


A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses

Muh-Chyun Tang¹  · Yun Jen Cheng¹ ·
Kuang Hua Chen¹

Received: 31 March 2017 / Published online: 5 September 2017
© Akadémiai Kiadó, Budapest, Hungary 2017

Abstract As digital humanities continues to expand and become more inclusive, little is known about the extent to which its knowledge is integrated. A bibliometric analysis of published literature in digital humanities was conducted to examine the degree of its intellectual cohesion over time (1989–2014). Co-authorship, article co-citation, and bibliographic coupling networks were generated so SNA based cohesion analysis can be applied. Modularity maximization partition was also performed to both co-citation and “author bibliographic coupling” networks to identify main research interests manifested in the literature. The results show that, as publications in digital humanities continue to grow, its diversity and coherence, two hallmarks of interdisciplinarity, have shown signs of becoming more robust. The co-author network, however, remained rather fragmented, with collaboration mainly limited by language and geographic boundaries. The domain specific practices in digital humanities that might contribute to such fragmentation was discussed.

Keywords Digital humanities · Co-citation analysis · Co-author network · Network cohesion · Interdisciplinarity · Bibliographic coupling · Intellectual cohesion · Knowledge integration

Introduction

Digital humanities (DH), formally known as humanities computing, is a field of research mainly concerned with the intersection between computing and various disciplines in humanities. On “What is humanities computing?” McCarty stated that “it is

✉ Muh-Chyun Tang
mctang@ntu.edu.tw

Yun Jen Cheng
yjcheng0314@gmail.com

Kuang Hua Chen
khchen@ntu.edu.tw

¹ Department of Library and Information Science, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan, ROC

methodological in nature and interdisciplinary in scope...focusing both on the pragmatic issues of how computing assists scholarship and teaching in the disciplines and on the theoretical problems of shift in perspective brought about by computing” (McCarty 2005). As an emerging area of research that draws research interests and expertise from multiple disciplines, DH presents a fertile ground for the study of “knowledge integration” process. One crucial aspect of interdisciplinarity, beyond the presence of knowledge of diverse origins, is the integration of them into a coherent enterprise (Rafols and Meyer 2010; Porter and Rafols 2009; Porter et al. 2007). The cognitive integration of concepts, theories, methods and/or results from diverse fields is considered as the hallmark of interdisciplinary research (Wagner et al. 2011; Levallois et al. 2012). According to Rafols and Meyer (2010), “knowledge integration” is “... a process that is characterized by high cognitive heterogeneity (*diversity*) and increases in relational structure (*coherence*)” where coherence is defined by the extent to which specific topics, concepts, tools are interconnected. In a bibliometrics context, coherence can be represented by whether the basic bibliographic elements (e.g. authors, articles, keywords, or publication sources) in a set of published literature form a tightly or loosely connected structure. While in Rafols and Meyer (2010) network coherence is used to represent how well knowledge is integrated across formally defined disciplinary boundaries, the network construct of “cohesion” has also been applied to characterize knowledge integration among different subspecialties *within* a discipline (Moody 2004; Carolan 2008) a research area (Gondal 2011; Levallois et al. 2012; Liu and Xia 2015) or a scholarly community (Rawlings et al. 2015).

While research initiatives in DH often share the common methodological outlook, it remains an empirical question whether DH as a field has consolidated or remained fragmented over time. Little is known about what kind of bibliographic network topologies might result from the collaboration between computer sciences and humanities, whose scholarly practices stand in direct contrast to each other in many aspects. Taking a bibliometric approach, this study aims to fill the gap.

Literature review

As a continuously evolving field, DH has attracted a wide range of knowledge interests and expertise, which is manifested by its disciplinary and institutional diversity (Svensson 2010). Yet, the degree of cross-fertilization among these research efforts essential for interdisciplinarity is less evident. A distinction is often made between multi-disciplinarity and interdisciplinarity as the later suggests, besides the presence of diverse bodies of knowledge, the integration and synthesis of them into a coherent whole (Wagner et al. 2011). Porter et al. (2007) defines interdisciplinarity as a mode of research activity that integrates theories, concepts, techniques or data from two or more bodies of specialization or research practice.

As cognitive integration taking place in the research process is difficult to observe at a large scale, researchers often opt to infer knowledge integration in a field from the outcome of the research, namely, its published literature (Wagner et al. 2011). Knowledge integration in literature can be observed both at the micro and macro level. At the micro level, the diversity of author background and cited references represented within individual articles are often taken as evidences of knowledge integration (Bordons et al. 2004; Porter and Rafols 2009; Rafols and Meyer 2010). At the macro level, researchers interested in the degree of intellectual cohesion of a field as whole often opt to examine the degree of

cohesion or interconnectedness of its published literature or collaborative network. (Moody 2004; Moody and White 2003; Acedo et al. 2006a; Vidgen et al. 2007; Carolan 2008; Gondal 2011; Levallois et al. 2012; Liu and Xia 2015; Rawlings et al. 2015).

Based on the assumption that the knowledge integration process in research communities depends heavily on the topology of the underlying social network, Moody used the structure of collaborative (i.e. co-authorship of journal articles) network to trace the macro structure of subspecialties in Sociology over time (Moody 2004). The concept of “structure cohesion” or “connectivity” in network analysis was used to measure the degree of social cohesion in Sociology. Moody (2004) discussed three types of network structures: star production, small world, and structurally cohesive and surmised on the corresponding collaborative practices each might represent. Moody (2004) believe that a structurally cohesive collaboration network model signals the presence of “permeable theoretical boundaries and generic methods” that allows scholars specialized in particular empirical or theoretical skills to collaborate freely. He added that, if enough scholars engage in this kind of cross-fertilization, mixing across multiple areas, there will be few clear divisions presented in the collaborative network (Moody 2004). A “star production network”, on the other hand, represents a scholarly community where the cohesion of the network hinges on a small set of prominent scholars or seminal works at the core, with most others located at the periphery. A small-world network is where the local clustering is high, yet thanks to a few boundary-spanning shortcuts, the average path length remains low (Milgram 1967; Watts 1999; Watts and Strogatz 1998). Similarly, Carolan (2008) used network structure of articles published by a leading journal in Education to examine how well the heterogeneous set of ideas and practices were integrated within the discipline. Besides aforementioned three models, a “plural world model” (Condliffe Lagemann 1989) was proposed. Being the most fragmented of all models, a “plural word model” manifested itself in isolated components that lack of linkages to each other, suggesting a variety of specialized research communities that contribute little to the integration of knowledge in the field. One novel aspect of Carolan (2008) was the use of server log data, instead of traditional bibliographic data, to generate the network, where the strength of relationship between two articles was determined by the degree of overlap of their readership. The resulting article interlock network exhibits the features of both small-world and structural cohesive models. Lately in their study of the co-authorship network in the interdisciplinary field of “evolution of cooperation”, Liu and Xia (2015) traced its trajectory from a few local structures to a global structure of “chained communities” that demonstrated the features of a small world model.

While there have been previous efforts to using co-citation (Leydesdorff and Salah 2010), and co-words (Wang and Inaba 2009) analysis to map the domain of DH, they were relatively small in scale and did not address specifically the question of its intellectual cohesion. In the present study, three types of network resulted from co-authorship, co-citation, and bibliographic coupling were generated from published literature in digital humanities. Social network analytical methods were then applied to measure the interconnectedness of these networks. Of particular interests are the degree of cohesion or integration manifested in these networks. It is hoped that the identification of network topology (see Fig. 1) in these networks would provide insights into the scholarly practices such as the collaborative patterns, the degree of interdisciplinarity, as well as the state of cognitive consensus within DH. Modularity-based community detection method was also applied to help identify major research interests in the field.

