Visualizing DHQ (Digital Humanities Quarterly) Bibliography

Mapping Cultures in the Big Tent: Multidisciplinary Networks in the Digital Humanities Quarterly

Authorship

Dulce Maria de la Cruz
Freelance Data Analyst
Dulce.Maria.delaCruz@gmail.com

Kristin Lewis
AAAS S&T Policy Fellow
kristin.l.m.lewis@gmail.com

Teh-Hen Yu
IT Professional
tehhenyu@hotmail.com

Jake Kaupp
Engineering Education Researcher
Queen's University
Kingston, ON, Canada
jkaupp@gmail.com

Max Kemman
PhD Candidate
University of Luxembourg
Luxembourg
maxkemman@gmail.com

Table of Contents

- Authorship
- Table of Contents
- Description of visualization goal/need, hand-sketch of the envisioned visualization, and discussion why this project is important
  - Goal/Need
  - Hand-sketch
  - Dataset
  - Importance
- Discussion of related work
  - Visualization of Citation Networks
  - Academic Cultures
- Simple statistics of the data sets used, e.g., number of entities, major entity attributes, etc.
  - Co-author network:
  - Paper Citation network:
  - Document Co-citation network:
  - Citation Network:
- Data analysis/visualization (algorithms) applied and resulting visualizations
  - Co-author network:
  - Document Co-citation network:
  - Word Clouds associated with the components of the Co-citation network:
- Discussion of key insights gained from the analysis/visualization
  - Insights gained from the co-author network:
  - Insights gained from the co-citation network:
  - Insights gained from the word clouds associated with the co-citation network:
- What problems surfaced during validation and how does your redesign resolve them?
  - Validation and problems
    - Data
    - Analysis of DHQ authors
    - Analysis of DHQ topics
  - Redesign
- Discussion of challenges and opportunities
- References

Description of visualization goal/need, hand-sketch of the envisioned visualization, and discussion why this project is important

Goal/Need

The Digital Humanities Quarterly (DHQ) journal covers all aspects of digital media in the humanities, representing a meeting point between digital humanities research and the wider humanities community [1]. It is the publication from ACH (Association for Computers and Humanities), which is part of ADHO (the Alliance of Digital Humanities Organization) -- a global alliance with constituent members in EU (EADH), US (ACH), Canada (CSDH/SCHN), Japan (JADH), Australia (aaDH), and an international network (centerNet) with 196 Digital Humanity Centers globally. Articles published in DHQ involve authors of multiple countries, institutions and disciplines who work on several subjects and areas related to digital media research.
Under a recent grant from NEH (National Endowment for Humanities), DHQ has developed a centralized bibliography which supports the bibliographic referencing for the journal. The client is looking for visualizations that show:

1. how citations reflect differences in academic culture at the institutional and geographic level
2. the changes to that culture over time.
3. correlations between article topics (reflected in keywords) and citation patterns.

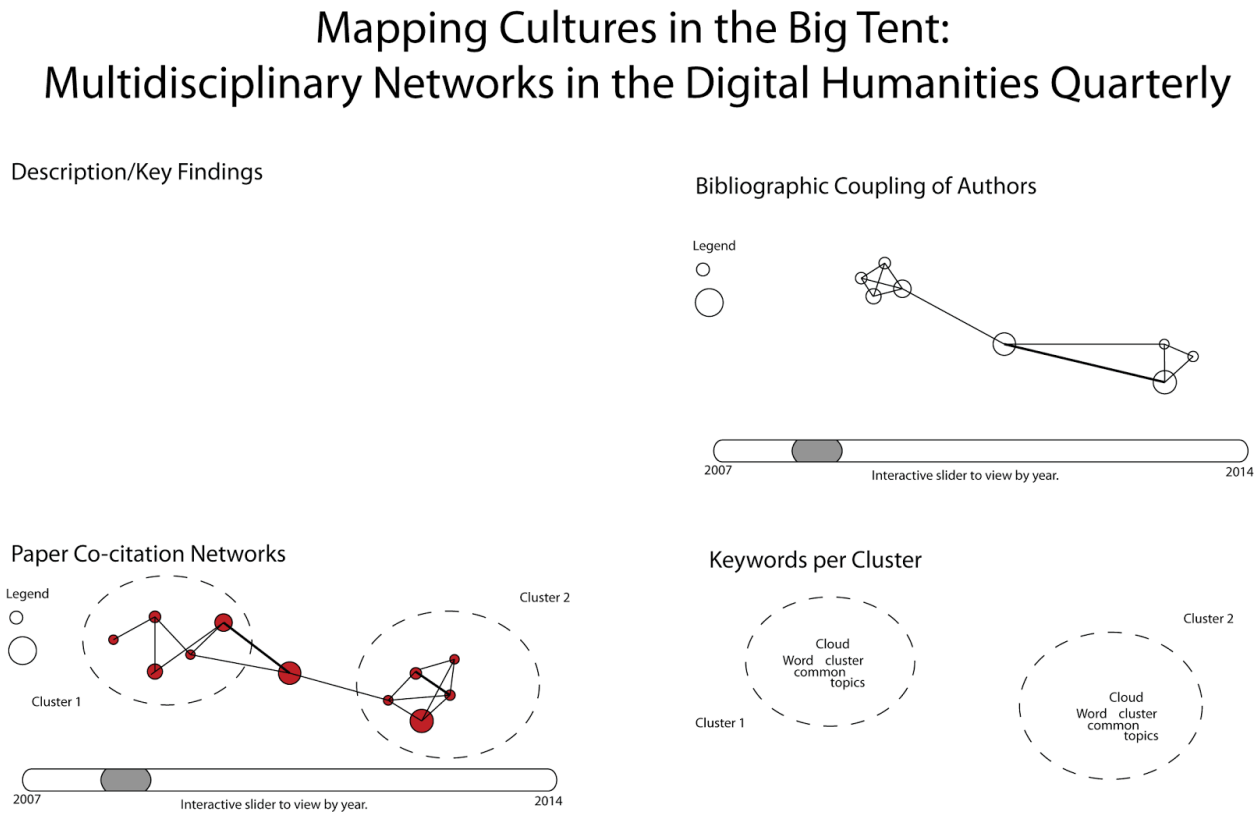
The identification of those subjects and areas and of their major contributors would be very important for any researcher involved or interested in digital media research. However, due to the collaborative, multidisciplinary nature of the digital media research, such identification becomes extremely difficult, if not impossible, to accomplish by merely analyzing the DHQ bibliographic database [2]. In such a case, visualization is the preferred approach.

Hand-sketch

Taking into account the available data for each DHQ article, we have planned a visualization that includes several components which may reveal clusters representing different cultures within this highly interdisciplinary research. First, we will create a bibliographic coupling of authors. The usual approach of bibliographic coupling is to cluster articles [3], we however deviate from this for two reasons: 1) we are interested in cultures, which is best represented by people rather than articles, and 2) due to the interdisciplinary nature of DH a single article with multiple authors can represent multiple cultures (e.g. a

collaboration between a historian and a computer scientist). Second, we will create a co-citation network of cited articles [4].

