STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

A multilayer exploration of the cognitive structure of publications in history[1]

Giovanni Colavizza*, Massimo Franceschet**

* *gcolavizza@turing.ac.uk*
The Alan Turing Institute, British Library, 96 Euston Road NW1 2DB, London UK.

** *massimo.franceschet@uniud.it*
Department of Mathematics, Computer Science and Physics, University of Udine. Via delle Scienze 206, 33100 Udine IT.

Introduction

Citation networks among journal articles are perhaps the most common object of investigation in bibliometrics. For example, citation networks are widely used for science mapping as a way to explore the cognitive structure of scientific fields (Börner 2010). Within this framework, the disciplines traditionally part of the humanities fare differently (Colavizza 2017b). Their main trait being the interplay of a broader array of publication typologies – monographs, edited volumes, journal articles – with a richer set of cited objects, including primary evidence. Consequently, when considered from a science mapping perspective, a community, field or specialism in the humanities might be represented as a multilayer network.

We consider here a specialism in history, the history of Venice, and represent it using a set of publications including both books (edited and monographs) and journal articles. This set of publications is interconnected using three similarity measures: bibliographic coupling over references to books, bibliographic coupling over references to primary sources and textual similarity. The result is a multi-relation network with three distinct dimensions (that we will call layers), one per similarity measure, connecting the same publications. Given this representation, we proceed to analyse the different communities emerging from the three layers, to qualify them and consider to what extent they overlap or instead provide for orthogonal conceptual spaces.

Methods

We start by describing the network representation we use. Take $B = (V, E, w)$, the bibliographic coupling network of publications, where $w : E \rightarrow \mathbb{R}^+$ is a function mapping each edge to a positive weight expressing the similarity between two nodes. W is the weighted, symmetric adjacency matrix representation of such network, where $W_{i,j} = W_{j,i} = w(e_{i,j})$ if there exist an edge between vertices i and j, 0 otherwise. The function w can be defined in a variety of ways, beyond the traditional reference overlap similarity. We consider three

[1] This work was in part supported by the Swiss National Fund with grants 205121_159961 and P1ELP2_168489.

definitions here: reference overlap (traditional bibliographic coupling, only considering references to books), reference overlap over primary sources and textual similarity.

For **traditional reference overlap**, we consider the cosine similarity over the references that two publications i and j have in common, and use this value to weight the edge connecting them in B:

$$w(e_{i,j}) = \frac{R_{i,j}}{\sqrt{R_i}\sqrt{R_j}} \qquad (1)$$

Where $R_{i,j}$ is the number of references in common between i and j, R_i the number of references of i and similarly for j.

With respect to the **reference overlap for primary sources**, we first need to discuss what does it mean to refer to a collection of documents in an archive. Archives are typically organized hierarchically and contain individual collections of documents with an internal tree structure, as shown in Figure 1. A publication can make reference to any level of the hierarchy of a collection; a hierarchical organization is a property of primary sources in other domains too, e.g. Classics. We therefore need a way to first, calculate the reference overlap similarity of two publications with respect to a given collection (tree), and then aggregate its score over the ensemble of cited trees (forest). Consider first a similarity score $s(a, b)$ where a and b are individual references made by two distinct publications:

$$s(a, b) = 2^{-d}$$

Where d is the distance in number of steps between a and b in a tree of the archive. If a and b are in different trees $s(a, b) = 0$ and if $a = b$, then $s(a, b) = 1$.

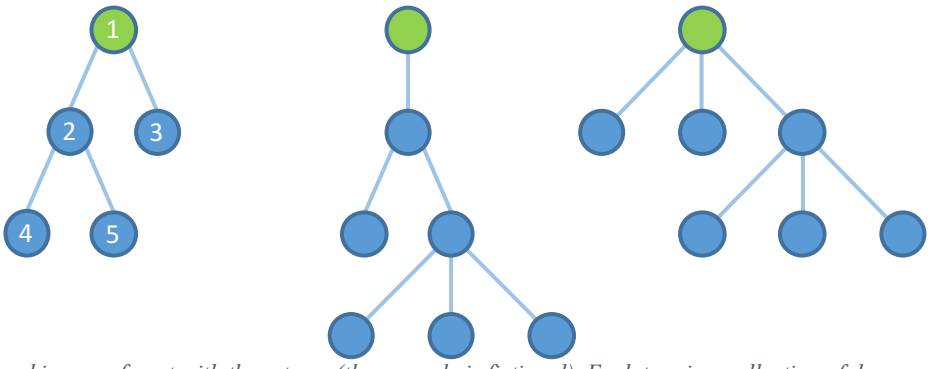

Figure 1: an archive as a forest with three trees (the example is fictional). Each tree is a collection of documents organized hierarchically, with a root (in green). In archival terminology, the root may be referred to as a record group and blue nodes as series or sub-series of documents. The similarity score s behaves as follows: $s(1,1) = 1$; $s(1,2) = 0.5$; $s(4,5) = 0.25$.

