Reconstructing Intellectual Networks: From the ESTC's bibliographic metadata to historical material

Mark J. Hill[1[0000-0001-7273-1775]], Ville Vaara[1[0000-0001-7924-4355]], Tanja Säily[1[[0000-0003-4407-8929]], Leo Lahti[2[0000-0001-5537-637X]], and Mikko Tolonen[1[[0000-0003-2892-8911]]

[1] University of Helsinki, Department of Digital Humanities, Helsinki, Finland
[2] University of Turku, Department of Mathematics and Statistics, Turku, Finland
mark.hill@helsinki.fi

Abstract. This paper demonstrates the use of the ESTC as a representation of material history through the extraction and parsing of its data in a way which allows it to be used in social network analysis. In doing this it makes two contributions. The first is methodological, outlining how such a transformation of data is possible. The second is historical, by demonstrating how this data can be used to support historical claims.

Keywords: Digital History; Social Network Analysis; Metadata; Book History; Bibliographic Data; Intellectual History.

1 Introduction

The English Short Title Catalogue (ESTC), according to the British Library, is the comprehensive, international union catalogue listing early books, serials, newspapers and selected ephemera printed before 1801. Covering over 480,000 documents, which are held by more than 2,000 libraries, it is an essential record of early English print culture, used by researchers to discover and locate early modern texts. However, like other library catalogues, it is "a greatly underestimated source of knowledge" (Tolonen et al. 2018: 1).

Historians of the booktrade have noted the relevance of shifting relationships between authors, printers, publishers, and other actors when studying the history of print. The early modern development of the English market for printed books has been described as a process from an individual cantered medieval craft towards a tightly organised wholesale business orchestrated through a London monopoly (Pollard 1978). Publishing networks naturally, therefore, play a central role when it comes to understanding the nature of these changes (Belanger 1975; Collins 1927). By extension, these networks can also be seen to represent intellectual, religious, and social groups connected by their shared endeavours and interests (Raven 2007). The aim of this paper, therefore - and following recent research which has highlighted the quantitative utility of library catalogues for historical research - is to demonstrate that the ESTC is a representation of these complex and changing historical relationships which can be analysed using quantitative methods (Lahti et al 2015; Tolonen et al.

2018). An important caveat to this claim, however, is that the changes in the publishing landscape are impossible to understand solely through information discovered on the imprint of a document. Thus, the claims made in this paper are exploratory rather than conclusive in this regard - they are proofs of concept and demonstrations as to how this data can be used in novel ways, but they should be understood as aiming to augment traditional historical research rather than being the end source of historical knowledge of booktrade itself. In this way, the ESTC is used not as a catalogue, nor as a research tool per se, but instead as historical material itself which adds to the existing evidence and expertise utilized in the field. In doing this, this paper makes two research contributions: methodological and historical.

2 Methodological Contribution

The first contribution comes from the transformation of historical record into historical relic; specifically, the process of taking a discrete set of catalogue records and turning them into a harmonized relational dataset rich in historical data.

2.1 The Data

As noted, the ESTC is the comprehensive, international union catalogue for English titles. Its construction aimed to include all books (1475-1801), pamphlets, broadsides, and other miscellaneous documents. Each document's record can contain as many as 420 discrete data points, covering details from the document's location in the collections it is held, to its former owner. It is, therefore, an incredibly rich source of historical material, all of which is recorded using the Machine-Readable Cataloging (MARC) 21 standard.[1]

While this means that the data is nominally machine readable, the process of creating a linked dataset should not be trivialized. Originally designed to make the reproduction of catalogues simpler, the content of MARC records was not designed with extraction of linked data, and quantitative analysis of this data, in mind. There are a number of issues which make using it for purposes other than locating texts problematic. Specifically, a more structured data model is required for catalogue data to be useful for systematic quantitative research.

To this end, the Helsinki Computational History Research Group (COMHIS) has an ongoing project which aims to extract all data points from the ESTC.[2] For the purposes of this paper, however, we focus on the extraction of MARC fields 100 and 700 (personal names); 110 and 710 (corporate names); and 260 (imprint), and how we have created a relational data model in which each actor is an unique and identifiable object connected to all titles they were reported to have been involved with, while maintaining the document-specific details of their particular role.[3]

---

1 The Library of Congress has an overview of MARC 21 (https://www.loc.gov/marc/).

2 This paper is a part of this larger research project being conducted by COMHIS (https://www.helsinki.fi/en/researchgroups/computational-history).

3 The code used to parse the ESTC was written in R and Python.

2.2 Extraction, Harmonization, and Preparation

Actors Fields

MARC tags 100, 110, 700, and 710 contain information with regard to actors involved with a specific document. While there may be as many as 20 discrete pieces of data, the key points of interest include names; letters, initials, abbreviations, phrases, or numbers used in place of a name; titles or other words associated with a name; dates associated with a name (birth/death years, active years); the relationship between a name and a work; attribution information when the responsibility is unknown, uncertain, fictitious or pseudonymous; and other miscellaneous information.

With regard to discrete records, we have been able to extract 557,847 references to actors from 397,061 documents (those for which there were actors named). Once extracted, these names were processed to correct numerous unicode issues, data entry errors, and standardize accents. The role of each actor was either extracted (when available), deduced from the MARC tag, or marked as unknown. If a single entry had multiple roles (e.g., "Publisher and printer"), new entries for the same actor were created to reflect these multiple roles on the document.

Importantly, there is no recorded relationship between references to actors recorded in the ESTC. Thus, there is no way of knowing whether one entry for William Shakespeare is those found elsewhere. Or, more problematically, whether any or all of the 137 records for "A. B." are the same person. Thus, before analysis was possible, extensive processing, harmonization, and validation were necessary.

In many cases records within the ESTC themselves were robust enough for this (for example, when an actor's name included specific years of activity). In many cases, however, such information was not available. In these situations we turned to the Virtual International Authority File (VIAF: https://viaf.org/), a harmonized record of multiple international library catalogues which has collected and unified individual records and identifiers for actors. The aim was to take these actor identifiers and apply them to records in the ESTC. To do this, multiple matching steps were made, for example: exact string matches; string matches minus particular types of punctuation; bag-of-word matches for records which contained data considered unique such as years of birth and/or death. Once these matches were exhausted, we made use of the VIAF API to conduct fuzzy searches matching both actor name and the title of the document they were linked to. This resulted in unification of 523,955 records into 68,230 actors, with 34,288 references to actors unmatched.

