Contextualizing controversies of the post-Lutheran reformation: A workflow for network analytics involving relational and graph databases

Aline Deicke and Anna Neovesky
Digital Academy, Academy of Sciences and Literature, Mainz, Germany
aline.deicke@adwmainz.de, anna.neovesky@adwmainz.de

Abstract

Data accumulated by long term research projects offers interesting perspectives for network analytical research when made available under a free license. In the research project "Controversia et Confessio", the datasets are stored in relational databases, whose structures do not lend themselves easily to the relationship-based thinking necessary for network research. The following paper presents an exemplary analytical workflow involving graph databases as an intermediary storage layer as it was employed in a current network analytical study on this data by the authors.

1 Introduction

Over the last decade, network analysis has become a widely popular method of examining historical data. With the growing number of research projects publishing their collected data under open licenses, it has become possible to conduct analyses that rely on big data sets of high quality. One of these projects is “Controversia et Confessio”[1], a long term research project of the Academy of Sciences and Literature | Mainz and the Leibniz Institute of European History in Mainz. Its data, published under a CC-BY-license, is stored in a MySQL-database that also powers a website. Because of its distinct purpose and structure, this relational database proves difficult as a basis for the data requirements needed for multiple, distinguished network analyses. Therefore, a workflow has been developed that relies on graph databases as an intermediary layer of data storage and analysis.

1.1 The research project “Controversia et Confessio” and use-case

The project “Controversia et Confessio - An Edition of Sources for the Development of Confessional Documents and Confessionalization” focuses on the controversies that occurred during the “Wittenberger Reformation” after Luther’s death. The events served to define the Lutheran doctrine and to consolidate the Protestant denomination. The main medium through which the discussions were carried out are pamphlets [1]. The project collects and redacts these pamphlets, which are published thematically and categorized according to controversies as for example the “Osiandrian Controversy”[2].

Another result of the ongoing research and a central element to access the information is the bio-bibliographic database, which contains the pamphlets with title, incipit, commentary, description, corresponding controversy, and further metadata, as well as biographies of the most important persons

[1] http://www.controversia-et-confessio.de/ (Accessed on 2016-04-09).

[2] Overview of the volumes published so far: www.ieg-mainz.de/forschungsprojekte/kontroversliteratur_und_streitkultur/controversia_et_cofessio (Accessed on 2016-04-09).

M. Düring, A. Jatowt, J. Preiser-Kapeller, A. van den Bosch (eds.): Proceedings of the 3rd HistoInformatics Workshop, Krakow, Poland, 11 July 2016, published at http://ceur-ws.org

involved in the controversies. Initially organized in two separate databases, these were combined into one relational database in 2015, when the Digital Academy[3], the Digital Humanities department of the Academy of Sciences and Literature | Mainz, started to collaborate with the project. Currently, the database contains 2063 pamphlets and about 800 persons or entities.

The use-cases portrayed in this paper belong to a research project conducted by the authors which aims to examine the nature of theological discussion through pamphlets from a network analytical perspective[4]. For this purpose, the relationships of authors and opponents as stated in the pamphlets are transformed into a one-mode network and analysed. Apart from this, the study also focuses on determining the potential of network methods for traditional editorial work. Additionally, the transformation of data into different formats and data models also helps to assist the process of data normalization that was undertaken to improve the quality of the project's database.

2 Relational database

The data is stored in a relational database management system, MySQL. The MySQL database is an integral part of a complex content management driven online platform. A restructuring towards another data model would have affected the application too much, but a graph based representation is ideal for network related concerns, as it emphasizes the relation between items. Therefore a separate graph database was created for the network analysis. As the steps for the transformation are fully documented, the scripts can be reused and in case of modifications or corrections in the data the graph database can be updated anytime with fresh data from the RDBMS.