In order to equally weight every distinct collection when establishing a similarity between two publications, we proceed as follows. Given two publications i and j, R is the set of distinct document collections the two publications refer to. For every collection r, we first calculate the average similarity $avg(r, i, j)$ among all possible pairs of references of i and j which refer to any node in the tree of r, using the similarity s defined above. We then take the mean of the scores of every tree, therefore:

$$w(e_{i,j}) = \frac{1}{|R|} \sum_r avg(r, i, j) \qquad (2)$$

We base the **textual similarity** among two papers on the BM25 measure, adopted to rank documents for the purpose of information retrieval and document clustering. This measure has already been applied to assess the textual similarity of scientific publications (e.g. Boyack et al. 2011). Given a publication i and another publication j, the BM25 similarity $s(i,j)$ is calculated as:

$$s(i,j) = \sum_{z=1}^{n} IDF_z \frac{n_z(k_1+1)}{n_z + k_1\left(1 - b + b\frac{|D|}{|\overline{D}|}\right)}$$

where n denotes the number of unique tokens in i, n_z equals the frequency of token z in publication j, and $n_z = 0$ for tokens that are in i but not in j. k_1 and b have been set to the commonly used values of 2 and 0.75 respectively. $|D|$ denotes the length of document j, in number of tokens. $|\overline{D}|$ denotes the average length of all documents in the dataset. The IDF value for every unique token z in the dataset is calculated as:

$$IDF_z = \log\left(\frac{N - p_z + 0.5}{p_z + 0.5}\right)$$

where N denotes the total number of publications in the dataset and p_z denotes the number of publications containing token z. IDF scores strictly below zero are discarded to filter out very commonly occurring tokens. BM25 is not a symmetric measure. We obtain a symmetric measure for the similarity of documents i and j as follows:

$$w(e_{i,j}) = \frac{s(i,j)+s(j,i)}{2} \qquad (3)$$

All similarity scores are within or normalized to range between 0 and 1.

Dataset

Our dataset is custom made and was extracted from library collections as part of a process of digitization and indexation of the scholarly literature on the history of Venice, following an approach detailed elsewhere (Colavizza at al. 2017). We consider here a corpus of 2863 publications, 1648 books and 1215 journal articles, citing at least one book and primary source each. This corpus is mainly composed of publications in Italian, fewer in English, French, German and other languages, including Latin (for edition of sources). The chronological coverage favors recent decades, with 96% of the books and the majority of articles published after 1945.

After digitization, all publication contents were extracted with commercial optical character recognition tools (ABBYY Fine Reader). The extraction of references from the full contents, thus including footnotes, was then carried out using supervised approaches (Colavizza & Romanello 2017, Colavizza at al. 2017). Most crucially, the extraction of citations from references (disambiguation) was done differently for citations to books and primary sources. With respect to books, the external resource of the Italian National Catalog was used to perform a lookup, with a precision evaluation score of 0.784 and recall of 0.904 (Colavizza at al. 2017). The limitation of this approach is that for a book to be cited, it needs to appear within the National Italian Catalogue. With respect to primary sources, we consider only references made to documentation stored at the Archive of Venice, for the reason that this archive is the main reference for historical studies on the city's past. We used a set of 21,031 manually

disambiguated references to documents at the Archive to train a system of classifiers and disambiguate the rest, making reference to the external resource of the information system of the Archive. The classifiers, evaluated against 41 randomly picked articles as validation dataset (691 references), have a precision of 0.985 and a recall of 0.912. It must be noted that the citation space of the two set of sources is significantly different in size. Citations to documents at the archive refer in practice to collections of documents, as indexed by the information system of the Archive, whilst citations to books refer to individual editions of works. As a consequence, the whole corpus cites 588 distinct primary sources and 67,848 distinct books, or secondary ones.

