in his book *Einstein, Picasso: Space, Time and the Beauty That Causes Havoc* (Miller, 2002), a historical research work that analyses their lives, works and common influences. Further important elements, shown but not highlighted in Fig. 1, are *Les Demoiselles d'Avignon*; Georges Braque; Maurice Princet, known as *le mathématicien du cubisme*; electromagnetism; modernism and Niels Bohr, among many others. In total, there are more than 75 elements that are present in both Miller's work and this study that provide a significant interaction and knowledge flow between both clusters. This qualitative agreement between traditional historical research and the proposed unsupervised machine learning approach opens interesting possibilities for this kind of interdisciplinary research.

It is also interesting to observe how Gertrude Stein is correctly located halfway between art and literature. Something similar happens with Brian Coffey, an Irish poet with a strong background in physical chemistry (he studied with Jean Baptiste Perrin, Nobel laureate in Physics) who appears somewhere between the literature- and science-related clusters. Salvador Dalí is located halfway between art and science, which agrees with his well-known interest in scientific concepts such as space, time and the fourth dimension, among many others. Interestingly, Picasso, Einstein and Joyce, even being the seeds of each cluster, do not play a relevant role as knowledge dealers.

Despite the impressive qualitative agreement between the knowledge map self-generated by the method proposed here and the careful historical analysis performed by Miller in his book, we can move further by using network science to quantify many of the structural characteristics of the network. To extract significant information from the topology of the network, we perform a threefold analysis: on the one hand, we analyse the whole network and its global metrics (modularity and assortativity); then, we focus on the cluster level, analysing their structure-domain dependence and their mutual interactions; and finally, we address the individual node connectivity assuming that in complex networks, nodes are connected according to the role they fulfil.

For the three clusters identified in Fig. 1, we can analyse their structural and relational characteristics. It is well-known from social network studies that nodes of one type prefer to link to other nodes of the same type. This behaviour is called assortative mixing (or homophily) and can be quantified by means of the assortativity coefficient as detailed in the Methods section. For a network with isolated clusters (with no links between elements of different groups), this coefficient is one; in the case analysed here, we find it to be 0.82 (see Table 1), which means that in spite of the high clusterization degree, there is still a non-negligible number of links connecting nodes from different clusters. The assortativity coefficient is calculated using the normalised assortativity matrix (see Table 2), in which diagonal elements account for the fraction of internal links within each cluster, whereas off-diagonal elements ( $a_{ij}$ ) indicate the fraction of links between clusters  $i$  and  $j$ . Thus, we can observe that almost 13% of the links are between elements of the art- and literature-related clusters, whereas only 1 and 2% of the links correspond to science-literature and science-art relationships, respectively. Another way of quantifying the partition of a network is by determining the so-called modularity ( $Q$ ). In this case, we obtain a modularity of 0.78 (with 0.89 being the maximum possible value for  $Q$ ); this means that the observed

modularity is ~88% of the expected maximum, which agrees with the observed high homophily.

Concerning cluster characterisation, we find that the three clusters constitute *weak communities* in the sense that the sum of the internal degrees of all nodes (for each cluster) exceeds the sum of their external degrees (see Methods). It is also possible to

define the *openness coefficient* (see Table 1), which represents the percentage of nodes with an external degree higher than its internal degree. The observed openness is 1.29%, 0.46% and 12.44% for the clusters related to Picasso, Einstein and Joyce, respectively. The science-related cluster is the most endogamic cluster, which is most likely related to its high intrinsic degree of specialisation. On the other hand, the literature-related cluster shows a large openness (compared to the other clusters), which is an indication of the broad range of interest to/from the literature domain.

Finally, at the level of individual nodes, we can use the well-known Stirling coefficient to identify the most relevant *knowledge dealers*. The Stirling coefficient, which has been successfully used for analysing diversity in science, technology and society (Stirling, 2007), gives an equilibrated measure of disparity, balance and variety. Table 3 shows the top ten nodes for each cluster according to the Stirling coefficient, which represent the most relevant knowledge dealers (either people, concepts or works) among different disciplines. The non-highlighted elements in the second column correspond to relevant people in the art-science relationship that were non-contemporaneous to the genesis of the Cubism and relativity ideas. Filtering these results with a temporal window (Hoffart et al., 2013) can improve the efficiency of the method in these cases.

**Table 1 Structural, relational and global properties of the network.**

Property	Cluster (seed name)		
	Picasso	Einstein	Joyce
<b>Structural</b>			
Size (nodes)	232	431	193
Diameter (max SPL)	1.23	1.21	1.45
Average (shortest) path length	0.74	0.67	0.80
Nodes density (dia)	189	356	134
Links density (dia)	0.16	0.50	0.07
Nodes density (APL)	314	646	241
Links density (APL)	0.26	0.90	0.13
<b>Relational</b>			
Community	Weak	Weak	Weak
Openness (%)	1.29	0.46	12.44
Average degree	1.36	0.66	1.25
<b>Global</b>			
Modularity	0.78 (0.89) [88%]		
Assortativity coefficient	0.82		

Structural, relational and global properties of the network obtained from the seeds 'Pablo Picasso', 'Albert Einstein' and 'James Joyce'. Diameters are defined as the maximum shortest path length. Densities are calculated using the diameter (dia) and the average shortest path length (APL). Weak (or strong) communities, openness and average degree are defined as specified in the Methods section. For modularity, the actual value, the maximum expected value and the percentage with respect to this maximum are shown. The maximum value for the assortativity coefficient is one. Unless specified, all quantities are in arbitrary units.