There were problems, however. First, VIAF has duplicate records for the same actor, and these duplicates made their way into our own data. Additionally, there are harmonized records in VIAF which represent multiple individuals with a shared name. Due to these issues, a number of verification steps were implemented to validate and further harmonize actors. Specifically, there is an extensive semi-automated verification process which looks for false positives (names unified under the same ID, but which are not recorded as identical in the ESTC) and false negatives (names which were identical in the ESTC, but do not share an ID following the harmonization process). In these cases, historical knowledge was used to make a judgement. For the actors which we were unable to unify, a final match was made between those with an

exact name string which were attached to records for multiple editions of the same text. Actors which did not meet any of the previous criteria were treated as distinct, regardless of similarity. At this point 558,243 references to actors have been harmonized into 92,044 unique actors. However, the process has been written to make additional unification a simple task, and thus improvement is an iterative process.

Imprint Field

MARC tag 260 covers the imprint field, notes relating to the publication, printing, distribution, issue, release, or production of a work. While the field itself contains fewer data points (nine) than the actor fields, it is recorded in a much more problematic manner. Specifically, the names of entities representing booktrade actors are contained within a text chunks replicating the full publisher statement. These can be very ambiguous for standard computational extraction methods.[4] To parse these statements, first relevant data was identified, differentiated, and separated from the rest of the imprint using natural language processing. This data is then identified as entities representing persons, roles, names of organisations, and locations.[5]

Once extracted, a similar unification process to that outlined above is followed. First, post processing of names is necessary. This includes correcting spelling errors and normalizing spelling variations when sensical (i.e., Iohn becomes John; VVoodcocke becomes Woodcock), and location information and years of activity are used to match initials with full names (e.g., J. Newbery and John Newbery). Entities are then unified as far as possible by cross-referencing names with existing records extracted from the ESTC, the VIAF, and the British Book Trade Index (BBTI: http://bbti.bodleian.ox.ac.uk/). As with the VIAF, there are issues with the BBTI – in particular duplicate entries which require further processing. Currently we are able to harmonize 35,252 unique actors from 332,410 discrete entries.

Output

To conclude, we have extracted extensive metadata from the catalogue, including titles, publication years and places, edition and imprint information, and all named actors involved or related to a publication. Although this process is largely automated, it has required extensive verification and post-extraction cleaning. Particular attention was given to careful extraction and unification of named entities. Additional enrichment steps are also applied. For example, genders were either extracted from external sources (VIAF) or assigned using a historical gender-name dictionary which we constructed using historical parish records.[6] Overall, however, the processing and cleaning of this data is an iterative process, and we continue to refine the data.

---

[4] Two offer two examples: "printed for Bernard Lintott at the Cross-Keys, between the two Temple-Gates, in Fleet-Street. The Double Gallant: Or, the Sick Lady's Cure. A Comedy. Written by Mr. Cibber"; “printed by E: Coates. 1655. Sould by Thomas Heath in Covent garden, and Henry Herringman at the Ancker on the lowest side of the New-Exchange.”

[5] The current method makes use of the Stanford Natural Language Parser.

[6] See: https://github.com/COMHIS/names_and_genders

2.3 Historical Network Data and its Validity

As the aim of this paper is to use the extracted data for social network analysis, we needed to further transform the bibliographic records. To this end, we created node and edge tables in which documents are edges and unified actors are nodes. The current outputs result in a network of 72,066 nodes connected by 328,996 edges.[7] However, due to the temporal complexity of the booktrade, further refinement was necessary. In particular, in order to discard actors (in particular authors) who continued to be published after their deaths, the dataset was refined to only create links during an actor's lifetime, or active years. Additionally, because we were interested in changes over the entire timespan of the ESTC, as well as during an individual's life, further subsets, covering ten year periods which overlap by five years, were created.

It is important to reflect on the meaning of these connections. The premise is that a shared effort on a particular document represents a real world historical connection. While it is difficult to know whether a printer, for example, could really be said to have had a close relationship with an author, the records nonetheless represent relationships between actors taken from historical objects (Raven 2001: 2). That is, they are not constructed out of curated (by us) information.[8] Additionally, these are a particular type of historical record - a record of the booktrade - which is particularly well suited for network research for at least two reasons: first, the importance of geography; and second, technology as a limiting factor in book production.

The history of the English booktrade is tied to London.[9] By a large margin the majority of early modern publications in English came from London (317,679 to runner-up Edinburgh's 31, 887). This is important because, although it was a large city (London's population grew to roughly one million people by 1800), the booktrade functioned on a human scale in which personal networks mattered, and geographical location reflected personal and intellectual connection. As Raven (2007: 155) reports, booksellers located themselves in proximity to their clientele, and thus clustered amongst competition. Additionally, these locations could be further linked through networks of publishers and printers specializing in given genres or types of publications. A printer, for example, could be attached to a specific bookseller or publisher, even sharing a premise. This is important because it means that the networks which emerge in the ESTC can reflect very real spatial and personal relationships.[10]

---

[7] Network construction and analysis were conducted using the iGraph package for R.

[8] The ESTC is an incredible record of the early modern booktrade. However, it is not entirely comprehensive (Raven 2007: 406-407). For further reflections on comprehensiveness and digital history see: Hill (2017).

[9] For more on London and the history of the booktrade see: Chartier (1987); Feather (1982); Greg (1956); Belanger (1975); Blagden (1951); McKenzie (1969); Myers and Harris (1982; 1987); Treadwell (1982).

[10] Even as publishing established itself within North America, imports remained dominant, and for those who wished to purchase English texts, personal relationships with individuals in London were key (Green 2009). Even in exceptional cases, such as Boston in the 17th and early 18th centuries, and Philadelphia and New York later in the 18th century, London actors remained entangled, be it via direct connections, such as with Benjamin Franklin and William Strahan, or as relocated Londoners, such as David Hall and James Rivington.