For each cluster, we will include a word cloud in order to illuminate correlations between articles topics (reflected in keywords) and citations patterns. An interactive visualization with a slider by year (or an automated animated visualization) will help reveal how these clusters change or grow over time. Potential visualizations are shown in the diagrams below:



The suggested visualization aims to allow DHQ's readers to gain insight into the citation networks that operate within the journal, showing them the major contributors and the subjects and areas involved in digital media research.

Dataset

Two tables were extracted from the Client dataset:

1. dhq_articles (178 records with unique columns: article_id, author, year, title, journal/conference/collection, abstract, reference IDs, isDHQ)
2. works_cited_in_dhq (3823 records; same columns as above table)

We initially encountered problems in the data, which required some pre-processing, pruning and reconciliation of the datasets. For instance:

- Incomplete, inconsistent and missing records in the dataset.
- There is no author institution affiliation available for both tables. That makes it quite difficult if we are to group the authors by institutions or GEOs.
- Keywords are only provided for a subset of the articles in the xml information.

The first issue will be easily remedied by contacting the client to verify the completeness of the provided data, and seeing if they can address any gaps; or by developing a consistent framework to exclude inconsistent or incomplete records. We will also cross reference the records with other information provided in the dataset, to ensure the completeness and accuracy of the dataset.

The second issue requires additional work, as the provided information is not sufficient to identify the department or disciplinary affiliation of the author. In addition, there is no geographic information or institutional affiliation in the dataset. In our group's initial efforts to address the second issue we were able to scrape the DHQ web site (http://www.digitalhumanities.org/dhq/index/author.html) to extract information about the affiliation of the first author for each DHQ article; this affiliation was then extended to provide geospatial information for the primary author. Affiliation data (at an institutional level) is also available in provided XML files, which requires additional processing. Continuing efforts will focus on identifying the departmental and disciplinary affiliation of the authors by search engines such as Google, Google Scholar, Web of Science and the scholarly database provided by the Cyberinfrastructure for Network Science center at Indiana University.

We also wish to contact the client, as the initial description of the project includes mention to additional items not included in the provided dataset, such as topic keywords for each paper. However, we can conduct topical analysis using provided abstract or title fields if keywords are not available.

Importance

Digital Humanities (DH) is a field of research difficult to define due to its heterogeneity[1]. With its inclusionary ambitions, DH is regularly referred to as a 'big tent' [5] encompassing scholars from a wide variety of disciplines such as history, literature, linguistics, but also disciplines such as human-computer interaction

[1] See e.g. http://whatisdigitalhumanities.com/ for a wide variety of definitions from different scholars

and computer science. This collaborative, multidisciplinary approach to digital media makes DH an interesting field, but also difficult to grasp. A question is to what extent the big tent of DH represents a single epistemic culture, or actually a variety of cultures [5, 6].

DHQ is arguably one of the largest journals specifically aimed at DH research. As such, it has attracted publications from across the big tent. To gain an understanding of the diversity of culture(s) in the DH, we are interested in how unique disciplinary cultures are represented in the DHQ. Considering cultures are self-referential systems, we might expect that scholars from a certain culture are more likely to cite scholars from their own discipline rather than from others [6]. As such, we expect citation behaviour to reflect disciplinary cultural norms. Therefore, visualizing and analysing the bibliographic data of DHQ not only gives insights into the specific bibliographies from DHQ, it might give insight into the way the different epistemic cultures in the DH big tent interact with one another, and how this interaction and collaboration impacts the networks over time.

Discussion of related work

Visualization of Citation Networks

Ever since Garfield's work on citation analysis [7], citations have been of interest to understand how scientists and scholars build upon one another's work. Generally, citations are a useful metric for understanding academia because they constitute the simplest relation between two publications [8, 9], representing that the authors have read and been influenced by another paper [10]. Citations are thus a useful way to understand the relations of a publication to previous work and its impact on subsequent research [11]. As such, understanding how knowledge diffuses through academic cultures can be approached by analysing citations. This was done manually in the past, but has become much more feasible and scalable thanks to computational analysis [12]. By showing citations as a graph, documents can be understood as part of a citation environment [9]. Clusters of citation environments may represent cultures of knowledge present in the journal. Therefore, we will cluster the citation data as described above in "Hand-sketch".

Many of these visualizations are build by scientists from Bibliometrics or Science and Technology Studies. However, there is an increasing interest in visualization also from DH scholars [13]. A DH project that aimed at humanities scholars that visualized bibliographies was RoSE (Research-oriented Social Environment) that

aimed to “storyboard” intellectual movement [14]. Likewise, our visualization will be aimed at humanities scholars.

Academic Cultures

Culture in higher education is both broadly used and difficult to define. It can pertain to many different social structures within and across an institution, providing many different units of analysis [15]. One of the most commonly studied units is the the departmental or disciplinary sub-culture, which can be represented equally across institution, country and geographic boundaries [16]. There is no universal definition and depiction of culture but it is typically comprised of common beliefs and perceptions that coalesce around similar values of norms of groups of individuals of common background, training and experience [17]. These values and norms comprising disciplinary cultures can be separated into, first; the epistemic (“ways of knowing and organizing knowledge”) cognitive aspect consists of the general topical area of expertise and established research methods and resources [12]. Second, the socio-organizational, hierarchical aspect which is predominantly defined by institutional structures and divisions, such as departments or faculties that comprise an institution [12].

With respect to our data, we can categorize or classify the topic of each paper according to a specific institution through textual analysis of abstracts or titles. Consistent with this approach, we can also identify the departmental affiliation of author through deeper research and web-scraping. This permits the classification of topic, author, institution and departmental affiliation, providing an approach to defining the cultural identity of dataset records. However, this approach is not without limitations and drawbacks with the most prominent being that departments can contain multiple disciplinary cultures [18]. This highlights the importance of focusing our classification along epistemic lines, which embody the definition set forth by Cetina:

> “*those sets of practices, arrangements and mechanisms bound together by necessity, affinity and historical coincidence which, in a given area of professional expertise, make up how we know what we know. Epistemic cultures are cultures of creating and warranting knowledge”* [19, p363].

By using this more malleable concept, we are able to use the publication history of an author alongside the departmental and institutional categories to produce a more holistic classification scheme which considers both the institutional disciplinary organization (the department) along with the knowledge-based classification of the discipline. We believe that this approach will permit the classification of different cultures within a journal as multidisciplinary and collaborative as the Digital Humanities Quarterly, even if different disciplinary cultures exist within a single department, according to citation behaviour.

Simple statistics of the data sets used, e.g., number of entities, major entity attributes, etc.