We construct three weighted networks from the dataset, as detailed above. For the text similarity network, we consider the first 1000 words of each publication, after lowercasing and the removal of punctuation. In Table 1 we provide an overview of the network layers *book* (weighted network of co-references to books, using Eq. 1), *source* (weighted network of co-references to primary sources, Eq. 2), and *text* (weighted network of text similarity, Eq. 3). In all networks, we included only edges with a weight that is larger than the median weight of all edges. It is worth nothing that the book network has the lowest density and transitivity, consequently faring higher in modularity (calculated on a run using a fast greedy approach, see below). Conversely, the primary sources network has high density as the space of citable sources is considerably smaller in this case and their concentration higher. It is also not surprisingly that the text network shows the highest density and lowest modularity, as it is likely most publications share a similar vocabulary to a considerable degree.

Table 1: An overview of the three networks, all have the same number of nodes (2863).

Network / Measure	Edges	Density	Transitivity	Modularity
1- Book	253,820	0.06	0.34	0.28
2- Source	1,164,808	0.28	0.60	0.19
3- Text	2,048,477	0.50	0.71	0.10

Results

For each layer, we computed the community structure via greedy optimization of modularity (the `cluster_fast_greedy` implementation of the package igraph in R v1.2.0)[2]. This method starts out with each vertex in the network in a one-vertex group of its own, then successively amalgamates groups in pairs, choosing at each step the pair whose amalgamation gives the largest increase in modularity, or the smallest decrease if no choice gives an increase. The community structure with the highest value of modularity is given as output. Table 2 contains information on the largest communities for each layer (four communities for the *book* layer, three for *source* layer, and two for *text* layer). We show the size (number of nodes) of the community, the average degree and average strength (weighted degree) of nodes in the community.

Table 2: Largest communities for every layer of the network.

Community	Size	Degree (avg.)	Strength (avg.)
Book1	795	182	9
Book2	760	255	13
Book3	674	188	8

[2] Experiments with different approaches including Louvain, leading eigenvalue and walktrap, yielded comparable results. We leave as future work to further investigate community detection methods for the case at hand.

Book4	299	100	5
Source1	1226	1139	135
Source2	1157	676	86
Source3	440	336	46
Text1	1599	1241	909
Text2	1259	1678	1201

We use the Jaccard index to investigate the overlap of communities in different layers. The Jaccard index is a statistic used for comparing the similarity and diversity of sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

The index runs from 0 to 1, with values close to 0 indicating low similarity and values close to 1 corresponding to large similarity. For each pair of the largest communities in different layers, we compute the Jaccard index of the corresponding node sets. We can represent all these coefficients with a *tripartite* graph, where the three node types of the graph correspond to the three layers (book, source and text), while the edges run from nodes of different type and are weighted with the corresponding Jaccard index, as shown in Table 3.

Table 3: The weighted adjacency matrix of the tripartite graph of the Jaccard index across the largest communities of the three layers.

	Book1	**Book2**	**Book3**	**Book4**	**Source1**	**Source2**	**Source3**	**Text1**	**Text2**
Book1	–	–	–	–	0.23	0.17	0.12	0.25	0.18
Book2		–	–	–	0.22	0.21	0.06	0.22	0.20
Book3			–	–	0.19	0.14	0.13	0.20	0.18
Book4				–	0.08	0.08	0.10	0.08	0.11
Source1					–	–	–	0.34	0.26
Source2						–	–	0.30	0.28
Source3							–	0.13	0.14
Text1								–	–
Text2									–

The community overlap of a tripartite graph can be effectively visualized using a hive plot, shown in Figure 2. In hive plots, nodes are assigned to one of three (or more) axes. In our case, the axes correspond to the layers: book in red (top), source in green (right) and text in blue (left). Nodes are ordered on an axis based on some properties; in our case the nodes (communities) are sorted by increasing size from the center to the periphery of the hive. Edges are drawn as Bezier curves, which can be annotated to communicate additional information. In our case, we use the transparency of edges to represent the Jaccard index (hence, darker edges correspond to pairs of communities with higher similarity).