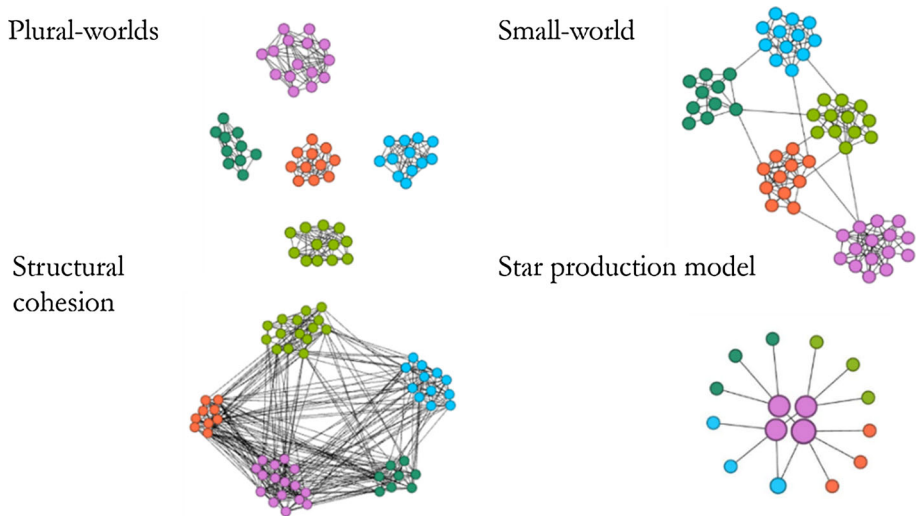


Fig. 1 Network topology in a bibliographic network

Data collection procedures

As DH is not a formally recognized field in the main citation databases, it poses a challenge to identify all relevant literature that constitutes its knowledge base. In this study, the publications of DH were identified by both keywords search with Scopus, one of the largest citation databases, and journals with an explicit digital humanities orientation. For the Scopus search, the query (“digital humanities” OR “digital humanity” OR “humanities computing” OR “humanity computing”), was used to search the field of title, keywords and abstract, which resulted in a set of 1967 articles and book chapters. As it is likely that publications in DH do not necessarily have those keywords, a complementary set of articles was created by retrieving all the articles published in the following journals from their websites: *Journal of Digital Humanity*, *Digital Humanities Quarterly*, *International Journal of Humanities and Arts Computing*, *Digital Medievalist*, *Digital Studies*, *Literary and Linguistics Computing*, among which only *Literary and Linguistics Computing* was indexed by Scopus at the time of data collection. The list of journals was selected because they are published by the members of the Alliance of Digital Humanities Organizations (ADHO). The union of the two sets constitutes the “target set” from which bibliographic networks can be generated and analyzed. The “target set” contains 2115 articles, 2787 authors, and 3469 keywords.

The bibliographic information of the articles, including title, author(s), source (journal, conference proceeding, and book), author keywords as well as citation counts in the target set were then downloaded for further analysis. Three types of bibliographic networks were generated: co-citation, bibliographic coupling and co-authorship networks. The co-author network was built based on the co-occurrence of author names appearing in the author field for all articles in the target set. Name disambiguation was facilitated by the author ID assigned by Scopus, for articles not indexed by Scopus, authors names were examined by the researchers to ensure consistency. To study how the co-author network evolves over time, a 5-year overlapping time slice was used to divide our target set, resulting in a total of

22 co-citation networks (i.e. from 1989–1993, 1990–1994... to 2010–2014. See Fig. 2). The same 5 year sliding window was applied to the co-citation and bibliographic networks.

To generate the co-citation networks, pair-wise matching of all the articles’ received citations has to be conducted. This is done by using Google Scholar’s citation tracing function. Google Scholar was chosen because it is the most comprehensive citation database. As many of the publications in our target set have not been indexed by the other two major citation indexes, WoS and Scopus, Google scholar became the only feasible option to trace the citation each article in our target set has received. The citations received by every article in our target set were identified and downloaded so pair-wise matching could be performed to identify shared citations. Bibliographic coupling was enabled by matching the reference list of article retrieved from Scopus, therefore only articles published by journals indexed by Scopus were included in this part of analysis.

Data analysis

The longitudinal approach allows us to trace the trajectories of knowledge diversity and integration in digital humanities over time. The “balance” aspect of diversity proposed in (Porter and Rafols 2009) is measured by Gini index, which calculated on the distribution of author assigned keywords. A higher Gini index indicates a few dominant topics addressed by many different articles, a lower Gini index, on the other hand, signals a more balanced and therefore diverse interested in the field. Social network analytical metrics were applied to measure the degree of cohesion of the bibliographic networks. There commonly used cohesion metrics were applied to assess the intellectual cohesion of the international community of DH: the percentage of nodes remains in the largest component, clustering coefficient, and average path length. A component is the maximal sets of nodes in which every node can reach every other by some path, no matter how long the paths are. The relative size of the giant component, the largest subgraph within a network, is often used to indicate how interconnected a network is. Cluster coefficient measures the degree to which one’s neighbors are also connected. Average path lengths represents, on average, the steps it takes for a node to reach another node in the network. A high clustering coefficient, coupled with short network diameter or average path length, signal the presence of a small-world phenomenon.

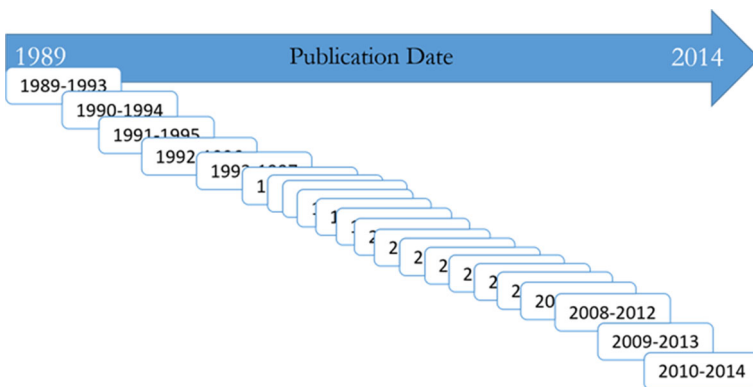


Fig. 2 Sliding publication window to generate multiple network

Another application of bibliographic analysis is the identification and visualization of the subtopics within a field (see, for example, Morris and Van der Veer Martens 2008; McCain 1990; White and McCain 1998). Modularity compares the observed fraction of links within the cluster to expected fraction if links were distributed randomly. Modularity maximization algorithm was performed to identify important sub-areas (Blondel et al. 2008). Centrality analyses were also conducted to identify the prominent actors in the networks. This part of analyses was conducted at journal, article, and author levels. Of particularly interest was to identify items that contribute most to the cohesion of the networks.

Results

The diversity of research topics in DH

Drawing on Stirling (2007), Rafols and Meyer (2010) pointed out when measuring diversity, one needs to consider three aspects, namely, the number of discipline cited (variety), the degree of their concentration (balance), and how dissimilar these categories (disparity) are. Previous classification based approach to measuring disciplinary diversity often involved utilizing subject categories (SC) assigned by Web of Science (WoS) at the journal level. (Rafols and Meyer 2010; Porter and Rafols 2009). Without the benefit of a formal classification system such as SC in WoS used in (Porter and Rafols 2009), it is not practical to determine the dis/similarity of subjects treated in the literature so the dimension of disparity was left out of our diversity analysis. Instead, we used author assigned keywords instead as the indication of subjects treated in the literature of DH. The variety and the balance of the keywords were measured to assess its diversity.

The keywords retrieved from our target set were first manually examined to control for spelling variations. Figures 3 and 4 shows, respectively, the growth of publications in DH and the gradual rise of the number of distinct author assigned keywords over time, indicating the growing variety of research topics in DH (Fig. 4).