There are two entities provided by the client, as stated in the above section. We have dhq_articles and works_cited_in_dhq.

The attributes for both tables are as below:

- article id
- authors
- year
- title
- journal/conference/collection
- abstract
- cited references
- isDHQ

A great deal of effort was devoted to mining, cleaning and validating the client-provided dataset prior to analysis and visualization. There were missing records, inconsistent fields and potential mismatches or errors in referencing. To reconcile this, the DHQ and other scholarly databases were scraped and mined to correlate and verify all records, and to populate institutional affiliation, country of origin, and provide the future basis for disciplinary categorization. This required a concerted effort in a variety of tools (R, Sci2, OpenRefine) which resulted in the additional variables listed below. The following attributes were compiled and added to the original table of dhq_articles:

- cite me as
- times cited
- affiliation
- country
- count cited references

Co-author network:

We first explored authorship by extracting a co-author network, which is an undirected-unweighted network.

The Network Analysis Toolkit (NAT) generates the following results for this network:

Nodes: 270
Isolated nodes: 97
Node attributes present: label, number_of_authored_works, times_cited

Edges: 373
No self loops were discovered.
No parallel edges were discovered.

Edge attributes:
Did not detect any nonnumeric attributes.
Numeric attributes:

	min	max	mean
number_...	1	6	1.07507
weight	1	6	1.07507

This network seems to be valued.

Average degree: 2.763
This graph is not weakly connected.
There are 139 weakly connected components. (97 isolates)
The largest connected component consists of 43 nodes.
Did not calculate strong connectedness because this graph was not directed.

Density (disregarding weights): 0.0103
Additional Densities by Numeric Attribute

Paper Citation network:

To understand the citation network as a whole, we extracted the paper citation network and used the Network Analysis Toolkit:

Nodes: 4870
Isolated nodes: 21
Node attributes present: label, localCitationCount, inOriginalDataSet, globalCitationCount

Edges: 5330
There are: 5 self loops.
They are as follows:
The node:
Greenspan2011 1 Greenspan, Brian 0
has an edge with itself.
The node:
Siemens2009 1 Siemens, Ray | Leitch, Cara | Blake, Analisa | Armstrong, Karin | Willinsky, John 1
has an edge with itself.
The node:

Rosenbloom2012 1 Rosenbloom, Paul S. 0
has an edge with itself.
The node:
Causer2012 1 Causer, Tim | Wallace, Valerie 0
has an edge with itself.
The node:
Crymble2013 1 Crymble, Adam | Flanders, Julia 0
has an edge with itself.

No parallel edges were discovered.

Did not detect any edge attributes.
This network does not seem to be a valued network.

Average total degree: 2.1889
Average in degree: 1.0945
Average out degree: 1.0945
This graph is not weakly connected.
There are 40 weakly connected components. (21 isolates)
The largest connected component consists of 4563 nodes.
This graph is not strongly connected.
There are 4870 strongly connected components.
The largest strongly connected component consists of 1 nodes.

Did not calculate density due to the presence of self-loops.
Many algorithms will not function correctly with this graph.

Document Co-citation network:

We next extracted the document co-citation network for the DHQ articles dataset (as described in the next section). The Network Analysis Toolkit (NAT) provides the following results for the produced network:

Nodes: 27
Isolated nodes: 0
Node attributes present: label, localcitationcount, inoriginaldataset, globalcitationcount, cited

Edges: 56
No self loops were discovered.
No parallel edges were discovered.

Edge attributes:
Did not detect any nonnumeric attributes.
Numeric attributes:

	min	max	mean
weight	1	2	1.08929

This network seems to be valued.

Average degree: 4.1481
This graph is not weakly connected.
There are 3 weakly connected components. (0 isolates)
The largest connected component consists of 22 nodes.
Did not calculate strong connectedness because this graph was not directed.

Density (disregarding weights): 0.1595
Additional Densities by Numeric Attribute

Citation Network:

According to the provided dataset, there are **4870** included publications, of which **195** are articles from DHQ itself. To count the number of references, we run the NAT on the extracted paper citation network: there are **5330** references, with an average in-degree of 1.0945. We find the highest cited document is **kirschenbaum2008** ("Mechanisms: New Media and the Forensic Imagination"), cited **15** times. The DHQ article which cites the most references is **Christine Borgman**'s "The Digital Future is Now: A Call to Action for the Humanities" with **130 references**. The author with the most DHQ publications is **Julianne Nyhan**, with **7 publications**.

Data analysis/visualization (algorithms) applied and resulting visualizations

Co-author network:

To analyze the authors, we generated the co-author network in Sci2 with the following workflow.

1. We opened Sci2 and load the dhq articles data file, so we ran:

```
File -> Load -> dhq_articles_2007_2014.csv
```

2. We selected the previous csv file and ran:

```
Data Preparation -> Extract co-author network
```

3. We selected in the Data Manager of Sci2 the network resulting from the previous step (*Extracted co-author network*) and ran:

```
Analysis -> Networks -> Network Analysis Toolkit (NAT)
```

4. We did the following to add node degree attribute:

```
Analysis -> Networks -> Unweighted & Undirected -> Node Degree
```

5. We selected in the Data Manager of Sci2 the network resulting from the previous step

```
Preprocessing -> Networks -> Extract Top Nodes
```

with Nodes=150 as parameter.

We created a file by viewing the resulting network which contains top 150 authors with totaldegree info.