The second aspect of the booktrade which suits it for network analysis is one of technological limitation - specifically that of the hand-operated printing press. For the entirety of the period covered by the ESTC, the key factor limiting increases in print production was labour. As the numbers of publications increased, the industry (and the humans who made it) had to grow in tandem. The upshot of this reality (as well as the centrality of London within the industry) is that the booktrade was made up of, and dependent on, human networks (Hirschfeld 2001; Dobranski 2014). This makes the ESTC a potentially rich source from which to construct historical networks.

3 Historical Contribution

The results presented in this paper are meant as proofs of concept that bibliographic metadata can be transformed into historical network data (not to form a cohesive historical argument with regard to the booktrade), and to this end, the next section of this paper is divided into four parts which apply network analysis to the ESTC: an overview of the network data extracted in relation to authors; individual case studies; authorial centrality over the early modern period; and historical community detection.

3.1 Authorial Networks Overview

As a starting point, an overview of data, as recorded and thus reported by the ESTC is offered (Fig. 1).

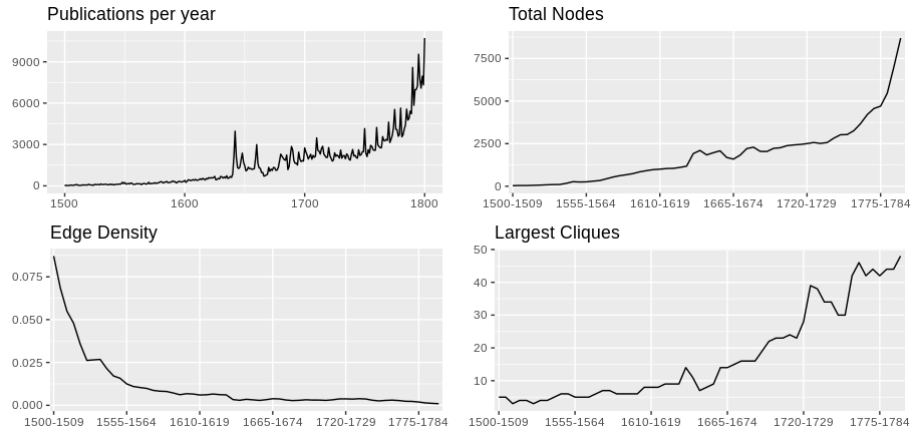

Fig. 1. ESTC network changes: 1500-1799.

While the data for much of the beginning of the period covered is sparse, this in itself allows for comparison.[11] As noted above, the growth of the booktrade was tied to

[11] The jump in the amount of records covering the era of the civil war may be related to an increase in political writings from the time, but the records of these writings can be tied to

increased labour - and this relationship is visible immediately: as the number of documents printed went up, the size the network grows in tandem. However, the transformation is more complex than this, as visible in the increase of cliques and drop in edge density. These are both signs that the network was becoming more diverse, rather than a singular densely interconnected network, indicating an emerging structure in which particular nodes played central roles.[12] That is, as new individuals were brought into the booktrade to increase production, aspects of the trade become increasingly specialized and vertically integrated with specific actors emerging as key in managing and developing the industry (Plant 1974: 59).

It should be noted, however, that there is much historical nuance and debate with regard to the details of the production end of the booktrade. Trade publishing, distribution networks, proofreading, imprecision in both the titles assigned to professions within the trade, and what was actually recorded on imprints, all make the history of publishing an incredibly difficult field to document, and this complexity is clearly visible in the figures above.[13] It is therefore important to, again, note that the data extracted from these networks do not represent claims as to a new understanding of the historical reality of the booktrade. Instead, this data should augment existing historical knowledge in a way which supports it and provides further insights.

With this in mind: when we investigate the relationships between roles specifically, we can see that this complexity is not universal. Publishers, needing to develop more robust networks to advance their commercial ends, grew larger networks over time, while printers became, if not less entangled in these networks, less visible in the imprints reporting them.[14] Authors, however, maintained a similar number of connections (Fig. 2); it was who these connections were with that shifted.

---

the existence of the Thomason Collection of Civil War Tracts, and thus claims with regard to this increase should be tempered. On the other hand, the Great Fire is clearly responsible for the loss of historical works, the records of works, and the potential for production.

[12] These changes also reflect the reality of imprint records not always acknowledging all involved in production. Additionally, the increase in the size of network cliques, particularly in the 18th century, is representative of the historical practise of shared copyrights – a situation which is clearly visible in Fig. 4.

[13] To note how these historical difficulties may impact our data: as the records being used are largely extracted from imprints, and as printers became less important in terms of the distribution of works to the audiences purchasing them, their details may have been less likely to be included. Additionally, as printing houses became larger, or the work more frequently distributed between houses, the pragmatics of including all printers may have been questioned. There is also ambiguity in imprint statements which has previously been noted as problematic. As Raven writes: “Recovering these trading relationships is not easy. Many distribution agents are unspecified.” Having said this, the ESTC remains the best record of these relationships we have, and as Raven importantly concludes: “Despite these reservations, we can at least be certain that in most cases inclusion of a name in an imprint line indicates some sort of financial involvement. This is particularly the case where the author is mentioned” (Raven 2001: 2).

[14] It should be noted that the distinction between publisher and printer during this time was not formalized as it is today, and this is an additional historical issue which needs to be taken into consideration.

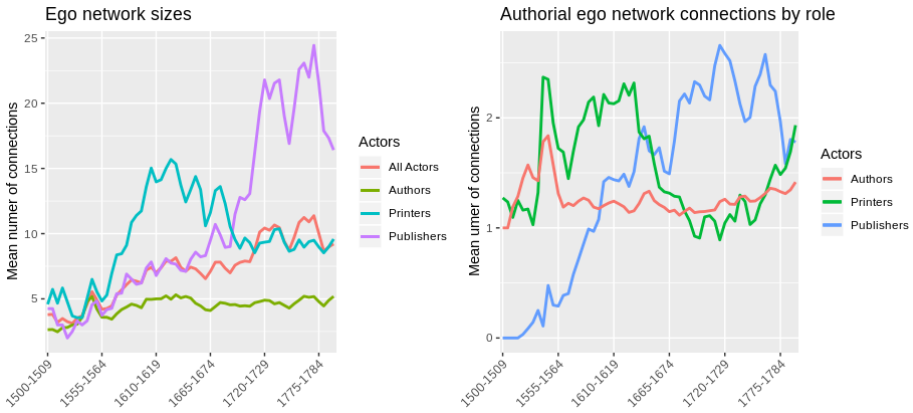

Fig. 2. Left: mean ego network size of authors, publishers, printers, and other actors. Right: mean connections from authors to printers, publishers, and other authors.