Gini index was then applied to measure the balance of the subjects covered in our target set. Gini index is commonly used to measure the skewedness of the distribution. High Gini index of the keyword distribution signal the existence of few dominant keywords. Figure 5 shows a gradually rising Gini indexing over time, which might be interpreted as a gradual consolidation of research efforts. Yet notice that even with the gentle rising slope, the degree of concentration remains low at below 0.35, which suggests considerably balanced research interests in DH.

To examine the evolving of research interests in DH, we further divided the most frequent author assigned keywords into three 9 year periods: 1987–1995, 1996–2004, and 2005–2014 (see “Appendix 1”). An interesting observation is the presence of keywords related to social media (e.g. “Twitter,” “social network”), digital culture (e.g. “Web 2.0,” “videogame, mobile devices”, and “design”) in recent years, which seems to suggest the expanding of DH from computer processing of literature to the study of all aspects of humanists concerns of different aspects of digital culture.

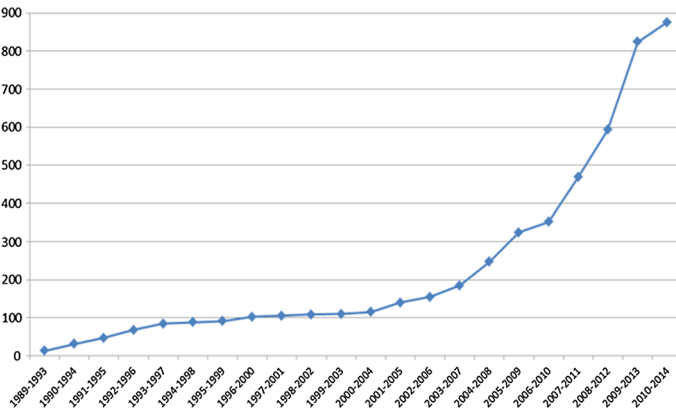


Fig. 3 The growth of articles in DH

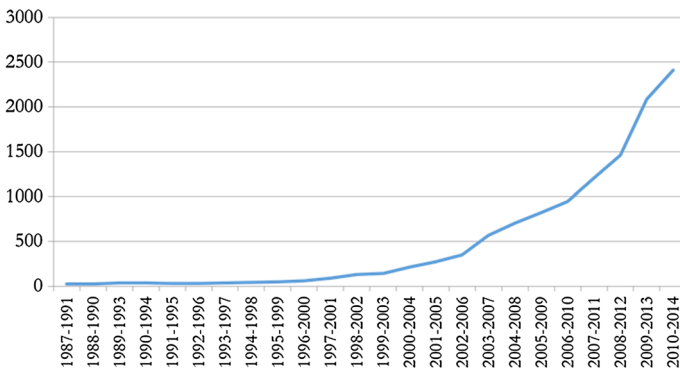


Fig. 4 The rise of distinct author assigned keywords over time

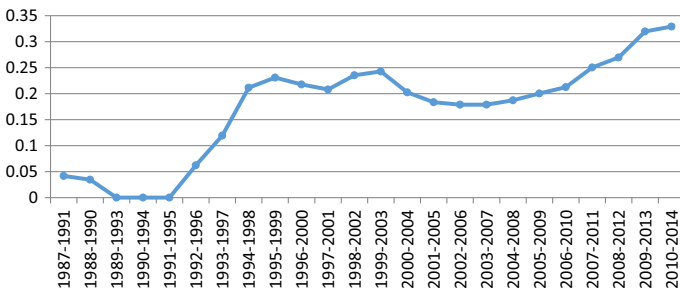


Fig. 5 Gini index of keyword distribution over time

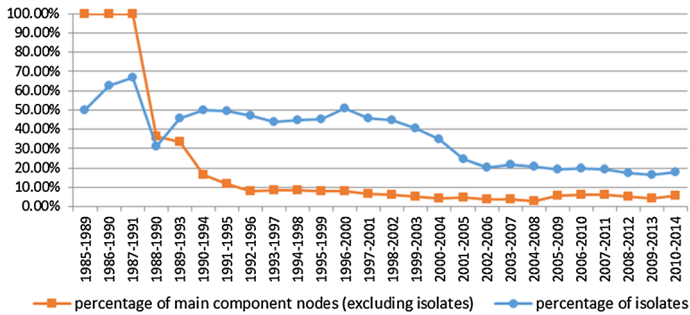


Fig. 6 Percentage of nodes in the main component in the co-authorship network

Co-author network in DH

Percentage of nodes constitutes the giant component is often used as a basic indication of network cohesion. As shown in Fig. 6, the percentage of the nodes in the giant component in the co-author network hovered only below 20% even after discounting the isolates, which is extremely low compared to other disciplines or research areas (see Table 1). Notice the contrast is especially striking with sciences, medicine, and IT. The low percentage of nodes in the giant component, coupled with extremely high clustering coefficient and modularity (Fig. 7), indicated that most collaborations took place at the local level, lacking global “shortcuts” found in the small world model to hold the network together.

A closer examination of the co-authorship network shows that, beyond the two largest components, there were relatively few international collaborations. The largest component was composed of scholars mainly from the U.S. (28.53), Canada (27.12%), U.K. (26.84%), and Germany (10.45%); the second largest component was composed of scholars from the U.S. (39.71%), the Netherlands (30.88%), and Japan (5.88%); and the third component was composed almost entirely of Italian scholars (96.97%) see (Fig. 8).

As shown in Fig. 9, beyond the three largest components, the rest of the components are very small. Furthermore, the dominance of single-country co-authorship becomes even more salient in smaller components and the network is mainly highly fractured along national boundaries. The distribution of the main participating countries in DH research is shown in Fig. 10.

Co-citation network

Figure 11 shows the long-term trend of nodes in the giant component of the co-citation network over time. A jump of the percentage of nodes in the giant component can be observed in the early 2000, which then gradually levelled off. Notice also that there is a significant portion of nodes are isolated. The dip in recent years might be more likely due to the artifact of citation window than a sign of disintegration. Another two cohesion measures, average geodesic distance and clustering coefficient seem to calibrate such interpretation. A rather short average distance between nodes in the giant component, hovering around 4. The cluster coefficient stayed steadily high at 0.60, indicating dense local collaboration, which, coupled with the short average distance, fit the parameters of a typical small world model.

Table 1 Components and clustering coefficient across different fields. Sources: ^a Acedo et al. (2006a), ^b Newman (2001), ^c Moody (2004) and ^d Liu and Xia (2015)

	DH	Management and Organization ^a	Medicine ^b	Physics ^b	High energy physics ^b	IT ^b	Sociology ^c	Evolution of cooperation ^d
# of nodes	2787	10,176	152,0251	52,909	56,627	11,994	197,976	3670
Average degree	3.8	2,43	18.1	9.7	173	3.59	–	3,409
Main component (size)	354	4625	139,5693	44,337	49,002	6396	68,285	1127
Main component (percentage)	12.7	45.4	92.6	85.4	88.7	57.2	34.5	30.71
Size of second largest component	68	23	49	18	69	42	–	–
Clustering coefficient	0.927	0.681	0.066	0.43	0.726	0.496	0.194	0.632