6. Then we visualize the previous network with GUESS algorithm

```
Visualize -> Networks -> GUESS
```

7. In GUESS window, we applied GEM layout and Bin Pack layout

8. Then we resize Linear from 1 to 20 with nodes attributes number_of_authored_work

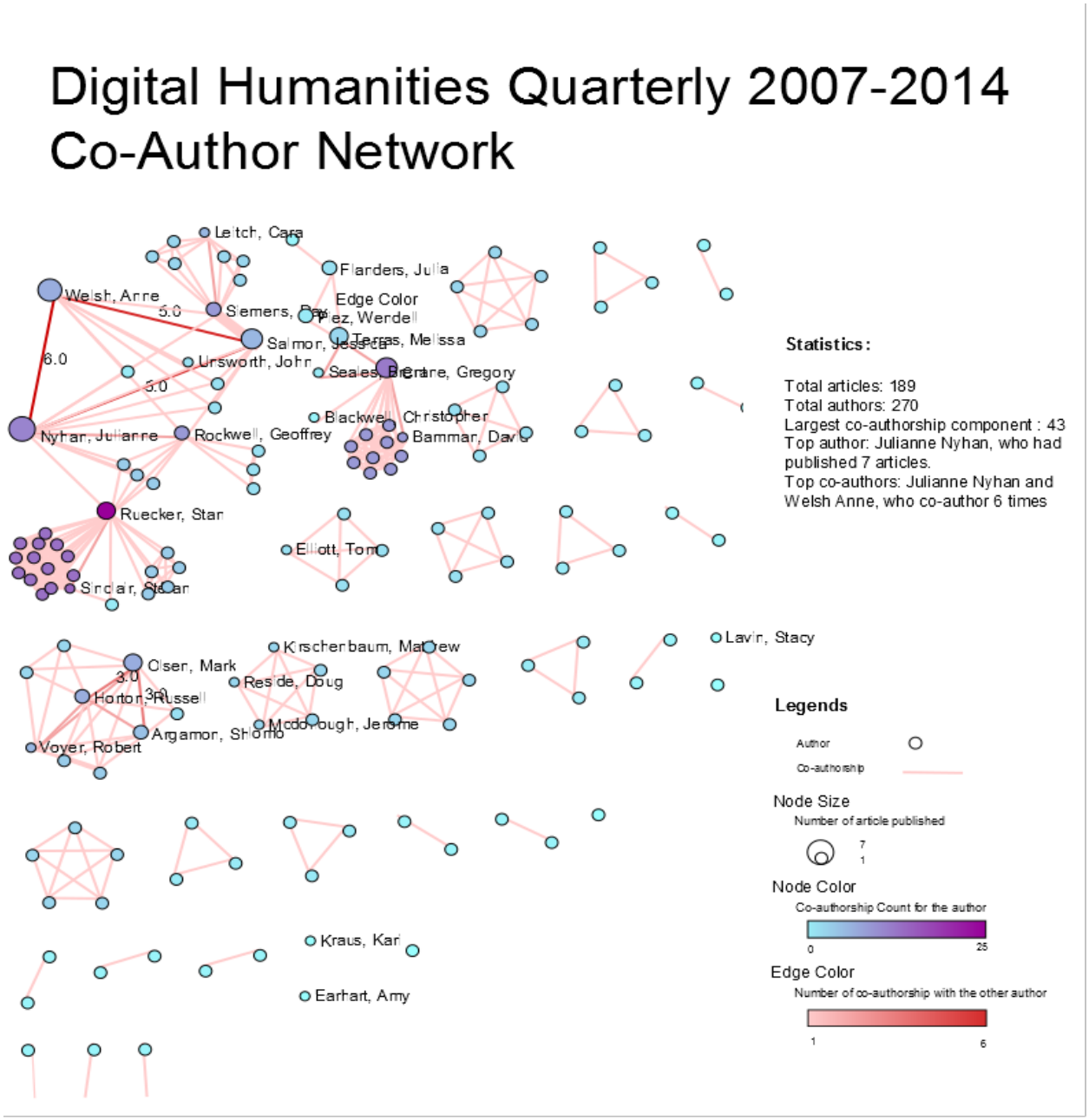
9. Then we colorize the nodes from light blue to purple for nodes attribute totaldegree

10. Then we colorize the edges from pink to red for edges attribute number_of_coauthored_work

11. Then we show label for the nodes based on attribute number_of_authored_work >=2

12. Then we show label for edges based on attribute number_of_coauthored_work >= 3

The resulting visualization is shown below:



Document Co-citation network:

The document co-citation network below is the result of the following workflow: [21]

1. We opened Sci2 and load the dhq articles data file, so we ran:

```
File -> Load -> dhq_articles_2007_2014.csv
```

2. We selected the previous csv file and ran:

```
Data Preparation -> Extract paper citation network
```

3. We selected in the Data Manager of Sci2 the network resulting from the previous step (*Extracted paper-citation network*) and ran:

```
Analysis -> Networks -> Network Analysis Toolkit (NAT)
```

4. The output of the previous step informed that self loops were discovered so, we selected the network *Extracted paper-citation network* and ran:

```
Preprocessing -> Networks -> Remove Self Loops
```

5. In order to extract an network that contains only the original DHQ papers, we selected the network resulting from the previous step (*Without self loops*) and ran:

```
Preprocessing -> Networks -> Extract Nodes Above or Below Value
```

With the following parameters:

- Extract from this number: -1.0
- Numeric Attribute: globalCitationCount

6. We selected the resulting file (*Nodes above -1.0 by globalcitationcount*) and ran:

```
Data Preparation -> Extract Document Co-Citation Network
```

7. We selected the network produced in the previous step (*Co-citation Similarity Network*) and ran:

```
Analysis -> Networks -> Network Analysis Toolkit (NAT)
```

The output of the previous step informed us that the there were 189 nodes, 56 edges and 162 isolates in the network.

8. To remove isolated nodes, we selected *Co-citation Similarity Network* in the Data Manager of Sci2 and ran:

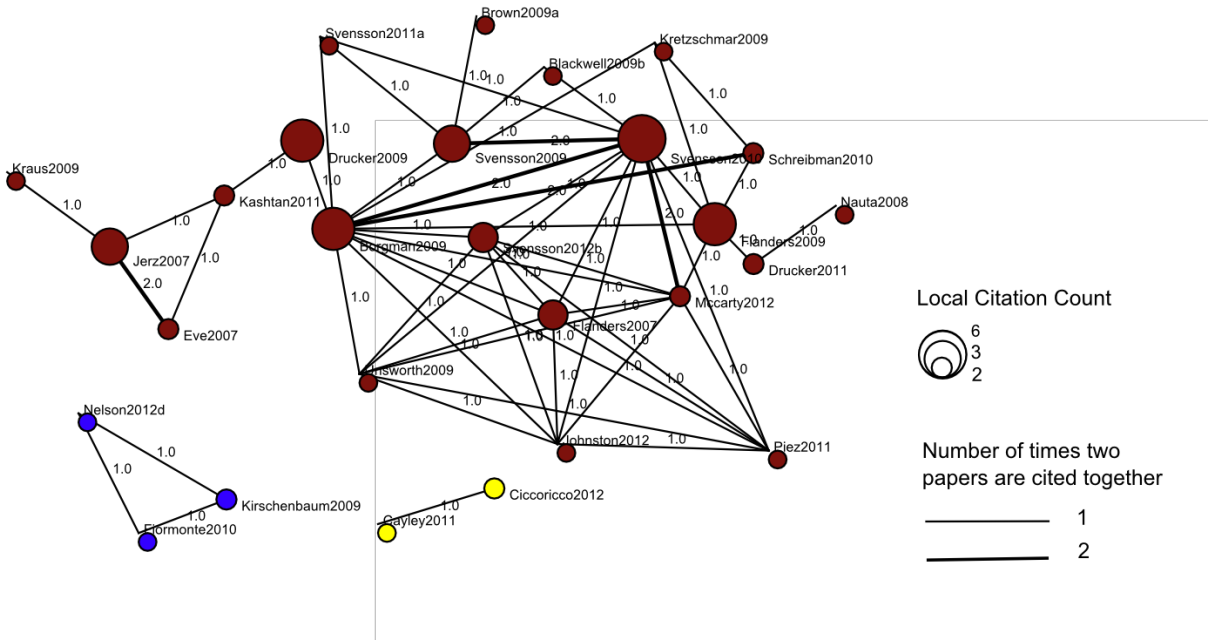
```
Preprocessing -> Networks -> Delete Isolates
```

9. We selected the network obtained in the previous step and visualize it with GUESS:

```
Visualization -> Networks -> GUESS
```

where the size of the nodes and the thickness of the edges were resized using the *localcitation* and *weight* parameters, respectively. Inkscape was used to add the title and legend to the network.