3 Graph database model

First, the relational data model is transferred to a graph-based model (figure 1). The central objects stay the same: Pamphlet, actor, year of publication, place of publication, and controversy. In the relational data model, all of the mentioned items are tables in the database. In the graph model, said tables are nodes containing the information necessary for the network analysis (the data model in the SQL-database contains much more information, for example text and name variants, comments on date, folio and material). The edges between the nodes are directed to introduce better semantics to the graph-based data model. Other than most nodes, the nodes with the labels "Actor" and "Pamphlet" can be connected via two different options, "AUTHOR_OF" and "OPPONENT_OF". According to this model, the relational data is imported into the graph database, in this case the open source community edition of Neo4j[5].

4 Import of MySQL-data and export of graph data

To transfer the original data, all necessary fields from the MySQL database are dumped into several CSV-files. Before importing the data it should be checked that the files are well-formed, for example by using appropriate software or the tools of the Neo4j-shell itself[6]. A detailed description of the import approach as it was employed can be found in Andreas Kuczera's article on graph databases for Historians [2, 3].

[3] www.digitale-akademie.de (Accessed on 2016-04-09).

[4] The authors would like to thank Dr. Jan-Martin Lies and Dr. habil. Kęstutis Daugirdas as well as Prof. Dr. Irene Dingel for their constructive criticism and support.

[5] http://neo4j.com; https://github.com/neo4j (Accessed on 2016-04-09). Version used: Neo4j Community Edition 2.3.2.

[6] http://neo4j.com/developer/guide-import-csv/#_csv_data_quality (Accessed on 2016-04-09).

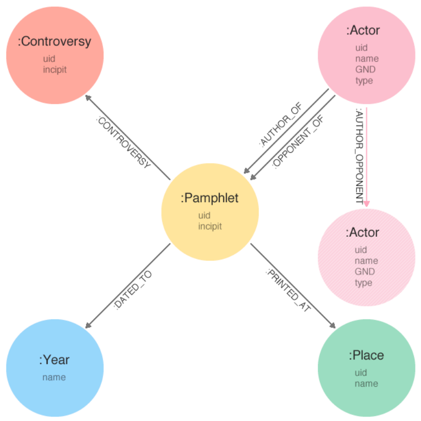

Figure 1: Graph-based data model of the "Controversia et Confessio" dataset

To ensure a better backwards compatibility and easier comparisons between the MySQL and the graph database, the original unique identifiers from the MySQL database are used to identify single nodes. Due to the fact that the table-based context of the RDBMS does not exist in the graph database, these uids are not actually unique by themselves. This problem can be dealt with through the introduction of labels[7] such as ":Pamphlets", especially in the commands creating edges between different types of nodes. Labels also allow for a greater deal of specificity and easier and more semantic handling of queries, as will be evident later on.

For the network analysis, relationships between authors of pamphlets and their opponents are explored. To export the appropriate data out of the graph database, two approaches are possible, namely the export as a CSV or a GraphML file

CSV has been the preferred data format for network research for a long time as it has an easy structure and is human readable. However, in some network visualization programs, problems can arise when more complex edge tables containing attributes are imported, especially if edges are to be summed up to create weighted relationships. This can be circumvented with the use of GraphML[8], an XML standard format used to describe graphs. While there is no native support in Neo4j for exporting GraphML, a plugin is available that among other features enables the export of GraphML from the Neo4j-shell[9]. Because GraphML does not just mark up nodes but also the edges themselves, an additional relationship has to be added to the graph data model which up to now was included only implicitly, namely the AUTHOR_OPPONENT-relation (figure 1).

5 Incorporating graph databases into network research

The exported CSV- or GraphML-files can be imported into a network analysis software for further visualization and processing. In the case of this project, one of the challenges posed by the source data is the handling of relationship attributes in combination with weighted edges. Because authors can have opponent-relations to the same actor in form of several pamphlets, the weighted author-opponent-relationship between them can span not just one date, but a range of years, and can include affiliations to several controversies. While this particular problem was not of importance to the current study, it can be solved by combining edges before the import into the network analysis software. In case of the "Controversia et Confessio" dataset, the import of author-opponent-edges as described above results in a graph containing 389 nodes and 827 weighted edges[10].