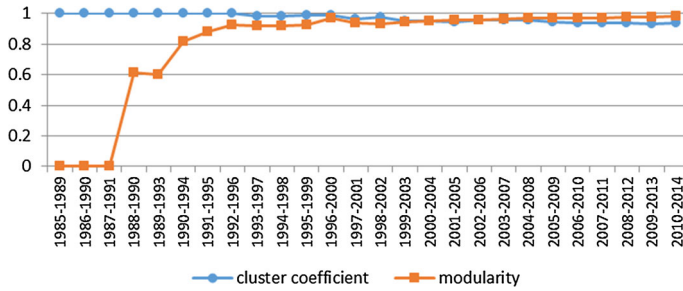


Fig. 7 Trends of clustering coefficient and modularity in co-authorship network

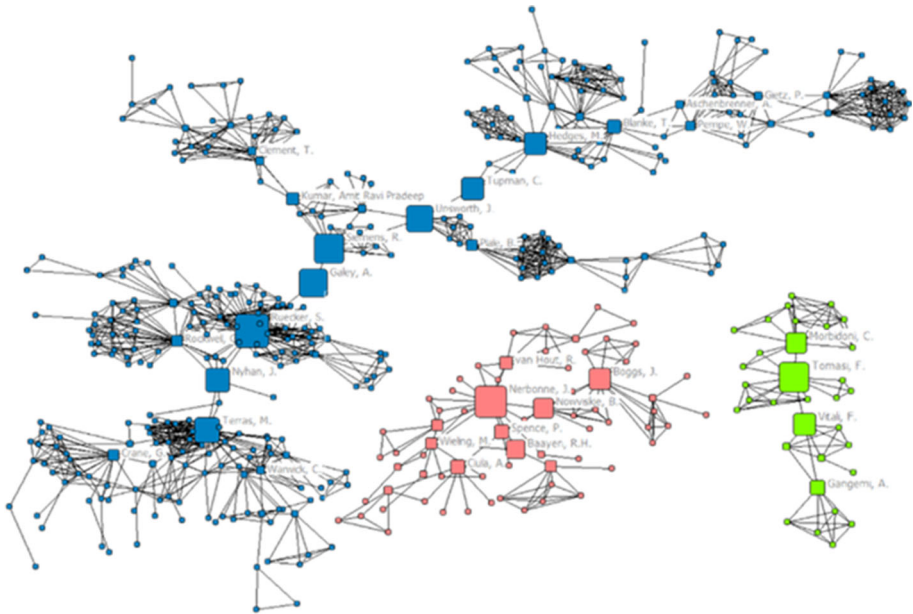


Fig. 8 Three largest components in the co-authorship network

Centrality analysis was also performed to identify important works in DH. Of particular interest are publications with highest betweenness centrality relative to their degree centrality as they play a relatively more important role in holding the network together (see “Appendix 2”).

One earlier study of a small set of articles in digital humanities identified two main co-citation clusters of journals, one made up of specialist journals devoted to computing application in humanities, and the other, a group of library and information science journals addressing the issues in digitalizing archives and libraries (Leydesdorff and Salah 2010). We visualized the source co-citation network using MDS where the distance between nodes signifying the similarity of their co-citation profiles (Fig. 12). The size of the nodes represents betweenness centrality, while different colors group frequently co-cited sources using fraction algorithm in UCINET (Borgatti et al. 2002).

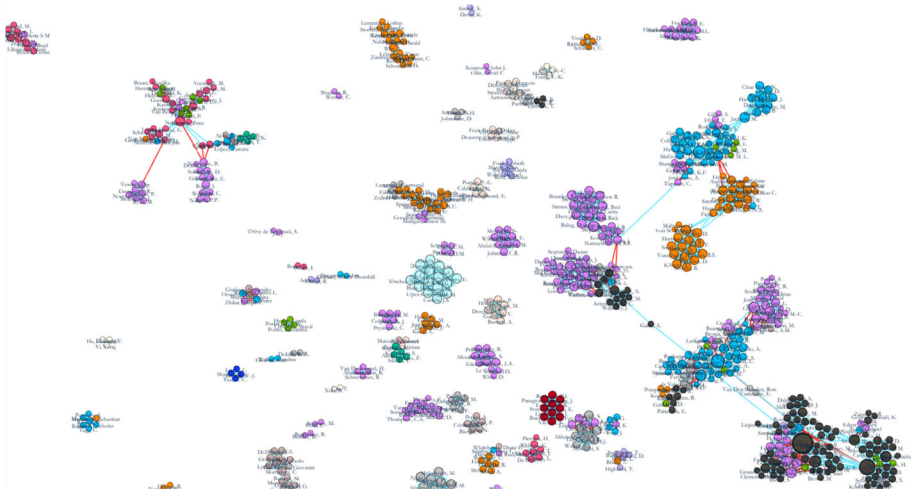


Fig. 9 Part of the co-authorship network. Nodes colored by nationality of author affiliations. Cross- country linkages are marked. (Color figure online)

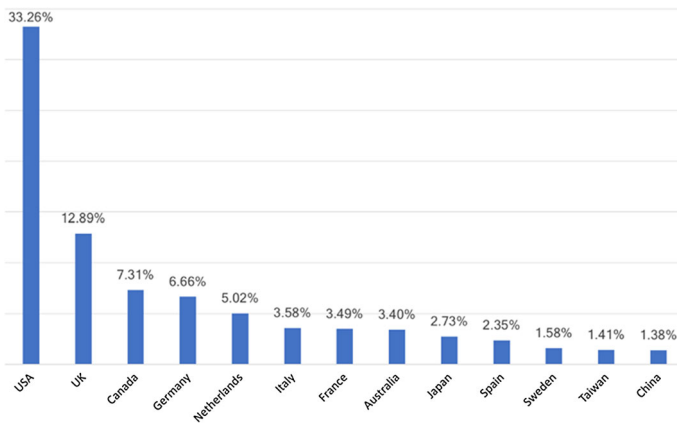


Fig. 10 Distribution of author country affiliation

Several sources that have an explicit focus in DH were in the red group; they also tended to have a higher betweenness centrality, indicating a strong bridging role that reflects the breadth and reach of their contents. Sources with a distinct emphasis in computation and computational linguistics tended to be marked in gray. The blue group represents the specialized interests in media, museum, archeology, biology and geography. The black group is not as easy to interpret, still we can spot several conference proceedings, especially those related to digital libraries in this group. Notice also that the stress value is quite high, signaling a high distortion of the two-dimensional visualization to the original data.

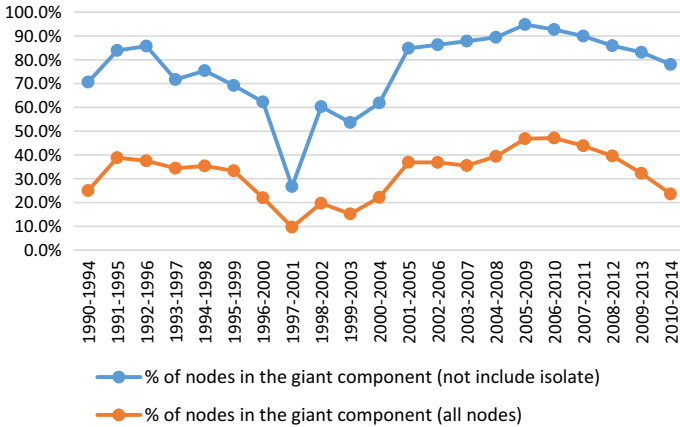


Fig. 11 Percent of nodes within the largest components in co-citation network over time