Digital Humanities Quarterly 2007 - 2014 Co-Citation Network



Word Clouds associated with the components of the Co-citation network:

The word clouds associated with the Co-Citation Network showed below is the result of the following workflow:

1. The article IDs of the nodes which belong to each component of the co-citation network were extracted and used to identify the articles associated with them.
2. A program in R language was written to produced the word cloud for each component of the co-citation network. This program used the abstracts of the articles associated with each component of the co-citation network, to extract a word cloud of the ten most common words in those abstracts. The procedure implemented by the R program involves steps like change the case of the text to lowercase, tokenization, stopword removal and identification of unique terms.
3. Inkscape was used to add the title and legend to the visualization of the word clouds.

Digital Humanities Quarterly 2007 - 2014 Word Clouds for the Co-Citation Network

Discussion of key insights gained from the analysis/visualization

The DHQ dataset contains data running from 2007-2014 with a total of 178 articles(from raw data). For years of 2013 and 2014, data were incomplete so we have to collect them manually from the digital journal website (http://www.digitalhumanities.org/dhq/). With all the data available from client's zip file and website, there are **195** articles that have been written by independent scholars and researchers of **148** institutions from **17** countries (Australia, Canada, Denmark, France, Germany, Ireland, Italy, Japan, Mexico, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, United Kingdom and the United States of America). Most authors of the DHQ papers are affiliated to institutions from the United States of America, Canada and United Kingdom.

Insights gained from the co-author network:

1. The largest co-author network component consists of 43 authors; which is about 16% of all authors (276 authors in all) who contributed to DHQ during this period. The second largest co-author network component consist of 18 authors.
2. The maximum number of authored works (articles) for single author is 7: Julianne Nyhan
3. The maximum co-authored articles for two authors are 6: by Welsh Anne and Julianne Nyhan
4. The average of the co-authorship per article is about 2.763
5. The number of articles with one author: 97; which is about 51% of the entire publications (189 articles in all) during this period.
6. The most collaborative author in this period: Ruecker, Stan; he co-authored 4 articles with 25 others.
7. The top article title with the most co-authors: "*Visualizing Theatrical Text: From Watching the Script to the Simulated Environment for Theatre (SET) "* with 14 co-authors.

Insights gained from the co-citation network:

1. The co-citation network extracted for DHQ documents published from 2007 to 2014 show three components.
2. The most cited DHQ document is Svensson, P. (2010). The landscape of digital humanities. *Digital Humanities.*
3. The document pairs that have been cited together most frequently are:
    - Borgman, C. L. (2009). The digital future is now: A call to action for the humanities. *Digital humanities quarterly*, *3*(4) and Schreibman, S., & Hanlon, A. (2010). Determining value for digital humanities tools: Report on a survey of tool developers. *Digital humanities quarterly*, *4*(2) .
    - Borgman, C. L. (2009). The digital future is now: A call to action for the humanities. *Digital humanities quarterly*, *3*(4) and Svensson, P. (2010). The landscape of digital humanities. *Digital Humanities quarterly*, *4*(1)
    - Svensson, P. (2010). The landscape of digital humanities. *Digital Humanities quarterly*, *4*(1) and McCarty, W., Nyhan, J., Welsh, A., & Salmon, J. (2012). Questioning, asking and enduring curiosity: an oral history conversation between Julianne Nyhan and Willard McCarty. *DHQ: Digital Humanities Quarterly*, 6(*3)*.
    - Svensson, P. (2010). The landscape of digital humanities. *Digital Humanities quarterly*, *4*(1) and Svensson, P. (2009). Humanities computing as digital humanities. *Digital Humanities Quarterly, 3*(3).

- Jerz, D. G. (2007). Somewhere Nearby is Colossal Cave: Examining Will Crowther's Original'Adventure'in Code and in Kentucky. *Digital Humanities Quarterly*, *1*(2) and Eve, E. (2007). All Hope Abandon: Biblical Text and Interactive Fiction. Digital Humanities Quarterly, 1(2).

Insights gained from the word clouds associated with the co-citation network:

1. The ten most common words associated to the the biggest component of the co-citation network (component with red nodes) sorted by decreasing frequency are:
   humanities, digital, article, visual, field, game, computing, new, also, project
2. The ten most common words associated to the the medium component of the co-citation network (component with blue nodes) sorted by decreasing frequency are:
   will, linking, systems, print, current, digital, encoding, nature, present, different
3. The ten most common words associated to the the smallest component of the co-citation network (component with yellow nodes) sorted by decreasing frequency are:
   writing, literary, reading, media, processes, part, found, google, close, digital

What problems surfaced during validation and how does your redesign resolve them?

Validation and problems

We have emailed the client contact (Professor Julia Flanders) to provide us the feedback from the write-up 1-8. She was very positive with our initial effort. As per our request, she provided us with the keywords for the DHQ articles from the draft metadata scheme they were developing.

Data

Some problems we encountered and fixed in the datafiles:

Missing articles

The client provided the article datafile with 178 articles. It was noticed that the client provided dataset was incomplete and contained both discrepancies and missing articles. To remedy this, the DHQ website was scraped for all up to date publication XML, excluding the single issue from 2015. This resulted in a complete dataset of 195 articles from the publication. There were still missing articles, perhaps due to misnumbering,

from the website XML (article id 186, 193). There was also a single article reprinted in two issues 2011:5.3 and 2012:6.3 that was reduced to a single record within its original and initial publication

Duplicate authors

The DHQ article XML generates duplicate authors, either due to inconsistent naming conventions or author error. These duplicate authors were address individually and unified to the version that was presented on the DHQ websites author listing.

1. Butts, Jimmy has duplicate Butts, J.J. **unified to Butts, Jimmy**
2. Jerz, Dennis has duplicate Jerz, Dennis G. **unified to Jerz, Dennis**
3. Kirschenbaum, Matthew has duplicate Kirschenbaum, Matthew G. **unified to Kirschenbaum, Matthew**
4. Terras, Melissa has duplicate Terras, Melissa M. **unified to Terras, Melissa M.**

After fixing duplicates, there are 276 unique authors in the dataset..