Feather (1994: 79) has written that:

> By the middle of the eighteenth century, it was possible to make a decent living from writing… Patronage was no longer essential even to literary authors, and the idea of the author as a participant in commercial activity was beginning to be accepted. At the same time as authors were becoming more overtly mercenary, there was also developing a more elevated idea of the author as artist or creator. The very concept of 'originality', in the sense of an 'original work' of art or literature, is an eighteenth-century idea.

What we may be witnessing in the figures above, then, is the shift from the "self-publishing" author (if such a conceptual category could be said to have existed) to the established "professional author" which, by the eighteenth century, turns to existing publisher-headed networks, rather than printers, to manufacture and distribute their works (Raven 2004: 4). For such a claim to be considered, however, we must acknowledge another complexity in the data: the recorded drop in authorial connections to printers is not uniform.

From the middle of the 18th century the relationship begins to reverse as publisher connections drop and printers once again increase in author ego networks. There are, at least, three historical reasons for this. First, the growth of the colonial booktrade in the 18th century, where publishing had yet to take the form it had in London, meant printers were more frequently identified on imprints than publishers (Fig. 3). The impact of this distinction towards the end of the 18th century was amplified as, by 1792, non-London-based locations outnumber London for the first time.

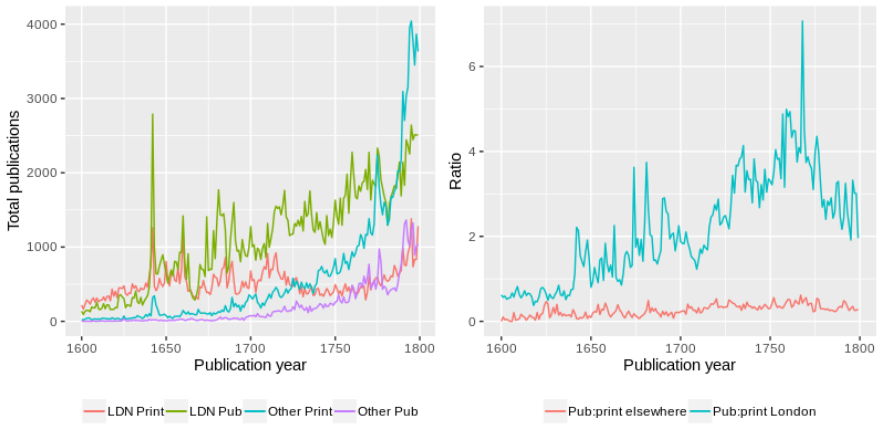

Fig. 3. Left: Total documents in ESTC with publishers or printers attributed to them by year. Right: Ratio of publishers to printers in London and elsewhere.

The second cause for the recorded increase in printers was the 1710 Copyright Act. While the legality of the statute is complicated, in theory it ended perpetual copyright after 28 years. This resulted in reprints of works which entered the public domain - first in the 1740s by printers outside London, and then in the 1770s by those in London (both events visible in Fig. 4) (Feather 1994: 81). The third reason was piracy; while a work legitimately tied to an established publisher would make this clear on the imprint, pirated works were, unsurprisingly, more vague in this regard (Feather 1994: 68).

Overall, it is clear that there is extensive historical data available in these network overviews which offer *longue durée* insights - which is an important contribution in itself from the historians perspective (Armitage 2012). These insights must, however, be tempered by detailed analysis of the finer points - and one way of doing just that is by turning to individual authors.

3.2 Individual Case Studies

The realities of the booktrade for individual authors are better seen when examined independently. To demonstrate this we provide two case studies: William Shakespeare and Bernard Mandeville. To begin, we offer a comparison of the ego network of Bernard Mandeville as extracted automatically from the ESTC, and as constructed by a scholar of Mandeville (Fig. 4).

The two networks are similar in construction: relationships and names reoccur, and a cluster of publishers who shared the copyright of Mandeville's *Fable of the Bees* is visible in both. When discounting unknown actors and booksellers from the ESTC, there is only one actor in the hand-crafted network which was not found in the ESTC network: the publisher John Peele.

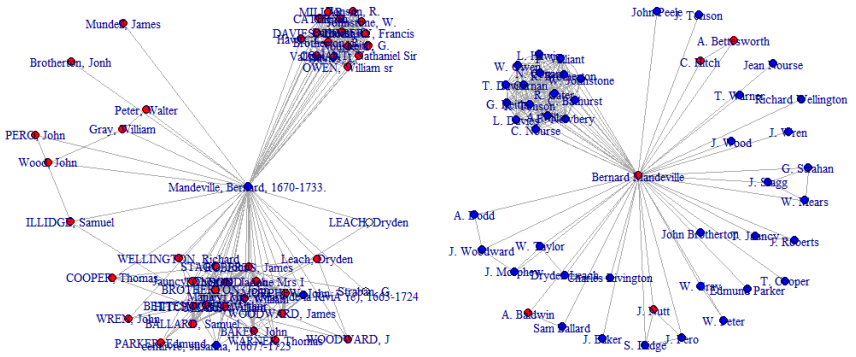

Fig. 4. Bernard Mandeville's (1670-1733) professional network. Left: constructed by a historian of Mandeville.[15] Right: constructed with parsed ESTC metadata.

There are, however, three additional nodes found in the ESTC network: two authors and one publisher. Additionally, the ESTC network contains nine additional publication records (five of which record Mandeville as "Unknown," and three as an attributed author) and 49 further unique connections between actors, a detail which would be exponentially more difficult record when constructing a network by hand. It is, therefore, both historically accurate and robust, and when turning to the makeup of the relationships further historical detail can be noted.

With regard to author popularity, death would perhaps be unsurprising as a limiting factor. While this may not be entirely intuitive in terms of the number of publications attached to an author, it is should certainly be the case with regard to social networks.