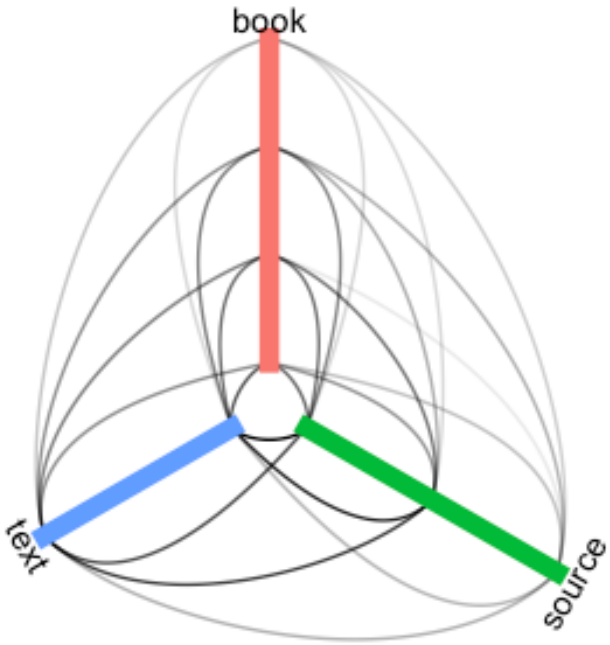

Figure 2: Hive plot of the Jaccard index across the largest communities of the three layers. The communities are ordered from largest to smallest, center to periphery.

In general, overlaps between communities are small, with a maximum of 0.34 for communities Source1 and Text1. This is a hint that the three layers – book source text – express largely orthogonal community structures and hence capture different types of similarity in our dataset. On average, communities in source and text layers have the highest similarity (0.24), while communities in book and text layers have intermediate similarity (0.18), and communities in book and source layers have the lowest similarity (0.14).

We proceed next to qualify the results of community detection using three different methods: I) expert judgement (by one of the authors), II) considering the most cited primary sources by the publications within each community and III) considering the aggregated topic distribution of each community of publications, calculated from a topic model trained on the whole corpus. Considering the former method first, the manual inspection of publication titles allowed to qualify the *book* communities only, in the following way:

1. Book1: Middle ages (history)
2. Book2: Early modern period (history)
3. Book3: History of art and architecture
4. Book4: Late early modern and modern period (history)

The reference overlap layer thus represents a coarse disciplinary organization (history and history of art and architecture), and a chronological subdivision of the former in three broad periods (up until the 15th century, from the 16th to the 18th and from the fall of the Republic in 1797 to the contemporary period). These results are in agreement with previous work (Colavizza 2017a). The *source* and *text* layers instead were too mixed to be qualified in a similar way.

Considering the second method, thus qualifying communities using their most cited primary sources, we have the following results for the *book* layer:

1. Book1: the most cited sources include the records of the most important public institutions of the Republic, plus the notary series (acts and wills; 8406 references to primary sources are given from this community).
2. Book2: the most cited sources include the same records as above, plus novel ones not existing for earlier periods. This community presents the largest variety of sources in use (9546 references).
3. Book3: the focus shifts to a rising importance of the notary series and of government records which allows to track the activities of artists, including the records of churches and confraternities (6031 references).
4. Book4: this community makes a wide use of records produced by the new governments existing after the fall of the Republic, including the Austrian dominions (2313 references).

With respect to the *source* layer instead:

1. Source1: this community makes an extensive use of primary sources and consider Venice directly as a topic (21,811 references).
2. Source2: here we instead find a marginal use of sources as Venice is considered within a broader topic (5126 references).
3. Source3: mainly gathers records which are relevant for modern history (19th century, 1322 references).

Once again, the *text* layer is hard to qualify using references to primary sources, as both its largest communities roughly use the same.

With respect to this last method, we use Latent Dirichlet Allocation (LDA, Blei et al. 2003) to train a model of 20 topics, using the first 25 pages of textual contents for every publication. We set-up a spaCy (v2.0) bag of words pipeline containing the following preprocessing steps: lowercase, extract and add named entities, filter to keep only alphanumeric tokens (no punctuation and numbers) and remove stopwords (for Italian, English, Spanish, German and French), lemmatize, remove tokens of just one character, add bigrams appearing 15 or more times. We eventually also filter the dictionary to include only tokens appearing at least 15 times and less frequently than the median (to avoid too rare or too frequent tokens), resulting in a dictionary of length 45,174. We then train a Gensim (v3.4.0) LDA model over 100 passes. The resulting 20 topics can be qualified as follows (by the expert):