Affiliations

Some authors are affiliated with more than one institution. For example, article id 169, and 176 had authors having being affiliated with two institutions. There are also some concerns with naming conventions for institutions that have multiple satellite campuses. Both of these issues had potential impact for geocoding records, as multiple locations for a single node can be difficult to interpret.

Many of these issues were handled manually. Upon correcting for double-affiliations and inconsistencies , the data has **148** unique institutions, including independent scholars. Geocoding of these records were conducted using the google maps api, with each independent author being geocoded according to their institutional affiliation. Independent scholars were also geocoded, according to information collected from web-searches, personal blogs or networking websites. For analysis and visualization purposes each record in the dataset (article.id) was assigned the geospatial coordinates of the primary author.

Disciplinary Identification of Authors

In order to determine and visualize any cultural differences or patterns in the DHQ dataset, we needed to determine this information for each DHQ author. Determining disciplinary affiliation by a web-search would take a great deal of time, and would be subject to a great deal of interpretation, affecting reliability of the information. To remedy this, we selected a method used previously to track cross-discipline information use [22] with the goal of identifying disciplines by departmental affiliation. The affiliation information in the the article XML only contained institution or company affiliations, and sparse mention of departmental or

disciplinary affiliation. To fully collect this information, the the DHQ Article Title listing was scraped and manipulated to extract more mention of departmental affiliation. Any remaining gaps in departmental affiliation was addressed by manual web searching.

Upon collecting the data there was still many variations in departments to determine disciplinary specifications. In order to produce a more detailed list of disciplinary culture, departmental affiliation was manually mapped to web of science subject areas. This resulting mapping was applied to the dataset, with the discipline of primary author representing the unique culture for each particular article. This resulted in 24 unique cultures being represented by the articles in the dataset. If all primary and co-authors are included in the analysis there are 29 unique cultures are represented in the dataset.

Duplicated documents / Self loops

We found that one of the DHQ documents - an introduction that was wrote by Mauro Carassai and Elisabet Takehana - was included in two different issues. Specifically, that introduction was published in the issues 2011 5.3 (http://www.digitalhumanities.org/dhq/vol/5/3/index.html) and 2012 6.3 (http://www.digitalhumanities.org/dhq/vol/6/2/index.html). To avoid duplicated records, only one those two publications - the first of them that appears in the issue 2011 5.3 - was included in the dataset.

We also had several of self loops in citation when a DHQ article and its reference ID were the same. The self loops were discovered during visualizing citation network. After analysing the documents (articles and references) involved in those self loops, we realize that they were different documents that shared the same author and year of publication. Due that, they also shared the same ID. In order to fix this problem, all the dataset was checked and we renamed those references that shared the same ID with DHQ articles despite they were different documents, by adding a letter (a, b, c,...) at the end of their original ID according with the rules used by DHQ and taking into account how these documents were referenced in other DHQ papers if so. As a result of that fixing work, the times cited value of the DHQ documents affected by this problem were updated.

Solutions to Data Problems

The raw dataset has been through more detailed examination. After validations, data mining/scraping, data processing with custom programs coding in R language and a lot of manual work, we have come up with a master dataset with additional info added (country, geocode, discipline, affiliations including departments info, and community, plus the keywords provided by client). To provide sufficient resolution, and categorical variables, for visualizations an author look-up table was created which contained the additional information

outlined above but for each separate author for each article ID. Both the master datafile and the author lookup table are our primary sources of data to load for visualization and analysis.

Analysis of DHQ authors

In the Hand-Sketch section we proposed to investigate cultures by bibliographic coupling of authors of DHQ articles. During our trials, we discovered that while this process is feasible in Sci2, focusing on authors introduces a serious bias in the analysis. Since co-authors reference the same articles, the relation between co-authors is stronger than between authors of separate articles referencing the same articles but also referencing other articles. Although this in itself could be argued not to be a problem, as co-authors perhaps should indeed be closer to one another than other authors, the problem is that co-authors of articles with many references are also closer to each other than co-authors of articles with few references.

A second issue with bibliographic coupling of authors is in the analysis of communities. We performed SLM community detection (weighted & undirected, default settings) on the network to create communities in the network. To check whether these communities provide additional insights with respect to cultures, we did a correspondence analysis in SPSS. From this we found that the identified SLM communities are strongly related to country (p<0.001, Cramer's V=.737), Web of Science subject area (p<0.001, V=.559), and year of publication (p<0.001, V=.763).

This suggests that the communities identified via SLM community detection of bibliographic coupled authors might introduce bias to our results, yet not give more insights than a co-author network with information regarding the author's country and WOS subject area. As a result, we have opted to create a co-author network to investigate the communities of authors of DHQ.

Analysis of DHQ topics

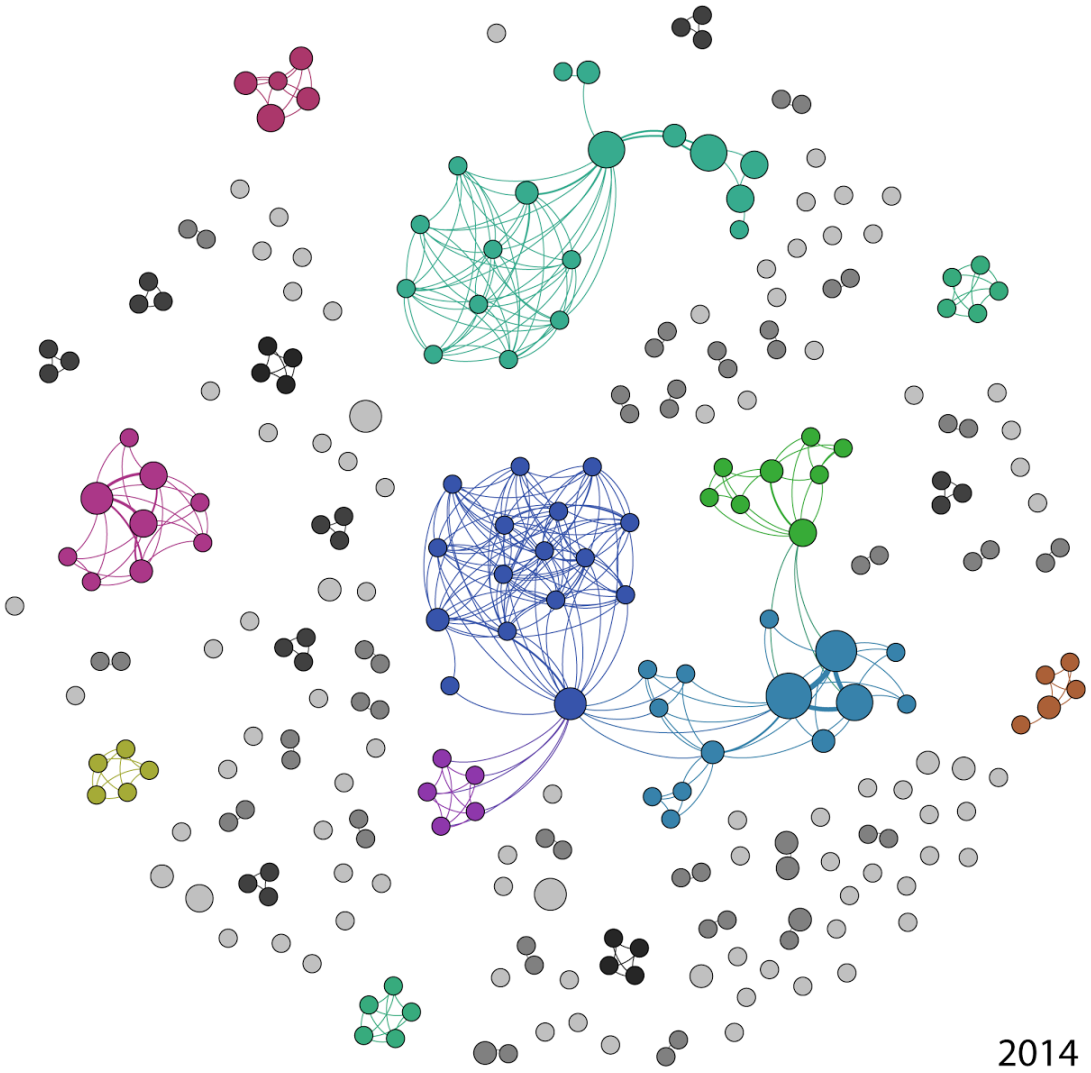
In the Hand-Sketch section, we proposed to extract word clouds for each cluster of the co-citation network of cited articles to illuminate correlations between articles topics (reflected by keywords) and citations patterns as required by the client. Despite the client contact (Professor Julia Flanders) providing us the keywords for the DHQ articles,we have elected to extract the word clouds from the titles and abstract of the articles. We believe that the keywords derived from the titles and abstracts may provide more insight into the values and ideas represented in individual cultures, than the thematic classification provided by the client.