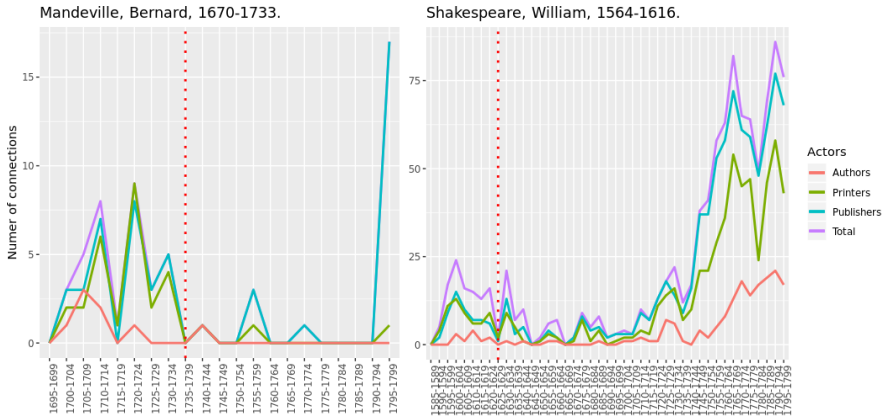

Fig. 5. Breakdown of types and number of connections over time. Dotted red line represents year of death.

[15] Data for hand-crafted ego network provided by Mikko Tolonen.

Our data shows, however, that death need not entirely hinder one's social network; posthumous publications - and those who worked on them - need to be taken into consideration when using this data (Fig. 5).

This is an important point. Methodologically, it highlights the issue of simply extracting all reported relationships. If one is interested in examining potentially real historical relationships, data must take into consideration life and/or active years. Historically, however, there are reasons one may be interested in exactly these types of posthumous records. While, as an author, Shakespeare remained in print after his death, there are noticeable peaks and troughs in his connection counts which are indicative of known editions issued by different publishers.

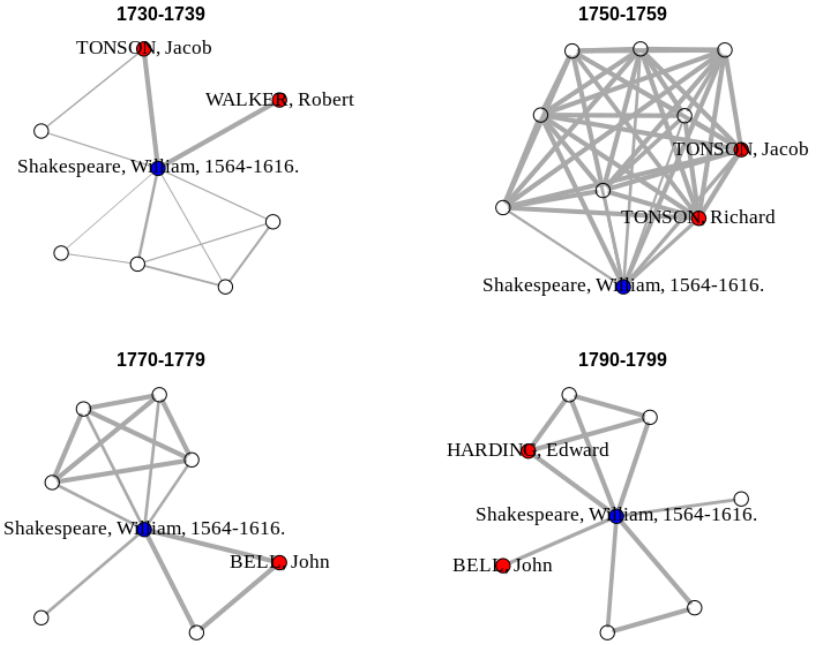

Fig. 6. Ego networks weighted by number of connections and the shift in key publishers over time

When looking at ego maps from different periods, we can clearly identify these publishers (Fig. 6). What is more, the networks themselves offer additional contextual detail - specifically, competition and shifting changes in importance. For example, in 1730-1739 we see distinct connections between feuding publishers Robert Walker and the Tonsons, the latter of which would come to dominate the network in the middle of the century. By 1770-1779, however, the market had changed. John Bell, whose more cheaply printed works, "satisfied public demand for inexpensive editions," emerged as a key actor, and remained visible until the end of the century - although further competition in the discount-edition market can be seen in Edward Harding's arrival (Rasmussen 2000: 343).

While these examples are, purposefully, built upon previous research, they offer demonstrations as to the novel ways in which the ESTC may be used to track authorial relationships at different historical moments.

3.3 Author Centrality

As Wasserman and Faust (1994: 169) have noted, "[o]ne of the primary uses of graph theory in social network analysis is the identification of the 'most important' actors" through centrality measurements – a quantitative method which, by taking into account the relative relationships between actors, can identify key nodes. What makes a node key depends on the measurement chosen, with each emphasizing particular types of relationships. Thus, careful assessment of one's data and the usefulness of each method is necessary when calculating centrality. With this in mind, this section of the paper applies four centrality measurements - degree, betweenness, closeness, and eigenvector - to the ESTC with the aim of both assessing their particular uses, as well as validating the historical richness of the data.

We extracted the ten most central authors per ten year time slices (to ensure contemporarily meaningful results) and compared them: first, against the most published authors for the same time period; second, for instances of direct overlap (i.e., were the same people identified by different measurements); and third, more qualitatively, in terms of how representative those authors were with regard to the historical period they were extracted from. Below is an overview of these results.

From 1500-1535, an era which one may identify with renaissance humanism, degree, betweenness, and closeness centrality measurements most frequently return to humanist authors (e.g., Erasmus, Thomas More, Sebastian Brant, Baptista Mantuanus) while eigenvector and the most-published pointed to religious authors and grammarians. During the reformation (1535 to the mid-16th century) all measurements identified reformation and counter-reformation figures (e.g., Martin Luther, William Tyndale, Thomas Beccon, John Knox, Jean Calvin, Theodore Beza). However, those found in the most-published category are almost exclusively English authors, some of who do not show up in any of the centrality measurements (John Hooper and Edmund Bonner). The dominance of the reformation continued for the first half of the Elizabethan era, although by the 1570s familiar dramatists, poets, and other literary figures emerge (although Puritan authors remain present throughout). Authors include: George Gascoigne, Thomas Newton, Thomas Nash, Robert Greene, Shakespeare, Thomas Heywood, and Michael Drayton. This trend generally continues during the Jacobean era (1605-1640) with the exception of eigenvector centrality which, interestingly, takes an explicitly Anglican and royalist turn, returning (as authors) James I and Charles I, Lancelot Andrewes, John White, William Laud, and John Williams. Unsurprisingly, during the civil war (1640-1655), literary authors disappear and are replaced by overtly political and religious authors (especially nonconformists). On the political end of the spectrum we find, amongst others, Charles I, Oliver Cromwell, Thomas Fairfax, Edward Montagu, John Pym, Robert Devereux, and Christopher Love. Religious figures include George Fox, James Naylor, William Prynne, Richard Baxter, and the first colonial-based author identified, John Cotton.