1. (English) late middle ages and early modern Venice
2. Cultural history, theatre and literature (late early modern period)
3. Napoleonic wars, first and second Austrian dominion, myth of Venice, Republic (19th century)
4. Economic history, mainland, rural administration, Friuli
5. Art and architecture, Palladio, Tiziano, Sansovino
6. Private and family history, wills, marriages
7. (Latin) middle ages, documentary editions
8. Relations with the Ottomans during the late middle ages, especially geopolitics
9. Cultural history, humanism, manuscript and printed books, Petrarca, Sanudo, Manuzio, Bembo
10. Geopolitics during the Renaissance
11. Commerce, maritime administration, Levant, empire
12. (French) late middle ages and early modern Venice
13. The role of and relations with the Catholic church, the Counter Reformation, Paolo Sarpi
14. The lagoon and its environment, the archaeological origins of Venice (middle ages)
15. Economic history, production, guilds, silk
16. University of Padua, the education of the patriciate, science and academies (late early modern and 19th century)
17. The literary origins of Venice, chronicles, Dandolo (middle ages)
18. (German) late middle ages and early modern Venice
19. Economic history, money and public finance
20. Religion, Catholic church, monasticism and churches

The resulting publication to topic distribution can yield a community to topic distribution by averaging the topic probability of the publications part of each community. Using this method, we can qualify the largest communities of the *book* layer as follows (giving the first 5 topics in decreasing order of probability within the community):

1. Book1: 7, 11, 1, 17, 14.
2. Book2: 4, 1, 3, 2, 6.
3. Book3: 5, 3, 1, 2, 9.
4. Book4: 3, 2, 4, 6, 16.

The topic to community structure in this case gives reasonable results, according to which the most probable topics for every community correspond to those associated with the relevant historical period and related topics. The same qualification for the *source* layer yields less clear results, besides a substantial overlap between Book4 and Source3 (19th century). Lastly, once again the *text* layer is difficult to qualify, with both of its communities presenting roughly similar topic distributions.

Discussion

In this contribution we considered the cognitive structure of a specialism in history, suggesting that it is worth representing it as a network made of multiple layers: similarity from references to books, from references to primary sources and textual similarity. We showed that the three layers under consideration are mostly orthogonal, providing different views on the cognitive structure of the specialism of the history of Venice. The *book* layer gives the most fine-grained representation, organized by discipline and historical period, with a clear structure also in terms of the use of primary sources and topics discussed within communities; the *source* layer is coarser, essentially distinguishing among publications heavily using primary sources, those marginally using them and those using recent modern sources. Lastly, the text layer essentially yields a unique community, given that its two emerging ones present a very low modularity and are essentially undistinguishable, providing evidence for the fact that at the level of the language in use, the historians of Venice overlap substantially.

This work has a number of limitations. Most notably, it does not consider citations to journal articles, which would yield an additional, perhaps still orthogonal layer to consider. Furthermore, the similarity measures we use, as well as other methods (community detection, topic modelling), require further empirical assessment, as does the effects of the amount of data used in experiments (especially regarding text similarity). We leave these as future work. Nevertheless, our results provisionally confirm the initial hypothesis that the cognitive organization of specialisms in history, and perhaps in the humanities more broadly, are made of multiple mostly orthogonal layers. Consequently, different layers should be considered according to specific analytical needs.

Bibliography

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research,* 3, 993–1022.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., Schijvenaars, B., Skupin, A., Ma, N. & Börner, K. (2011). Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. *PLoS ONE,* 6(3), e18029.

Börner, K. (2010). *Atlas of Science: Visualizing What We Know*. Cambridge, Mass.: MIT Press.

Colavizza, G. (2017a). The Core Literature of the Historians of Venice. *Frontiers in Digital Humanities,* 4(14).

Colavizza, G. (2017b). The Structural Role of the Core Literature in History. *Scientometrics,* 113(3), 1787–1809.

Colavizza, G. & Romanello, M. (2017). Annotated References in the Historiography on Venice: 19th–21st Centuries. *Journal of Open Humanities Data,* 3.

Colavizza, G., Romanello, M. & Kaplan, F. (2017). The References of References: A Method to Enrich Humanities Library Catalogs with Citation Data. *International Journal on Digital Libraries,* 18, 1–11.