Redesign

Following the problems identified during validation, we propose a redesign focusing on the following aspects:

1. A co-author network showing a cumulative network of authors collaborating
2. A bibliographic coupling network showing a cumulative network of DHQ articles with similar references, identifying communities through SLM community detection, and showing word clouds per community based on abstracts.

Co-Authorship of Articles in Digital Humanities Quarterly: 2007-2014



Number of DHQ-Published Works

7

1

Number of Co-Authored Papers

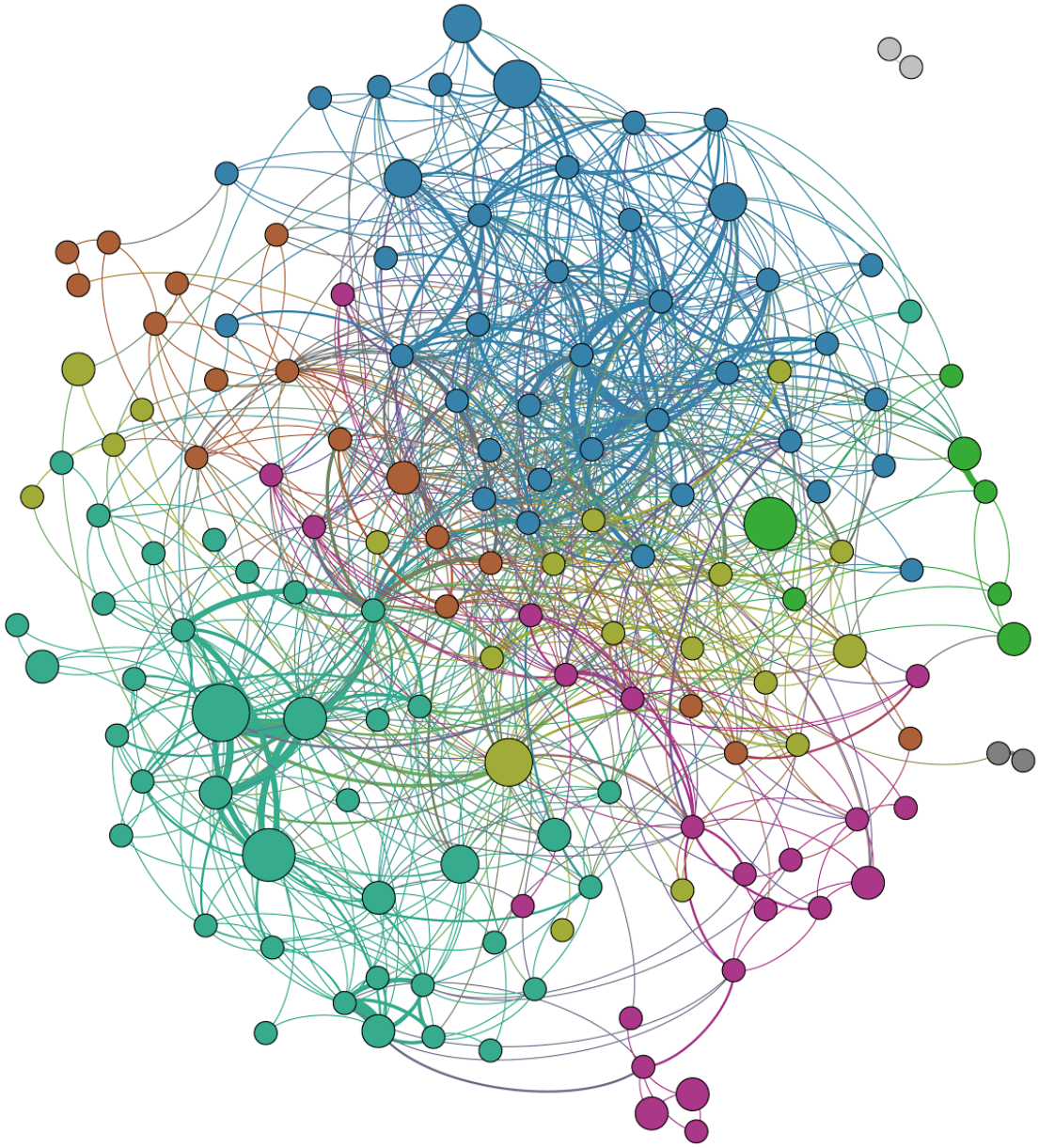
6

1

Fill colors are determined through SLM community detection on the entire network. Communities with fewer than 5 nodes are shown in shades of gray.

A video, and cumulative time slice images are available in Canvas.

Bibliographic Coupling in Digital Humanities Quarterly: 2007-2014



Number of Times Cited by Other DHQ Papers

6

0

Number of Papers Jointly Cited

17

1

Fill colors are determined through SLM community detection.