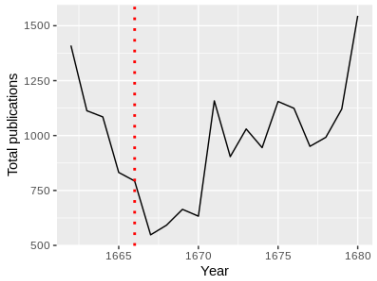

Fig. 7. Great Fire of London's impact on publications. The redline marks the year of the fire (1666).

From 1660-1675 the data is, unfortunately, tarnished by the Great Fire of London which, while taking place in 1666, resulted in the loss of texts going back to at least 1660, and impacted production for years to follow (Fig. 7). There is little cohesion in terms of authors identified as central during this period, although non-conformists dominate the publication records.[16]

The restoration period and beyond - the "Augustan Age," lasting roughly 1660 until 1785 - can be conveniently divided into three authorship-defined sub-eras: the Age of Dryden, the Age of Pope, and the Age of Johnson. The centrality of the namesakes for these literary eras is quickly confirmed. During the first, Dryden is more frequently identified than any other author until 1695 (although other important literary authors returned include Roger de Coverley, Elkanah Settle, Thomas D'Urfey, Abel Boywer, and Gilbert Burnet). This is followed by the Age of Pope (1700-1745) which, again, includes Pope himself, Daniel Defoe, Jonathan Swift, Richard Steele, Eliza Haywood, Colley Cibber, Ambrose Philips, and many others. The Age of Johnson, from 1750 until his death in 1784, is again dominated by the literary giants of the era, such as Johnson himself (although he first appears in 1765), Oliver Goldsmith, Tobias Smollett, John Lockman, George Colman, Hannah Cowley, etc.. This era of literary centrality ends abruptly, however, as political turmoil again emerged – this time round the American and French revolutions – and key figures shift to the likes of Thomas Paine, George Washington, Benjamin Franklin, Joel Barlow, William Cobbett, and Edmund Burke.

For the most part, these results are remarkably in line with historical expectations. To better compare the measurements, however, they were correlated into a series of tables from which the following visualization was constructed (Fig. 8). Each bar represents a specific measurement, and segments ten historical eras. The opacity of each segment represents how well each measurement reported authors deemed to be relevant to a given historical era. The aim is to offer a simplified representation from which the usefulness of the measurements can be assed.

[16] There are two potential reasons for this: one is the prominence of the Dutch printing nonconformist texts in Europe during this time; another being nonconformist social groups in London being located in the East End, away from the fire.

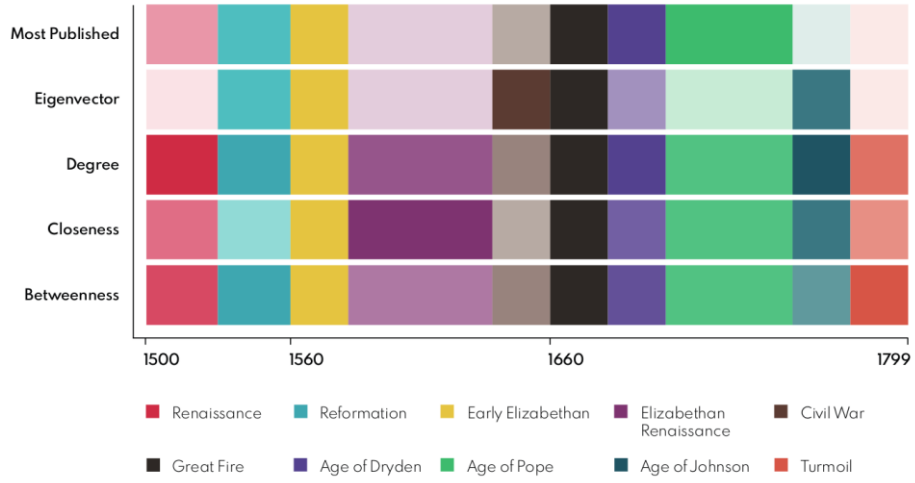

Fig. 8. Centrality measurements and the historical representativeness of the authors returned.

The first finding to draw one's attention to is that the centrality measurements seem to successfully identify key figures, and often better than simple publication records (although Eigenvector is an outlier). Additionally: there is a large amount of diversity between who measurements report as central - especially towards the end of the period covered. Eigenvector was, by far, the most unique, with nearly 50% of all results not found in other measurements. This was followed by closeness (36%), betweenness (27%), and degree (11%). Additionally, 27% of the authors - largely religious - returned as most frequently published were not found in any of the centrality measurements. Interestingly, authors often appear in the publication count category after they are first noted by a centrality measurement with one exception: the first half of the eighteenth century, a time coinciding with the development of the publisher. While beyond the remit of this paper, this points to a complex structural relationship between publishing as an industry, and the popularity of an author.[17]

The results generated from the eigenvector centrality measurements are also worth noting. Figure 8 reports that, as a measurement, eigenvector was quite poor at identifying authors representative of a given period. There were exceptions to this, however - in particular, in the lead up to, and during the, civil war. To understand the reasons for this one must understand that eigenvector identifies centrality by weighing each node by the centrality of all other nodes. That is, any given node is as central (or as important) as the nodes it is connected to. The upshot of this is immediately visible in the data: nodes identified as central by eigenvector measurements often represent

[17] There are, of course, even larger questions which one should ask: for example, do the results represent a central author because of their reception while alive, or do the connections which we are now identifying play a role in our perception? Additionally, the records for those who we have come to expect to be important are likely to be the bibliographic records which are more robust.

more established institutional figures. That is, in the lead up to the civil war we find royalist and Anglican authors, but at the start of the civil war they are replaced by parliamentarian authors. This is perhaps further highlighted by Henry Scobell being identified in the 1650s. While certainly an important figure in his own right, what perhaps contributes to him being identified as central in the entire network was his role as editor of parliamentary proceedings and licensor of newspaper and political pamphlets. This institutional centrality, however, means that the measurement is perhaps less appropriate in situations - or with data - in which a clear hierarchy is not visible. For example, at the end of the 18th century, while most measurements identified leading political authors, eigenvector returned various literary figures - with one exception: the very much institutional figure Edmund Burke.