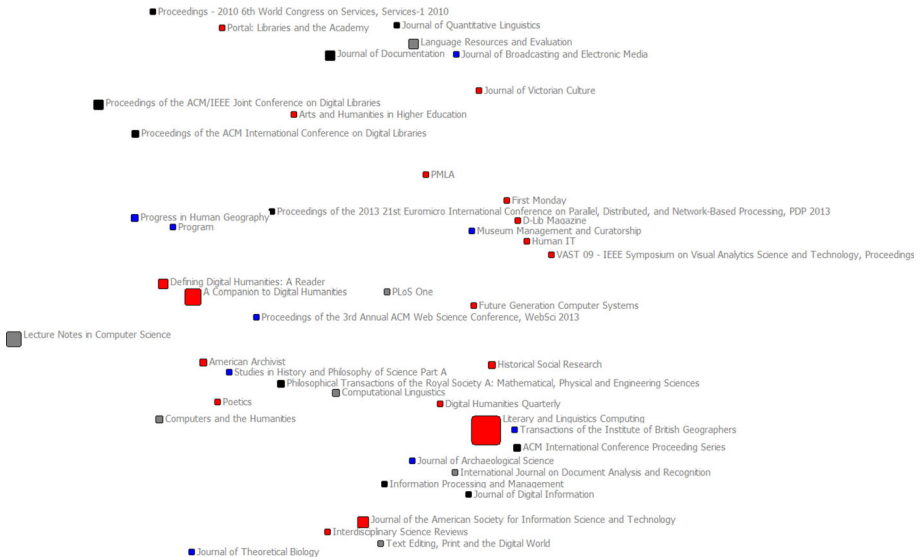


Fig. 12 MDS representation of source co-citation profiles. Stress value = 0.417

Modularity maximization based community detection procedure was performed to identify densely connected subgroups within the co-citation network (Blondel et al. 2008) (see Fig. 13). To interpret the modularity classes, author assigned keywords were extracted and their frequency tallied. The keywords with higher discriminative value (i.e. TF-IDF weight), that is, keywords appear frequently in certain classes but infrequently anywhere else, were then selected by the researchers. The success of the modularity based classification was mixed as some classes are more interpretable than others. The distinction of the classes was especially difficult among class 6, 7, and 8 because of the significant overlapping of keywords related to various digitalization efforts. However, keywords explicitly

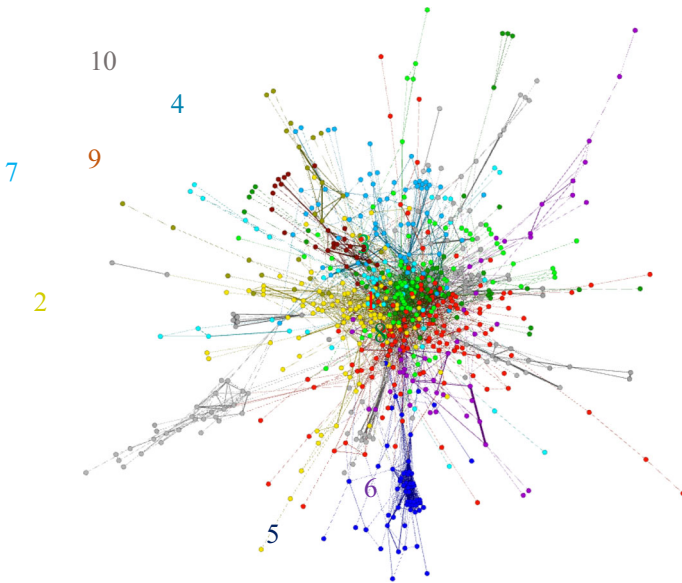


Fig. 13 Visualization of modularity classes in co-citation network, modularity = 0.62

associated with computational techniques were present in class 6, while class 7 seems to be more about collection development (see Table 2).

Bibliographic coupling network

The bibliographic coupling network was generated by pair-wise comparison of cited references retrieved from Scopus. A threshold of shared 4 citations in the reference lists was set to dichotomize the network as a largest drop of the percentage of the nodes in the giant component was observed between 3 (25.30%) and 4 (15%) shared items in the reference lists. The threshold of 4 was used because it generated the highest modularity value. Figure 14 shows the percentage of nodes contained in the giant component relative to the size of the network over time. A sudden rise of nodes in the giant component was observed in the early 2009–2013 period, signaling increased consensus about knowledge base in DH.

Likewise, modularity maximization graph partition (Blondel et al. 2008) was applied to the bibliographic coupling network to identify subspecialties in DH, only instead of individual articles/book chapters, authors were used as the unit of analysis (Fig. 12) McCain (1998). The author bibliographic coupling network comprised individual authors as the nodes (Zhao and Strotmann 2008). This is done so to increase the interpretability of the partitions as the original network is more fragmented. To interpret the modularity classes, prominent authors in each class, as ranked by their h-index, citation count and centrality, were identified. Their specialties were determined by the researchers according to their publications and affiliations listed in Scopus. As shown in Table 3, the larger the classes, the more heterogeneous the topics were present, with the exception of “author attribution”, which is also readily recognizable in the co-citation network. Notice that the resulting modularity classes are more well-defined than those produced by the co-citation network, especially those involved specific computing or digital technologies (Fig. 15).

Table 2 Top 10 modularity classes in co-citation network

Class	Percentage (%)	Specialties	Discriminative keywords
1	6.4	General interests	Infrastructure, digitalization, GIS, literary studies, interdisciplinarity, HCI, data science, information retrieval, digital heritage, social network, visualization, handwriting recognition, genealogy, historical GIS, maps, mobile devices, phonetics, cultural criticism, spatio-temporal analysis
2	4.84	Metadata	Annotation, linked data, XML, concurrent markup, RDF, TEI, ontology, electronic editing
3	4.4	User studies	User studies, research tools, information retrieval, semantic web, virtual research environment, scholarship
4	3.62	Authorship attribution	Authorship attribution, computational stylistic, stylometry, artificial intelligence, text-mining, textual analysis, feature selection,
5	3.13	Literary theories	Literary criticism, literary theories, computer criticism, Bakhtin, electronic texts, versioning, computer games, multi-media, structuralism, reception theory
6	2.98	Digital libraries/text analysis	Digital libraries, digital archives, cervantes project, computational linguistics, corpus linguistics, algorithmic criticism, automatic text analysis, authorship attribution, orthographic variation, support vector machines, machine learning, N-gram, object-oriented modeling, spelling, simulation, training
7	2	Digital libraries/infrastructure	Metadata, archive, electronic editions, markup SGML, digital library, collection management, cultural heritage, cultural institutions, earmark, image restoration, manuscript restoration, visualization, multispectral imaging, text collation
8	1.42	Digitalization	Humanities computing, digitalization, collaboration, digital preservation, research infrastructure, digital curation, digital methodology, digital philology, digital ecosystem
9	1.22	Scholarly communication	Scholarly communication, bibliometrics, citation analysis, disciplinary differences, web 2.0, interdisciplinarity, computer-mediated communication
10	1.22	Dialectology	Sociolinguistics, dialect, dialectology, dialectometry, phonetic (dis)similarity, language change in time, spectrogram, cochleagram, barkfilter, Danish, Dutch, German, American English

Discussion and conclusion

In this study we set out to examine the degree of topical diversity and intellectual cohesion in the emerging field of digital humanities as manifested in its published literature. Digital humanities presents an interesting case for a bibliometric based domain analysis because, firstly, as a research area strongly influenced by technological advance, its identity and scope are continuously redefined and debated; secondly, as an interdisciplinary field it combines various disciplines in humanities and computing, whose scholarly cultures and communication practices are in many aspects drastically different from each other (Archambault et al. 2005; Nederhof 2006). Without a domain specific index also makes it difficult to demarcate the domain of DH's published literature. We approach this issue by

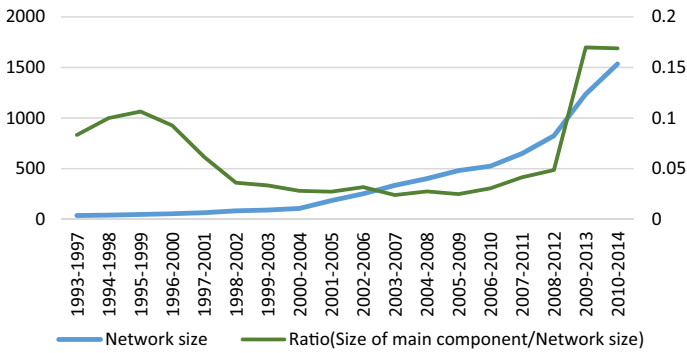


Fig. 14 Percentage of nodes in the giant component in the bibliographic coupling network