**Table 2 Normalised assortativity matrix.**

0.316	0.020	0.129
0	0.283	0.011
0	0	0.241

The elements of the normalised assortativity matrix ( $a_{ij}$ ) represent the fraction of weighted links that connect nodes between clusters of types  $i$  and  $j$ . Only the upper triangular matrix is defined.

### Conclusions

The complex networks approach proposed here shows the need to consider interdisciplinary knowledge as a whole instead of focusing on local and specific information. Moreover, it highlights the emergence of collective knowledge that can arise from individual uncoordinated actions. Thus, millions of small contributions from many different people lead to recovering firmly established historical and epistemological facts, as well as unveiling the flow of information among disciplines and understanding its structure and dynamics.

This method provides both qualitative and quantitative tools for the analysis of the relationships between different *elements*, such as people, ideas or works. The nature of these elements and their interactions makes necessary a systemic approach that considers subtle connections and structural characteristics of the networks. This approach has been shown to successfully address all these features by means of a multiscale analysis that allows studying individual elements, clusters and the whole network at the same time. It is expected that this proposal not only helps answer old questions but, even more importantly, opens the door to new interrogations so far hardly imaginable.

**Table 3 Stirling ranking for nodes on each cluster.**

Pablo Picasso	Albert Einstein	James Joyce
<b>Jean Metzinger</b>	Salvador Dalí	Brian Coffey
<b>La Femme au Cheval</b>	Wolfgang Paalen	Jim Norton (Irish actor)
<b>Du 'Cubisme'</b>	J. Robert Oppenheimer	Thomas Pynchon
Proto-Cubism	<b>Ernst Mach</b>	<b>Literary modernism</b>
<b>Le goûter (Tea Time)</b>	Genius (U.S. TV series)	The Third Policeman
L'Oiseau bleu (Metzinger)	<b>Niels Bohr</b>	Finnegans Wake
<b>Cubism</b>	<b>Ludwig Boltzmann</b>	Watership Down
Le Port (painting)	<b>Science and Hypothesis</b>	<b>Modernism</b>
<b>Section d'Or</b>	Enrico Fermi	<b>T. S. Eliot</b>
<b>Georges Seurat</b>	<b>Henri Poincaré</b>	<b>Twentieth-century English literature</b>

Top ten elements for each cluster according to the Stirling coefficient. Highlighted items (in the first two columns) correspond to those people, works or concepts mentioned by Miller in his research. Non-highlighted items (in the last two columns) represent those elements out of the temporal window (1900-1915) analysed here that did not take part in the cross pollination process of that period and therefore they were neither cited by Miller.

## Data availability

The supported data sources are publicly available, and their sources are presented in the 'Methods' section. Additional information regarding the analysis can be obtained from the author upon request.

Received: 21 December 2020; Accepted: 28 April 2021;  
Published online: 25 May 2021

## References

- Börner K, Chen CM, Boyack KW (2003) Visualizing knowledge domains. *Ann Rev Inform Sci Technol* 37:179–255
- Cilibrasi RL, Vitanyi PMB (2007) The google similarity distance. *IEEE Trans Knowl Data Eng* 19(3):370–383
- Consonni C, Laniado D, Montresor A (2019) WikiLinkGraphs: a complete, longitudinal and multi-language dataset of the wikipedia link networks. *Proc of the Thirteenth Int AAAI Conf on Web and Social Media*. Association for the Advancement of Artificial Intelligence ([www.aaai.org](http://www.aaai.org)) pp. 598–607
- Cory KA (1997) Discovering hidden analogies in an online humanities database. *Comput Humanit* 31(1):1–12. <https://doi.org/10.1023/A:1000422220677>
- Fortunato S (2010) Community detection in graphs. *Phys Rep* 486:75–174
- Fruchterman T, Reingold E (1991) Graph drawing by force-directed placement. *Softw Pract Exper* 21(11):1129–1164
- Gabrilovich E, Markovitch S (2007) Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *Int Joint Conf Artificial Intell* 7:1606–1611
- Gabrilovich E, Markovitch S (2009) Wikipedia-based semantic interpretation for natural language processing. *J Artificial Intell Res* 34:443–498
- Gates AJ, Ke Q, Varol O et al. (2019) Nature's reach: narrow work has broad impact. *Nature* 575:32–34. <https://doi.org/10.1038/d41586-019-03308-7>
- Henry S, McInnes BT (2017) Literature based discovery: models, methods, and trends. *J Biomed Inform* 74:20–32. <https://doi.org/10.1016/j.jbi.2017.08.011>
- Hoffart J, Suchanek FM, Berberich K et al. (2013) YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artif Intell* 194:28–61. <https://doi.org/10.1016/j.artint.2012.06.001>
- Ibrahim M, Danforth CM, Dodds PS (2017) Connecting every bit of knowledge: the structure of Wikipedia's first link network. *J Comput Sci* 19:21–30
- Miller AI (2002) *Einstein, Picasso: space, time, and the beauty that causes havoc*. Basic Books, New York, NY
- Newman MEJ (2003) The structure and function of complex networks. *Soc Indus Appl Math* 45(2):167–256
- Small H, Boyack KW, Klavans R (2014) Identifying emerging topics in science and technology. *Res Policy* 43:1450–1467
- Stirling A (2007) A general framework for analysing diversity in science, technology and society. *J R Soc Interface* 4(15):707–719

## Acknowledgements

The author acknowledges the financial support from the Spanish Government 'Ministerio de Ciencia e Innovación' (PID2019-104650GB-C21) and from the Donostia International Physics Center (Programa Mestizajes), as well as the support of NVIDIA Corporation with the donation of a Quadro RTX 6000 GPU used for this research.

## Competing interests

The author declares no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to G.A.S.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021