3.4 Community Detection

Centrality is not the only way in which institutions or establishments can be identified in networks. With algorithmic community detection the ESTC can be used to identify different and changing intellectual groups. To do this, we extracted (up to) the ten largest communities for each overlapping ten year slice using the fastgreedy algorithm.[18] This provided 577 communities varying from as few as two actors to over 1,200. To test whether these represented actual historical communities we developed a semi-automated classification method. First, we extracted all actor IDs attached to a given community, purging non-authors (e.g., publishers).[19] These IDs were then linked back to ESTC IDs, which allowed us to extract the titles and subtitles for the entirety of each community. The titles representing the 577 communities were still too numerous to investigate manually (the total collection of raw text titles was over 350 megabytes). We therefore tokenized titles, counted the occurrences of each type per community, and converted them into proportions. We then measured each individual community against its peer-communities as a single entity. From this, we extracted the tokens for each community which had the highest distance measurement, and a Z score for two population proportions which was statistically significant ($p < 0.05$). The results were less than a megabyte, and much easier for a human to parse.

When examining the time slices covering the English civil war, the contextual meaning of the tokens for two historical groups looked immediately promising. In one community we identified words such as: army, lieutenant, parliament, forces, commons, protector, souldiers, victory, warre, and cromwell. In the other, we found: king, majesties, proclamation, parliament, declaration, gracious, subjects, charles, royall, kingdom, and dominion. That is, the two communities appeared to represent the parliamentary and royalist sides of the civil war (plotted below in Fig. 9).

[18] Fastgreedy was chosen due to the amount of data that needed to be processed. Better results are likely to be achieved using a different method (Yang et al 2016).

[19] This was done, first, because we are primarily interested in authors for this study, second, to minimize duplicate titles, and third, limit the amount of data being processed.

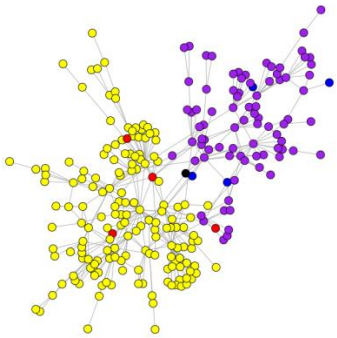

Fig. 9. 1645-1655: English Civil War communities (royalists purple; parliamentarians yellow).

Examining the actors who made up each community confirmed expectations. To offer some examples: the red nodes are prominent parliamentarian military leaders Thomas Fairfax and Percy Algernon, soldier Nicholas Culpeper, and republican pamphleteer William Stafford. Blue nodes represent prominent royalist military leaders Henry Worcester, John Humfrey, William Hamilton Hamilton, and, one-time-parliamentarian-turned-royalist, Edward Massey. Finally, the black node represents an interesting edge case connected to both communities: Marchamont Nedham, a pamphleteer who wrote for both sides during the civil war.

To turn to another example, one which does rely on as clear historical divisions: In the 1645-1654 time slice one can find a community of religious authors, including Quaker founders George Fox, Edward Burrough, and Francis Howgill, as well as various other nonconformists and Puritans. Interestingly, when one follows these actors through time slices, a Quaker-specific community develops, made up of many more prominent Quakers (including Margaret Fell, George Whitehead, James Nayler, and nearly half of the 'Valiant Sixty').

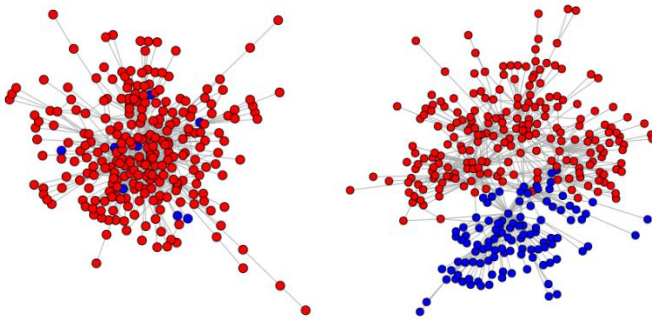

Fig. 10. Left: Key Quakers (blue) in general religious community (1645-1654). Right: Distinct Quaker community (blue) and previous religious community (red) (1650-1659).

While exploratory, these methods, again, offer a number of promising avenues for researchers interested in identifying, exploring, and tracking historical authorship.

4 Conclusion

The goal of this paper has been to offer an overview of the processes used to turn the ESTC into a historical relic, rather than produce a historical argument with regard to authorship as a whole. This remains important future research. However, we believe that there are many promising contributions in this work with regard to both its methodological and historical claims. With regard to the former contribution, we envision a wide variety of scholars who could make use of this work - including historians, linguists, and literary scholars. For example, historical sociolinguists aiming to reconstruct and analyse authorial social networks may find the processing and parsing procedures particularly informative (Bergs 2005; Sairio 2009). Additionally, there is the possibility of combining this type of metadata with massive full-text databases like Early English Books Online (EEBO) and Eighteenth Century Collections Online (ECCO). The latter - historical - contribution made by this paper is of particular relevance to book history. By introducing the statistical measurements unique to social network analysis, this paper augments and improves upon the existing, albeit sparse, quantitative analysis of the early English booktrade.[20] Additionally, we are particularly interested in its potential contribution to intellectual history, and the possibility of using these methods to reconstruct intellectual contexts, such as the relationships between specific actors involved in producing and disseminating philosophical and political texts. Bibliographic records have already shown themselves to be an important foundation for research in this field (Tolonen 2013), and this paper hopes to push this research further.