Table 3 Top 10 modularity classes in author bibliographic coupling network

Class	Percentage	Specialties	Author specialties
1	8.99	General interests	HCI, cultural geography, anthropology, ancient history, archive, social media, literary theories, rhetoric and narratives, mobile phone, video game
2	5.86	Digital infrastructure	Data mining, ontology, semantic web, metadata, OCR, name entity recognition, cultural heritage
3	5.35	Author attribution	Stylometry, author attribution, dialects
4	5.10	Digital libraries	e-learning, e-book, digital libraries, HCI, information seeking, mobile learning, mobile users, internet studies, intelligent user interface, data curation, linked data, informatics, 3D imaging
5	3.71	Social media	Social media, social network, new media and learning, mobile data, social studies of technology
6	3.31	Data mining	Machine learning, topic modeling, information retrieval, collaborative research platform, data sharing, bibliometrics, word sense disambiguation, visualization, social network analysis
7	2.86	3D visualization	3D modeling, 3D scanning, remote sensing, GIS
8	2.7	Topic modeling	Topic modeling
9	1.4	Library and Information Science	Digital libraries, information seeking, virtual reference services, metadata sharing, open access
10	1.37	Text recognition and analysis	Handwritten text recognition, historical spelling variants, character recognition

combining the inclusion of all journals affiliated with ADHO, DH’s most established scholarly society, and keywords search with Scopus, one of the most comprehensive citation databases available. Five year overlapping time slice was applied so longitudinal trends could be observed. It has been pointed out that previous studies in interdisciplinarity has been limited in their static view of the state of knowledge at a certain point in time (Wagner et al. 2011). A longitudinal approach such as ours allows us to trace DH’s trajectory of topic diversity and cohesion over time. The results show that the degree of knowledge diversity is high, as demonstrated by the increase of distinct author assigned keywords and the balance of its distribution. The presence of increasing publication

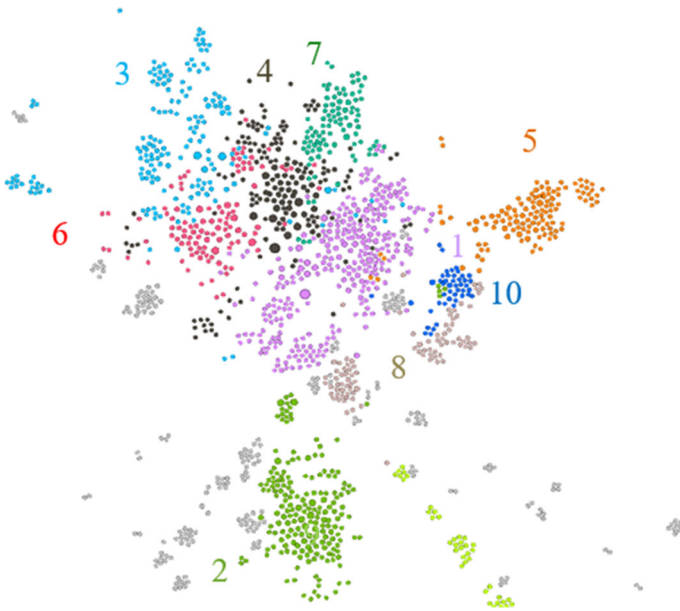


Fig. 15 Visualization of modularity classes in bibliographic coupling network, modularity = 0.51

sources from diverse disciplines also attests to its expansion. A recent broadening of interests from using computing as a methodology to addressing various humanists engagement with digital culture was also reflected in the keywords (Svensson 2010).

The case for knowledge integration is, however, mixed. Network cohesion analysis showed that, while both co-citation, and to a less degree, bibliographic coupling networks have gradually grown to be more cohesive, co-authorship network has remained extremely fragmented. The co-citation network is the most cohesive and exhibits the features of the “small world” model of high local clustering and short average geodesic distances. The trend toward more integration started in the early 2000 as demonstrated by noticeable increase of the percentage of nodes in the giant component. Being relatively more cohesive, the co-citation network afforded more domain analyses. Centrality analysis was performed to identify articles, monographs and other publication sources that contribute most to the integration of the network.

As for bibliographic coupling network, a sudden increase of the percentage of nodes in the giant component was spotted in the 2009–2013 period, and continued to rise in the latest, 2010–2014 period, which suggests a growing consensus on the field’s knowledge base.

Besides structural cohesion, modularity maximization partition was performed on the relative more cohesive co-citation and author bibliographic coupling networks to identify important areas of research interests. For co-citation analysis, author assigned keywords were used to help interpret the modularity classes. Keywords with higher discriminative value (i.e. TF-IDF weight) to each class were considered to be more representative. As for bibliographic coupling network, an “author bibliographic coupling” (Zhao and Strotmann 2008) approach was adopted where authors, instead of citing articles, were used as the basic unit of analysis. This is to help increase the cohesion and interpretability of the classes as the original network is rather sparse. Compared to topics or subjects, authors

have shown to give a more refined representation of a particular research interest in previous works of author co-citation analysis (McCain 1990; Åström 2007). Indeed, in the case of DH, the bibliographic coupling induced author network gives a more well-defined mapping of specialties than the co-citation network.

The co-authorship network was shown to be highly fragmented, consisting of numerous small components that resemble the “plural worlds” model, without an extensive giant component often observed in neighboring fields, such as digital libraries, for example (Liu et al. 2008). The clustering coefficient is very high, indicating that collaborators tends to form closely-knitted clusters. Furthermore, it was shown that they are fractured mainly along geographic and language boundaries, with very few international collaborations. We suspect that this is mainly due to the national or regional character in humanities (Nederhof 2006). The other reason might be the fact that “humanities” is an aggregate term that encompasses a wide variety of disciplines who might not share the same research concerns or methodologies. Additionally, it has been well known that humanities scholars tend to work independently and co-authorship is relatively rare, which might also contribute to the fragmentation. Even though we do observe a gradual increase of the average number of authors per paper in recent years to about 2.3, it is still much lower than other fields.

One novel aspect of this study is to juxtapose three types of network: co-authorship, co-citation, and bibliographic coupling, all of which are widely used for analyzing a research domain, but have rarely been combined and compared (Boyack and Klavans 2010; Yan and Ding 2012), especially for the specific purpose of examining the degree of intellectual cohesion of an academic field. The fact that three different networks exhibit distinct topologies in the case of DH suggests that cautions need to be taken when only one type of network is used to assess disciplinary cohesion. It also points to the need to take into consideration the domain specific scholarly practices that undergird its production and dissemination of knowledge. Among all the forms of knowledge flow or exchange of ideas, co-authorship is the most formal and demands the highest commitment. It is also most explicitly social and arguably more effective for the exchange of implicit knowledge. Because of its higher threshold for connection, it might not come as a surprise that DH’s co-author network turned out to be the most disintegrated. Yet one wonders whether such a great discrepancy in cohesion between the citation based (i.e. co-citation and bibliographic coupling) network and the co-authorship network is peculiar to DH or widely observable in other emerging areas of research. For example, a densely interconnected “core” of authors was found to play a crucial role in network cohesion in other emerging areas of research (Liu et al. 2008; Gondal 2011; Liu and Xia 2015) is ostensibly missing in DH. And it remains an empirical question whether the degree of cohesion in citation based network and co-authorship network are better aligned with each other in those fields. We suspect that several factors relating to domain specific scholarly practices in DH might help explain its lack of far-reaching collaboration. In Whitley’s theory of intellectual and social organization of academic fields, the dimensions of “mutual dependence” and “task uncertainty” were used to characterize the culture and practices of an academic field (Whitley 2000; Fry 2006). The “task uncertainty” dimension refers to the degree to which research questions, theoretical framework and methodological procedures are shared among scholars, whereas “mutual dependence” refers to the extent to which scholars have to rely on colleagues’ contribution to advance one’s work. Traditionally, humanities research has been characterized by high task uncertainty and low mutual dependence. It has also been shown that “corpus linguistics”, an important branch in digital humanities, is characterized by considerable variations in research goals and work procedures as well as high “technical uncertainty” (Fry 2006). These domain specific circumstances might help explain the sparsity of the co-author network.