While conducting the analysis of the author-opponent-graph, it frequently proves beneficial to return to the graph database due to a number of circumstances, e.g. to export new data with different edges, attributes or other variables to answer new research questions arising in the context of the conducted research, to further examine relationships in their original context – because the data is kept in a relational structure in the graph database, there is no cognitive dissonance as with databases that are relational only in the computational sense – or to conduct further non-network analysis on the data.

[7] http://neo4j.com/docs/2.3.2/graphdb-neo4j.html#graphdb-neo4j-labels (Accessed on 2016-04-09).

[8] http://graphml.graphdrawing.org/index.html (Accessed on 2016-04-09).

[9] https://github.com/jexp/neo4j-shell-tools (Accessed on 2016-04-09).

[10] Preliminary results of the study as presented at the 10. Workshop on Historical Network Research in Düsseldorf, 28.–30.04.2016, can be found at http://prezi.com/7kan3xmiarmw/?utm_campaign=share&utm_medium=copy&rc=ex0share (German; Accessed on 2016-04-09).

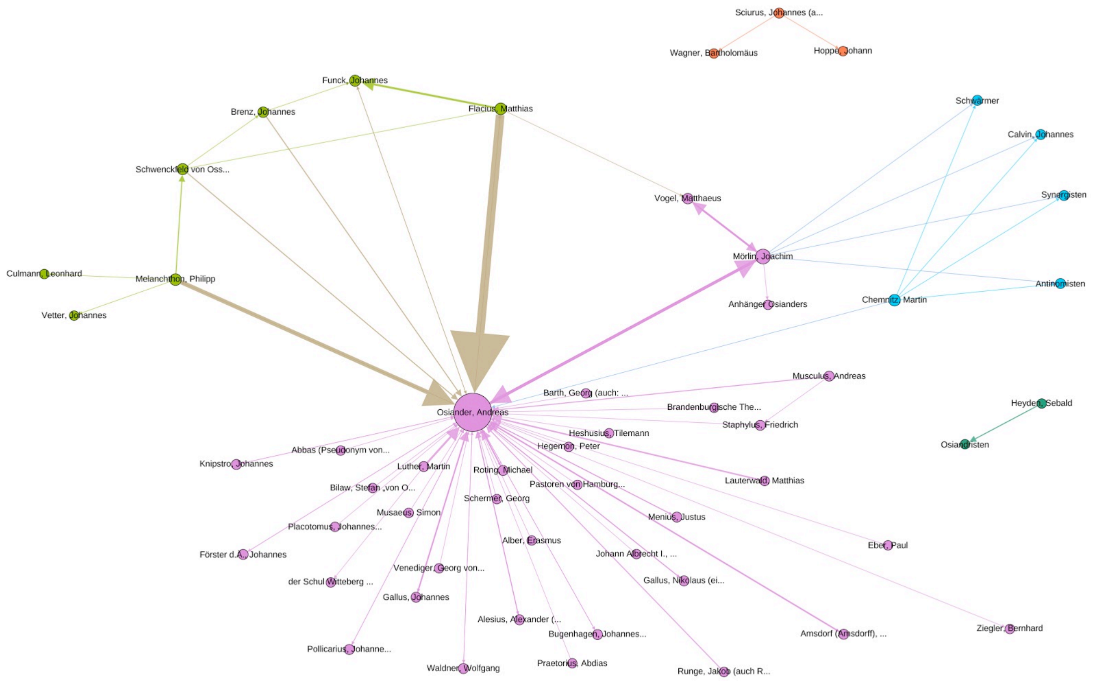

Figure 2: Subgraph visualizing the "Osiandrian Controversy" (Made with Gephi 0.9.1. Size of nodes: Degree, color of nodes: Modularity).

An example for the first case would be the export of subgraphs representing single controversies or time spans, or the examination of different relationship types like co-authorship. In the third case, one application could be a statistical examination of the data such as contrasting the number of pamphlets an author has written against an opponent with the number written without such an explicit reference.

For the second case, a simple example shall be discussed. While analyzing the subgraph of the controversy pertaining to Andreas Osiander's view of justification, the so-called "Osiandrian Controversy", several features of the network turned out to require more in-depth attention.