References

1. Armitage, D.: What's the Big Idea? Intellectual History and the Longue Durée. History of European Ideas 38:4, 493-507 (2012).
2. Belanger, T.: Booksellers' trade sales 1718–1768. The Library, 5th ser., 30, 281–302 (1975).
3. Bergs, A.: Social networks and historical sociolinguistics: Studies in morphosyntactic variation in the Paston letters (1421–1503), Topics in English Linguistics, vol. 51. Mouton de Gruyter, Berlin (2005).
4. Blagden, C.: Booksellers' trade sales 1718–1768. The Library, 5th ser., 5, 243–257 (1951).
5. Blayney, P.W.M.: The prevalence of shared printing in the early seventeenth century. Papers of the Bibliographical Society of America 67, 437–442 (1973).
6. Chartier, R.: Publishing strategies and what the people read, 1530–1660. In: Chartier, R. Cultural uses of print in early modern France, pp. 145–182. Princeton University Press, Princeton, NJ (1987).
7. Chen. D and Manning, C.: A Fast and Accurate Dependency Parser using Neural Networks. Proceedings of EMNLP (2014).
8. Collins, A. S. Authorship in the Days of Johnson: Being a Study of the Relation between Author, Patron, Publisher and Public, 1726-1780. London: Robert Holden, 1927.

[20] In this way it is a continuation of the work done by Lahti et al (2015).

9. Csardi, G. and Nepusz, T.: The igraph software package for complex network research. InterJournal, Complex Systems 1695. 2006. http://igraph.org
10. Dobranski, S.B.: Authorship in the seventeenth century. In: Oxford handbooks online. Oxford University Press, Oxford (2014). doi:10.1093/oxfordhb/9780199935338.013.002
11. Feather, J.: The English Book Trade and the Law 1695-1799. Publishing History 12, 51-76 (1982).
12. Feather, J. Publishing, Piracy and Politics: Historical Study of Copyright in Britain. Bloomsbury, London (1998).
13. Fitzmaurice, S., Robinson, J.A., Alexander, M., Hine, I.C., Mehl, S., Dallachy, F.: Reading into the past: Materials and methods in historical semantics research. In: Säily, T., Nurmi, A., Palander-Collin, M., Auer, A. (eds.) Exploring future paths for historical sociolinguistics, Advances in Historical Sociolinguistics vol. 7, 53–82. John Benjamins, Amsterdam (2017). doi:10.1075/ahs.7.03fit
14. Grafton, A.: The Importance of Being Printed. Journal of Interdisciplinary History II, 265-86 (1980).
15. Green, J.: The British book in North America. In: Suarez, M., SJ & Turner, M. (eds.) The Cambridge history of the book in Britain, vol. 5: 1695–1830, 544–559. Cambridge University Press, Cambridge (2009). doi:10.1017/CHOL9780521810173.030
16. Greg, W.W.: Some aspects and problems of London publishing between 1550 and 1650. Clarendon Press, Oxford (1956).
17. Hill, M.J.: Invisible interpretations: reflections on the digital humanities and intellectual history. Global Intellectual History 1(2), 130-150 (2017).
18. Hirschfeld, H.: Early modern collaboration and theories of authorship. PMLA 116(3), 609–622 (2001).
19. Hunter, I., and D. Saunders.: Lessons from the 'Literatory': How to Historicise Authorship. Critical Inquiry 17, 479-509 (1991).
20. Lahti, L., Ilomäki, N., and Tolonen, M.: A quantitative study of history in the English Short-Title Catalogue (ESTC), 1470–1800. Liber Quarterly 25(2), 87–116 (2015).
21. McKenzie, D.F.: Printers of the mind: Some notes on bibliographical theories and printing-house practices. Studies in Bibliography 22, 1–76 (1969).
22. Myers, R., Harris, M. (eds.): Aspects of printing from London. Oxford Polytechnic Press, Oxford (1987).
23. Myers, R., Harris, M. (eds.): Sale and distribution of books from London. Oxford Polytechnic Press, Oxford (1982).
24. Plant, M.: The English book trade: An economic history of the making and sale of books. 3rd edn. George Allen and Unwin, London (1974).
25. Pollard, G.: The English market for printed books. Publishing History 4, 7–48 (1978).
26. Rasmussen, Eric.: The Years Contributing to Shakespeare Studies: Editions and Textual Studies. In: Shakespeare Survey: Volume 53 (2000).
27. Raven, J.: The book trades. In: Rivers, I. (ed.) Books and their readers in eighteenth-century England, vol. 2: New essays, 1–34. Continuum, London (2001).
28. Raven, J.: The business of books: Booksellers and the English book trade 1450–1850. Yale University Press, New Haven (2007).
29. Raven, J.: The export of books to colonial North America. Publishing History 42, 21–49 (1997).
30. Raven, J.: The importation of books in the eighteenth century. In: Amory, H., Hall, D. (eds.) A history of the book in America: The colonial book in the Atlantic world, vol. 1, 183–197. Cambridge University Press, New York (2000).

31. Raven, J.: London booksellers and American customers: Transatlantic literary community and the Charleston Library Society, 1748–1811. The University of South Carolina Press, Columbia, SC (2002).
32. Rose, M.: Authors and owners: The invention of copyright. Harvard University Press, Cambridge, MA (1993).
33. Saunders, D.: Authorship and copyright. Routledge, London (1992).
34. Simpson, P.: Proofreading in the sixteenth, seventeenth and eighteenth centuries. Oxford University Press, Oxford (1970; orig. 1935).
35. Tolonen, M.: Mandeville and Hume: Anatomists of civil society. Voltaire Foundation, Oxford (2013).
36. Treadwell, M.: London printers and printing houses in 1705. Publishing History 7, 5–44 (1980).
37. Treadwell, M.: London trade publishers 1675–1750. The Library, 6th ser., 4, 99–134 (1982).
38. Wasserman, S. and Faust, K.: Social Network Analysis. Cambridge University Press (1994)
39. Wolf, E.: The book culture of a colonial city: Philadelphia books, bookmen, and booksellers. Oxford University Press, Oxford (1988).
40. Woodmansee, M.: On the author effect: Recovering collectivity. In: Woodmansee, M., Jaszi, P. (eds.) The construction of authorship: Textual appropriation in law and literature, pp. 15–28. Duke University Press, Durham (1994).
41. Yang, Z., Algesheimer, R., Tessone, C.J.: A Comparative Analysis of Community Detection Algorithms on Artificial Networks. Scientific Reports 6 (2016)