There are obviously limitations to our study, most noticeable of which is the lack of analysis of knowledge integration at the micro level, namely, the drawing of expertise from diverse disciplines to tackle a research question, which can be measured by looking into the composition of authors' disciplinary backgrounds (Levallois et al. 2012) and distribution of subject categories appearing in the reference list (Bordons et al. 2004; Porter and Rafols 2009; Rafols and Meyer 2010). Future studies of the knowledge integration at the work or team level will greatly complement the present research to provide a better understanding of the nature of knowledge integration in DH. Such an endeavor, however, will have to first overcome the challenge of the lack of coverage of DH literature in the major citation databases. Related to database coverage issue is how the bibliographic universe of a field can be defined, especially one that is highly interdisciplinary and continuously evolving like DH. Lacking a comprehensive bibliography, as one applied in (Leefmann et al. 2016) for the bibliometric analysis of Neuroethics, makes it difficult to claim comprehensiveness of the publication of a field. Even though we have tried to be as inclusive as possible in our selection of relevant literature by combining keyword search and key journal selection, it is inevitable that some of the highly representative sources or works might still be missed out, especially those published in non-English languages. Using Google scholar as the source of co-citation data also entails inherent bias in the scope of citing documents (Falagas et al. 2008). While these data limitations demand cautions when interpreting the results, there is little reason to see an even more comprehensive dataset would lead to drastically different conclusions. If anything, the inclusion of non-English literature is more likely to increase rather than reduce the fragmentation of the co-author network. One might also point out the perils of not using a fixed citation window when generating the co-citation networks. Indeed, using a fixed citation window is more theoretically sound, especially for fields where articles continue to be cited years after its publications. But in our case, we believe that it does not pose a threat to the conclusion we draw from the longitudinal data. It is concluded that the field of DH, when represented by its co-citation network, has become more cohesive over time because the percentage of nodes in the giant component has grown steadily (except the latest few periods because it takes time for the later publications to be cited) and the average path length has also become shorter, despite the fact that earlier networks were allowed more time to accumulate citations, therefore, everything being equal, should have formed more cohesive networks than later periods. However, this is not the case. On the contrary, the co-citation in later periods were shown to be more cohesive. Furthermore, even with the growing number of publications, the clustering coefficients remained steady, which is another indication of growing cohesion over time.

Appendix 1

See Table 4.

Table 4 The most frequent author keywords in three stages

1987-1995		frq	1996-2004		frq	2005-2014		frq
1	Humanities computing	2	Humanities computing	8	Digital humanities	202		
2	Applications of microcomputers	1	SGML	6	Humanities	49		
3	Art history	1	Dialect	4	Digital libraries	39		
4	Artificial intelligence	1	Dialectology	4	Big data	32		
5	CAI in literature	1	Dialectometry	4	Ontologies	25		
6	Chaos theory	1	Digital libraries	4	Technology	22		
7	Cognition	1	Electronic texts	4	Text mining	20		
8	Collaborative writing and exams	1	Literary criticism	4	GIS	19		
9	Computing	1	Literary theory	4	Metadata	19		
10	Conceptual convergence	1	Metadata	4	Annotation	18		
11	Concordances	1	XML	4	Collaboration	18		
12	Creativity	1	Archive	3	Digitization	18		
13	Databases	1	Link	3	History	18		
14	Discourse	1	Multi-media	3	Linked open data	18		
15	Education	1	TEI	3	XML	17		
16	Electronic text editions	1	Authorship attribution	2	Archives	16		
17	Extrovert	1	Bakhtin	2	Information retrieval	16		
18	Fuzzy sets	1	Browsing and navigation in large hypermedia	2	TEI	16		
19	Gender	1	Commercial text	2	Digital history	14		
20	Human computer interaction	1	Computer criticism	2	Semantic web	14		
21	Humanities	1	Computer games	2	Social media	14		
22	ideation	1	Computer-aided learning	2	Visualization	14		
23	Interdiscursivity	1	Computer-assisted language learning	2	Cultural heritage	13		
24	Intertextuality	1	Content tagging	2	Database	13		
25	Introvert	1	Corpus-based rule development for lexical acquisition	2	Social network	13		
26	LAN	1	Digital images	2	Authorship attribution	12		
27	Language systems	1	Digital preservation	2	Interdisciplinarity	12		
28	Lexicography	1	Document analysis	2	New media	12		
29	Literary criticism	1	Document type definition (DTD) design	2	Open access	12		
30	Literary theory	1	Education	2	Corpus linguistics	11		
31	Microcomputers	1	Electronic editing	2	Libraries	11		
32	Minorities	1	Expert systems	2	Narrative	11		
33	Musicology	1	French theory	2	Digital archives	10		
34	Networks	1	Higher education	2	Humanities computing	10		
35	Pseudonyms	1	Hypertext	2	Internet	10		
36	Questionnaire survey	1	Image restoration	2	Methodology	10		

Table 4 continued

	1987-1995	frq	1996–2004	frq	2005–2014	frq
37	Reader response	1	Image-based humanities computing	2	Scholarship	10
38	Riddles	1	Improvisation	2	Videogames	10
39	Semantics	1	Industrial text	2	Crowdsourcing	9
40	SGML	1	Interactive	2	Natural language processing	9
41	Social history	1	It concepts	2	RDF	9
42	TEI	1	Linguistics	2	Twitter	9
43	Text	1	Literary studies	2	Artificial intelligence	8
44	Text encoding	1	Literary text	2	Computational linguistics	8
45	Textual criticism	1	On-line delivery	2	Computer science	8
46	University computing	1	Pedagogy	2	culture	8
47			Performance	2	Design	8
48			Posthuman	2	Digital	8
49			Protocols	2	Digital media	8
50			Reception theory	2	digital scholarship	8
51			Research	2	Digital technologies	8
52			Resource discovery	2	Education	8
53			Sator	2	eResearch	8
54			Semantic mark up	2	Evaluation	8
55			Signifying	2	Human–computer interaction (HCI)	8
56			Statistics	2	Information technology	8
57			Structuralism	2	Manuscripts	8
58			Stylistics	2	Modelling	8
59			Text encoding and rendering	2	Research	8
60			Text/image coupling	2	Research infrastructures	8
61			Textual criticism versus control	2	Social sciences	8
62			Textual studies	2	Academic libraries	7
63			Transcription/editing tools	2	augmented reality	7
64			Undergraduate	2	Data mining	7
65			Versioning	2	Digital preservation	7
66			Winbrill	2	Epistemology	7
67			Womens writing	2	Historical GIS	7
68			4D-CAD	1	Information visualization	7
69			ABET 2000	1	Linguistics	7
70			Absorption	1	Literature	7
71			Accreditation	1	scholarly communication	7
72			Aesthetics	1	Arts	6
73			AFS algebra	1	Audience	6

Table 4 continued

1987-1995	frq	1996-2004	frq	2005-2014	frq
74		AFS structure	1	Computational stylistics	6
75		American English	1	Data Analysis	6
76		Analysis Technique	1	Data visualization	6
77		Animation	1	Ebooks	6
78		Architecture	1	eHumanities	6
79		Artificial intelligence	1	Ethics	6
80		Artists-as-researchers	1	Mobile devices	6
81		Autocorrelation	1	Multimedia	6
82		Barkfilter	1	Museums	6
83		Best educational practices	1	Newspapers	6
84		Bibliometrics	1	Storytelling	6
85		Carboniferous	1	User studies	6
86		Cervantes digital library (CDL)	1	Web 2.0	6

Appendix 2

See Table 5.