As figure 2 shows, there is a cluster of several actors (visualized in blue) in one corner of the graph who are all connected to two persons – Joachim Mörlin and Martin Chemnitz – but neither to each other nor to other actors in the graph. Because of the nature of pamphlets, which can be written by several persons against a large group of recipients, the context of this feature was examined in the graph database:

```
//Match pamphlets written by Martin Chemnitz in the context of the
'Osiandrian controversy' and all connected actors
MATCH (n:Actor)-[a:AUTHOR_OF]->(m:Pamphlet)<--(o:Actor), (m)-
[c:DATED_TO]->(p:Year), (m)-[e:CONTROVERSY]->(r:Controversy
{uid:'6611'})WHERE n.name =~ '.*Chemnitz.*'
RETURN n, m, o, p
```

A query is written which matches all pamphlets written by persons whose names contain the string "Chemnitz" and which are connected to the controversy with the uid "6611", that is to say the "Osiandrian Controversy". To catch all actors connected to these pamphlets, the relation type between the pamphlets m and the actors o is unspecified.

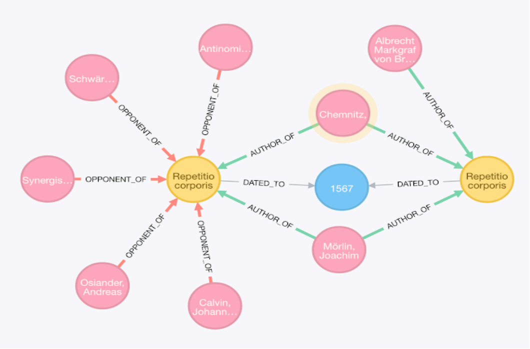

Figure 3: Graph view of the Neo4j browser interface showing pamphlets written by Martin Chemnitz and connecting nodes

The visualization output, when viewed in the graph view, can be further adjusted by assigning different colors and sizes to nodes as well as edges, which can serve to clarify the resulting graph structure. As seen in figure 3 Chemnitz wrote two pamphlets regarding the "Osiandrian Controversy", both in the year 1567. Because one of them has no opponent, it is not included in the network analysis. The other one, however, was written with Mörlin and directed against five actors, which conform to the ones seen in the subgraph above (figure 2).

Going back to the network analysis program, we can now confirm that the blue cluster consists of just one pamphlet involving two authors and five opponents. In addition to this, its late date – 1567 – shows that it was written years after the main discussion regarding Osiander's view of justification had ended.

To clarify the matter even further, more information regarding the overall context can be queried, e.g. which other controversies are addressed by the pamphlet in question. In a next step, these findings can be checked against the bibliographical database of the project "Controversia et Confessio" to find out which role, if any, this specific pamphlet played in the "Osiandrian Controversy" or if it was a later paper dealing with general issues or summaries of prior discussions.

6 Conclusions

Frequently, project data is stored in relational databases, which is not ideal for a network analytical workflow. In order not to compromise web applications running on these datasets, it is advisable to transfer this data into a second, better suited architecture. Graph databases allow for an easy handling of data based on relationships and for an effective workflow going back and forth between storage and analysis software without having to switch between different types of data modeling. As shown, the graph model enables segmentation as well as further independent analysis of selected network data. Moreover, by providing a new perspective on the data set, this workflow yields new impulses for research as well as data curation and normalization processes.

Literature

1. Dingel, I. Zwischen Disputation und Polemik. „Streitkultur" in den nachinterimistischen Kontroversen. in Jürgens, H. P. and Weller, T. ed. *Streitkultur und Öffentlichkeit im konfessionellen Zeitalter*, Vandenhoeck & Ruprecht, Göttingen, 2013,17-29.

2. Kuczera, A. Graphbasierte digitale Editionen, *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte*. Retrieved April 9th, 2016, from http://mittelalter.hypotheses.org/7994.

3. Kuczera, A. Graphdatenbanken für Historiker. Netzwerke in den Registern der Regesten Kaiser Friedrichs III. mit neo4j und Gephi, *Mittelalter*. *Interdisziplinäre Forschung und Rezeptionsgeschichte*. Retrieved April 9th, 2016, from http://mittelalter.hypotheses.org/5995.