**TANG — Longitudinal study**

<7 figures; 2 tables>

**A Longitudinal Study of Knowledge Integration in Digital Humanities Using Co-Citation Analysis**

Tang, M.-C., Cheng, Y.-J; Chen, K. and Hsiang, J

Although digital humanities continues to expand and become more inclusive, little is known about the extent to which its published knowledge is integrated. A longitudinal co-citation analysis of publications in digital humanities was conducted to examine the degree of cohesion of its published knowledge over time (1989–2014). The measurement of cohesion was performed at source and individual article levels. The results show that, while the publications in digital humanities continue to grow, its diversity and coherence, two hallmarks of interdisciplinarity, remain robust. Betweenness centrality was used to identify the nodes that contribute most significantly to the cohesion of the source and individual networks.

\* \* \*

One crucial aspect of Interdisciplinarity is the integration of different bodies of knowledge into a coherent enterprise (Rafols and Meyer, 2010; Porter et al., 2007). As an emerging field that draws research interests from multiple fields, DH provides a fertile ground for the study of whether and how different bodies of knowledge are integrated. The cognitive integration of concepts, theories, methods, and/or results from diverse fields is considered as the hallmark of interdisciplinary research. Such knowledge integration is much easier to observe at the micro level, where the diversity of authorship and cited references present within an individual article are often taken as evidences of interdisciplinarity. The level of integration at the macro level, which can be measured by the degree of interconnectedness among published works within a field, is not as readily accessible. While research initiatives in DH often share the common methodological outlook, there is little empirical study regarding whether DH as a research field has consolidated or remained fragmented. Taking a bibliometrics approach, this study aims to fill the gap. We used a co-citation network to examine the degree of interdisciplinarity in the published knowledge in DH. Three types of bibliographic elements resulting from the co-citation network were examined: keywords, publication sources, and individual articles. Social network analytical methods were applied to measure the diversity and interconnectedness of these elements, as well as the most prominent actors in these networks.

**Data Collection Procedures**

The publications of DH were identified by both keywords search and journals with an explicit digital humanities orientation. Keyword search with Scopus resulted in a set of 1,967 articles and book chapters. As publications in DH do not necessarily have those keywords, a complementary set of articles was created by retrieving all the articles published in the six journals published by the members of the Alliance of Digital Humanities Organizations (ADHO). The union of the two sets constitutes the target set (N = 2509) from which article co-citation networks were formed and analyzed.

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. The citations received by every article in our target set were identified and downloaded; then pair-wise matching was performed to identify shared citation. To study the co-citation network over time, a five-year overlapping time slice was used to divide our target set into 22 co-citation networks.

**Results**

Figure 1 shows the growth of articles and number of nodes in the co-citation networks over time. The graph indicates that the interests in DH took off in early 2000 and has continued to grow to this date. Notice also that there is a significant gap between the number of total articles and nodes in the co-citation network in our target set, indicating a large proportion of isolates in the whole network.

Figure 1. The growth of publications and size of the co-citation network over time.

We used author assigned keywords as the representation of subjects treated in the articles. The keywords thus extracted were then normalized manually. Gini index was then applied to measure the balance of the subjects covered in our target set over time. High Gini index of the keyword distribution signals the existence of few dominant keywords. Figure 2 shows a gradual rising of Gini indexing over time, which might be interpreted as a gradual consolidation of research efforts. Yet notice that even with the gently rising slope, the degree of concentration remains low at below .35, which suggests rather disperse research interests in DH.

Figure 2. Gini index of keyword distribution over time.

Figure 3 shows the long-term trend of cited articles remained in the giant component. A jump of the percentage of nodes in the giant component can be observed in the early 2000s; it then gradually levelled off. The dip in recent years might be more likely due to the citation window for recent publications than a sign of disintegration. Another two measures, average geodesic distance and clustering coefficient that often associate with the ‘small world’ phenomenon, also showed similar patterns. Figure 4 shows a rather short average distance between nodes in the giant component, hovering around 4. The cluster coefficient stayed steadily high at 60 percent, indicating dense local clustering, which, coupled with the short average distance, fit a typical small-world model.

Figure 3. Percent of nodes within the giant components.

Figure 4. Clustering coefficient.

Figure 5. Average geodesic distance.

Another way to assess the cohesion or integration of a network is through degree centralization, as it measures the degree to which network cohesion is hinged on particular focal importance (Borgatti et al., 2013). High degree centralization signals inequality of node importance. Figure 6 traces the trends of degree and betweenness centralization of nodes in the co-citation main component over time. Both degree and betweenness centralization leveled off gradually from their peaks in the mid-1990s, when the idea of DH was first introduced by a few seminal works. The degree concentration, however, has since declined gradually, which echoes our previous findings, suggesting the presence of many local clusters that represents a wide variety of specialized interests.

Figure 6. Degree and betweenness centrality of the giant components over time.

Among all the centrality measure, betweenness centrality is arguably the most closely related to network cohesion. Table 5 ranked the top 40 sources in our target set by their betweenness centrality, accompanied by their degree centrality. Of particular interest are the sources able to achieve high betweenness centrality despite relatively low degree centrality as they play a significant ‘bridging’ role in the network that contributes the most to the network cohesion.



Table 1. Top 40 sources with highest between centrality.

We visualized the source co-citation network using MDS where the distance between nodes signifies the similarity of their co-citation profiles (Figure 7). The size of the nodes represents betweenness centrality, while different colors group frequently co-cited sources using fraction algorithm.



Figure 7. MDS representation of source co-citation profiles. Stress value = 0.417.

To identify individual works that contribute most to the cohesion of the network, we performed the betweenness centrality on the article co-citation network. The works that have a particularly strong bridging character relative to their degrees were highlighted.



Table 2. To 40 works with highest betweenness centrality.

**Discussion and Conclusion**

In this study we applied social analytical methods to measure the degree of knowledge diversity and cohesion of the bibliographic elements extracted from the article co-citation network. The results show that the degree of knowledge diversity is high, as demonstrated by the evenness of the keyword distribution, as well as the presence of publication sources from diverse disciplines. It was shown that DH has gradually evolved into a diverse yet cohesive area of research, where, despite high degree of local research interests, the reachability remains high globally.   
 A longitudinal approach allows us to observe the evolving of the knowledge integration and the shifting of research topics in an area of research. There are obviously also several limitations to our study. First, even though we have tried to be inclusive in our selection to target set, some of the important works are inevitably missed, especially those published in non-English languages. Using Google Scholar as the source of co-citation data also entails inherent bias. Currently our interpretation of the visualization of the field is quite preliminary due to the lack of domain knowledge in DH. In the future, author co-citation network will also be constructed, which, with the help of expert inputs, will help us better interpret the sub-domains and their relationships in DH.

**References**

**Borgatti, S. P., Everett, M. G. and Johnson, J. C.** (2013). *Analyzing Social Networks*. SAGE.

**Porter, A. L., Cohen, A. S., Roessner, J. D. and Perreault, M.** (2007). Measuring

Researcher Interdisciplinarity. *Scientometrics,* **72**(1): 117–47.

**Rafols, I. and Meyer, M.** (2010). Diversity and Network Coherence as Indicators of Interdisciplinarity: Case Studies in Bionanoscience. *Scientometrics*, **82**(2): 263–87.