Table 5 Significant works in DH co-citation network

Betweenness centrality		Degree centrality
1	The history of humanities computing	1
2	What is digital humanities and what's it doing in English departments	3
3	If you build it will they come The LAIRAH study: quantifying the use of online resources in the arts and humanities through statistical analysis of user log data	7
4	Killer applications in digital humanities	2
5	Humanities computing as digital humanities	6
6	Stylistic analysis and authorship studies	8
7	A data structure for representing multi-version texts online	18
8	Information seeking by humanities scholars	16
9	Text-encoding, theories of the text, and the 'work-site'	19
10	Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces	4
11	The state of authorship attribution studies: some problems and solutions	11
12	Marking texts of many dimensions	12
13	Toward modeling the social edition: an approach to understanding the electronic scholarly edition in the context of new and emerging social media	53
14	Enabled backchannel: conference twitter use by digital humanists	61
15	All the way through: testing for authorship in different frequency strata	15
16	Electronic scholarly editing	10

Table 5 continued

Betweenness centrality		Degree centrality
17	Human computing - modelling with meaning	65
18	Digital infrastructure and the Homer multitext project	42
19	The inhibition of geographical information in digital humanities scholarship	26
20	Pliny: A model for digital support of scholarship	29
21	eHumanities desktop—an online system for corpus management and analysis in support of computing in the humanities	122
22	Deploying general-purpose virtual research environments for humanities research	77
23	An evaluation of text classification methods for literary study	32
24	Open source critical editions: a rationale	24
25	The state of the digital humanities: a report and a critique	13
26	Meaning and mining: the impact of implicit assumptions in data mining for the humanities	9
27	Speculative computing: aesthetic provocations in humanities computing	34
28	Texts into databases: the evolving field of new-style prosopography	62
29	Thinking about interpretation: pliny and scholarship in the humanities	22
30	Digital visualization as a scholarly activity	14
31	Blobjects: digital museum catalogs and diverse user communities	127
32	Transcription maximized; expense minimized crowdsourcing and editing the collected works of Jeremy Bentham	38
33	The structure of the arts and humanities citation index: a mapping on the basis of aggregated citations among 1157 journals	37
34	Visual GISing: bringing together corpus linguistics and geographical information systems	70
35	Disciplined: using educational studies to analyse ‘Humanities Computing’	27
36	Beyond digital incunabula: modeling the next generation of digital libraries	23
37	Understanding the information and communication technology needs of the e-humanist	28
38	Methodological commons: arts and humanities e-science fundamentals	44
39	Rabbit/duck grammars: a validation method for overlapping structures	59
40	Computational contributions to the humanities	217
41	Computer-mediated texts and textuality: theory and Practice	5
42	Towards information retrieval on historical document collections: the role of matching procedures and special lexica	74
43	Mapping the English Lake District: a literary GIS	347
44	Making web annotations persistent over time	86
45	American digital history	169
46	A controlled-corpus experiment in authorship identification by cross-entropy	21
47	Quantifying evidence in forensic authorship analysis	228
48	A framework for contextual information in digital collections	235
49	The digital humanities and humanities computing: an introduction	82
50	Annotating digital libraries and electronic editions in a collaborative and semantic perspective	170

References

- Acedo, F. J., Barroso, C., Casanueva, C., & Galán, J. L. (2006a). Co-authorship in management and organizational studies: An empirical and network analysis. *Journal of Management Studies*, 43(5), 957–983.
- Acedo, F. J., Barroso, C., & Galan, J. L. (2006b). The resource-based theory: dissemination and main trends. *Strategic Management Journal*, 27(7), 621–636.
- Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingras, Y. (2005). Welcome to the linguistic warp zone: Benchmarking scientific output in the social sciences and humanities. In *Proceedings of the ISSI 2005 conference* (pp. 24–28).
- Åström, F. (2007). Changes in the LIS research front: Time-sliced co-citation analyses of LIS journal articles, 1990–2004. *Journal of the American Society for Information Science and Technology*, 58(7), 947–957.
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.
- Bordons, M., Morillo, F., & Gómez, I. (2004). Analysis of cross-disciplinary research through bibliometric tools. In H. F. Moed, W. Glänzel & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 437–456). Dordrecht: Kluwer.
- Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for windows: Software for social network analysis*. Harvard, MA: Analytic Technologies.
- Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12), 2389–2404.
- Carolan, B. V. (2008). The structure of educational research: The role of multivocality in promoting cohesion in an article interlock network. *Social Networks*, 30(1), 69–82.
- Condliffe Lagemann, E. (1989). The plural worlds of educational research. *History of Education Quarterly*, 29(2), 183–214.
- Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, scopus, web of science, and google scholar: Strengths and weaknesses. *The FASEB Journal*, 22(2), 338–342.
- Fry, J. (2006). Scholarly research and information practices: A domain analytic approach. *Information Processing and Management*, 42(1), 299–316.
- Gondal, N. (2011). The local and global structure of knowledge production in an emergent research field: An exponential random graph analysis. *Social Networks*, 33(1), 20–30.
- Leefmann, J., Levallois, C., & Hildt, E. (2016). Neuroethics 1995–2012. A bibliometric analysis of the guiding themes of an emerging research field. *Frontiers in Human Neuroscience*, 10, 336.
- Levallois, C., Clithero, J. A., Wouters, P., Smidts, A., & Huettel, S. A. (2012). Translating upwards: Linking the neural and social sciences via neuroeconomics. *Nature Reviews Neuroscience*, 13(11), 789–797.
- Leydesdorff, L., & Salah, A. A. A. (2010). Maps on the basis of the Arts & Humanities Citation Index: The journals Leonardo and Art Journal versus “digital humanities” as a topic. *Journal of the American Society for Information Science and Technology*, 61(4), 787–801.
- Liu, P., & Xia, H. (2015). Structure and evolution of co-authorship network in an interdisciplinary research field. *Scientometrics*, 103(1), 101–134.
- Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2008). Co-authorship networks in the digital library research community. *Information Processing and Management*, 41(6), 1462–1480.
- McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6), 433.
- McCain, K. W. (1998). Neural networks research in context: A longitudinal journal cocitation analysis of an emerging interdisciplinary field. *Scientometrics*, 41(3), 389–410.
- McCarty, W. (2005). *Humanities computing*. Basingstoke: Palgrave Macmillan.
- Milgram, S. (1967). The small world problem. *Psychology Today*, 1, 61–67.
- Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69(2), 213–238.
- Moody, J., & White, D. R. (2003). Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, 103–127.
- Morris, S. A., & Van der Veer Martens, B. (2008). Mapping research specialties. *Annual Review of Information Science and Technology*, 42(1), 213–295.
- Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, 66(1), 81–100.
- Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Science*, 98(2), 404–409.

- Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719–745.
- Porter, A. L., Cohen, A. S., Roessner, J. D., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, *72*(1), 117–147.
- Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, *82*(2), 263–287.
- Rawlings, C. M., McFarland, D. A., Dahlander, L., & Wang, D. (2015). Streams of thought: Knowledge flows and intellectual cohesion in a multidisciplinary era. *Social Forces*, *93*(4), 1687–1722.
- Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society, Interface*, *4*(15), 707–719.
- Svensson, P. (2010). The landscape of digital humanities. *Digital Humanities*, *4*(1).
- Vidgen, R., Henneberg, S., & Naudé, P. (2007). What sort of community is the European Conference on Information Systems? A social network analysis 1993–2005. *European Journal of Information Systems*, *16*(1), 5–19.
- Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., et al. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, *5*(1), 14–26.
- Wang, X., & Inaba, M. (2009). Analyzing structures and evolution of digital humanities based on correspondence analysis and co-word analysis. *アート・リサーチ*, *9*, 123–134.
- Watts, D. J. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of sociology*, *105*(2), 493–527.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, *393*(6684), 440.
- White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, *49*(4), 327–355.
- Whitley, R. (2000). *The social and intellectual organization of the sciences*. Oxford: Oxford University Press.
- Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326.
- Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science and Technology*, *59*(13), 2070–2086.