Bibliographic Coupling in Digital Humanities Quarterly: Word Clouds of Top Four Clusters

Discussion of challenges and opportunities

Special characters

Given the global nature of digital humanities, the data contained special characters, e.g. in author names, or in titles and abstracts. Sci2 renders several of these characters incorrectly, and during the time periods analyzed, some authors with special characters had their names rendered in multiple ways, resulting in multiple nodes for the same author in the co-author network. Moreover, for the word clouds special characters were a difficulty since words including special characters could not be normalized. This forced us to manually change special characters to normal ones, or removal words from the word cloud. Although this was feasible for our current analysis, this is not a scalable solution.

Answering the client's questions

One of the major challenges of this project is whether the dataset we have allows us to satisfyingly provide an answer to the client's questions. To gain a full understanding of the academic cultures present in DHQ is a complex undertaking, and we see several sources of information that might prove needed to fully address this question that we do not currently have available. First, we have limited information about the authors of DHQ papers, since we only know the papers they wrote for DHQ, their institution, country, and Web of Science area subject. This information is not very granular, nor very extensive, and might be deemed too shallow to provide insight into author's cultures. Second, for cited documents we have information regarding author, year, title, and journal. However, to gain a deep insight into how academic cultures influence the papers published in DHQ this too might prove limited, since we know little about the authors influencing the DHQ papers, as well as about the contents of the documents cited. Although we believe our visualizations provide insight into how authors and papers relate to one another within DHQ, it is difficult to say whether we can describe this as insight into cultures.

Future Directions

We see several opportunities for extending this work in the future:

1. Developing an authorial bibliographic coupling network less biased towards the number of references from co-authored papers.
2. Including DHQ keywords for each bibliographic cluster and comparing them with the word clouds generated from the abstracts.
3. Including additional attributes such as country, affiliations, and Web of Science subject areas to the provided visualizations.
4. Developing an interactive visualization, perhaps using Tableau public, to allow users to explore the data more fully with time sliders and mouse-overs with additional information.
5. Creating document co-citation networks that include non-DHQ articles to understand how those play a role in the culture clusters within DHQ.

References

[1] Digital Humanities Quarterly (n.d.). *About DHQ*. Retrieved from http://www.digitalhumanities.org/dhq/about/about.html

[2] Borner, K., Chen, C. M., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review Of Information Science And Technology*, *37*, 179–255. doi:10.1002/aris.1440370106

[3] Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation*, *14*(1), 10-25.
[4] Weingart, S. (2013) Networks Demystified 4: Co-Citation Analysis. *the scottbot irregular.* http://www.scottbot.net/HIAL/?p=38272
[5] Svensson, Patrik. (2012) Beyond the big tent. *Debates in the Digital Humanities*, 36-49.
[6] Knorr Cetina, K. (2007). Culture in Global Knowledge Societies: Knowledge Cultures and Epistemic Cultures. *The Blackwell Companion to the Sociology of Culture*, *32*(4), 361–375. doi:10.1002/9780470996744.ch5
[7] Garfield, E. (1955). Citation indexes for science. A new dimension in documentation through association of ideas. *Science*, *122*(3159), 108–111.
[8] Brüggemann-Klein, A., Klein, R., & Landgraf, B. (1999). BibRelEx: exploring bibliographic databases by visualization of annotated content-based relations. *2000 IEEE Conference on Information Visualization. An International Conference on Computer Visualization and Graphics*, *5*(11), 19–24. doi:10.1109/IV.2000.859731
[9] Lu, W., Janssen, J., Milios, E., Japkowicz, N., & Zhang, Y. (2006). Node similarity in the citation graph. *Knowledge and Information Systems*, *11*(1), 105–129. doi:10.1007/s10115-006-0023-9
[10] Elmqvist, N., & Tsigas, P. (2007). CiteWiz: a tool for the visualization of scientific citation networks. *Information Visualization*, *6*(3), 215–232. doi:10.1057/palgrave.ivs.9500156
[11] Schäfer, U., & Kasterka, U. (2010). Scientific authoring support: a tool to navigate in typed citation graphs. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids* (pp. 7–14). Association for Computational Linguistics. Retrieved from http://portal.acm.org/citation.cfm?id=1860659
[12] Becher, T., & Parry, S. (2005). The Endurance of the Disciplines. In I. Bleiklie & M. Henkel (Eds.), *Governing Knowledge* (Vol. 9, pp. 133–144). Springer. doi:10.1007/1-4020-3504-7
[13] Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly*, *5*(1), 1–21. Retrieved from http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html
[14] Liu, A., et al. (2012) Friending the Humanities Knowledge Base: Exploring Bibliography as Social Network in RoSE. *White paper for the NEH Office of Digital Humanities.* Retrieved from https://rosedocumentation.files.wordpress.com/2012/07/rose-white-paper-as-submitted-to-neh.pdf
[15] Tierney, W. (1988). Organizational Culture in Higher Education: Defining the Essentials. *The Journal of Higher Education*, *59*(1), 2–21. doi:10.2307/1981868
[16] Tierney, G., & William, J. E. (2011). Culture Organizational in Higher Education. *The Journal of Higher Education, 59*(1), 2–21. Retrieved from http://www.jstor.org/stable/1981868 .
[17] Martin, Joanne. *Cultures in Organizations: Three Perspectives*. Oxford University Press, 1992.
[18] Tuunainen, J. (2005). When disciplinary worlds collide: The organizational ecology of disciplines in university department. *Symbolic Interaction*, *28*(2), 205–228. doi:10.1525/si.2005.28.2.205
[19] Knorr Cetina, K. (2007). Culture in Global Knowledge Societies: Knowledge Cultures and Epistemic Cultures. *The Blackwell Companion to the Sociology of Culture*, *32*(4), 361–375. doi:10.1002/9780470996744.ch5
[20] Herr, Bruce W., Huang, Weixia, Penumarthy, Shashikant, Börner, Katy . (2007) Designing Highly Flexible and Usable Cyberinfrastructures for Convergence. In William S. Bainbridge and Mihail C. Roco (Eds.) Progress in Convergence – Technologies for Human Wellbeing. Annals of the New York Academy of Sciences, Boston, MA, volume 1093, pp. 161-179. CI Shell Manual: Extract Document Co-Citation Network, retrieved from http://wiki.cns.iu.edu/display/CISHELL/Extract+Document+Co-Citation+Network April 6, 2015.
[21] Börner, Katy, et al. (updated 2013). 5.2.4 Mapping Scientometrics (ISI Data). Science of Science (Sci2) Tool Manual v1.1 beta. Retrieved from http://wiki.cns.iu.edu/pages/viewpage.action?pageId=2785308 April 6, 2015.

[22] Ortega, L., & Antell, K. (2006). Tracking Cross-Disciplinary Information Use by Author Affiliation: Demonstration of a Method. *College & Research Libraries*, *67*(5), 446–462. Retrieved from http://crl.acrl.org/content/67/5/446.short