

# Social Network Analysis and Co-Occurrence: Identifying the Gaps

Jens Dörpinghaus<sup>1,2</sup>

<sup>1</sup>Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany,

<sup>2</sup>University of Koblenz, Germany

## Abstract

Social Network Analysis is widely used in the humanities. However, historical and narrative texts in ancient languages are usually challenging for NLP (natural language processing) methods and AI (artificial intelligence) technologies developed for modern languages due to their complexity and missing models. In this article, we will focus on biblical texts. Here, linguistic resources are already available. However, no approaches for the automated linking of actors and other information, e.g. spatiality, are available. Thus, in this paper, we will analyze if co-occurrence might improve the linking of data in manual exegetical work. We will provide a detailed analysis and evaluation to identify the gaps for further research. The results of this paper are not limited to theology, but can be applied in all fields working with textual information.

## Keywords

Text Mining, NLP, Social Network Analysis, Co-Occurrence

## 1. Introduction

### 1.1. Motivation

Social networks play an important role in the social sciences and have been widely used for several decades, both in theory and in application. Understanding social interactions and networks and how they influence society are important issues. In the last few years there has been a growing interest in using social networks in historical sciences. Quite recently, considerable attention has been paid to social networks in religious studies and especially in theology. It was shown that social network analysis (SNA) helps to understand the ancient literature on the early religious movements and social identity.

Collar [1], for example, as an archaeologist, was among the first to combine religious studies and archaeology using SNA. In her work “Religious Networks in the Roman Empire” she examines why some cults and religions within the Roman Empire either vanished or became meaningless while maintaining the same popularity. After an introduction, she examines various cults, including the Jewish diaspora after 70 A.D. It is not yet known whether the SNA can be generalized in all cases, since the lack of data is a challenge. Regarding New Testament research the studies of [2], [3] and [4] should be mentioned. One of the main issues in what we

---


2nd Workshop on Humanities-Centred AI (CHAI-2022)

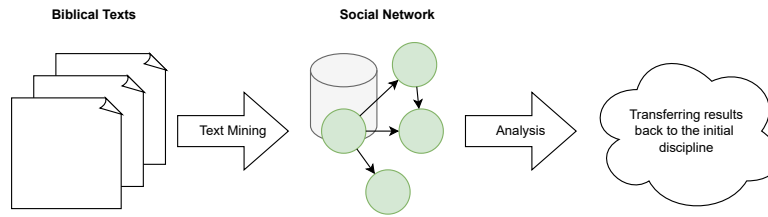
✉ jens.doerpinghaus@bibb.de (J. Dörpinghaus)

ORCID 0000-0003-0245-7752 (J. Dörpinghaus)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** Illustration of the proposed workflow. Narrative texts are transformed to social networks, which can be analyzed with methods from SNA. However, in the humanities these results must be interpreted within the framework of the initial scientific domain. In this work, we focus on the first step.

do *not* know about a social network is in particular what we can not reproduce. Thus, “network analysis needs to be embedded within traditional research both to produce results and not to be ignored. For this achievement, Collar should be praised.” [5, 226]

Most studies from historians and exegetes have only focused on understanding how the New Testaments constructs networks and identity. Thus, on the one hand we have developed mathematical computational social networks using exegetical methods. On the other hand these results should also raise new questions, show new perspective on biblical texts, and in particular on how these texts can be analyzed with AI methods. Previous work has been limited to only one of these goals, while very little work has been carried out on AI approaches for understanding biblical texts. In particular, techniques to build a large computational social network of early Christianity based on biblical, or early church texts, and other sources are time-consuming and require an interdisciplinary approach.

From a linguistic perspective, SNA always includes persons. These may well be fictitious or the information about them can be worked out by exegetical steps from historical sources, which Rollinger [6] did for the epoch of antiquity. Thus, the social network paradigm can technically be applied to narrative texts without any problem. As an example for a first systematization of these relations, the so-called figure configuration which Cornils [7, 75] uses for Acts may serve: This is a pure listing of characters appearing at the same time in a narrative. Thus, the computer-based evaluation of this data already used in the literary analysis is merely another logical step.

However, since little work has been done in the field of AI approaches for biblical texts, see [8] and [9], there is a gap in research. While more digital methods to analyze the results of manually curated analyses are used within theology, there is an important overlap to other methods from the humanities and their approaches to understand texts. But no evaluation of these methods are yet carried out.

This paper seeks to address these gaps on a particular research question: Can we use AI methods within the field of theology on biblical texts to build a social network on these narrative texts to apply methods from SNA?

## 1.2. Research Question

This work is embedded within the steps to generate a complete computational social network reconstruction of biblical texts, in particular Luke-Acts, for an analysis with various sociological

distance measures to establish a better understanding of the biblical text. In Figure 1 we present the ideal workflow which comprises two steps: (1) building a social network, and (2) analyzing the social network. However, in this paper, we focus on the first step. Since we have already stated that basically no AI methods for the automated detection of figure constellations exist – we will prove this statement in the next section – we will present some naive algorithm to detect the challenges, gaps, and problems to build a social network. In addition, we will provide a detailed quality control on a manually curated social network representation of the Gospel of Luke and Acts with classical exegetical methods. Exegetical methods are used to explain or interpret – not only religious – texts and literature within a given hermeneutic framework.

## 2. Related Work

Network approaches have been used in historical studies for some decades. Here they are often called *historical network analysis* (HNA). Reitmayer and Marx [10] note that many methods are used and no common formal structure exists. Only selected methods of network analysis are used, a full network analysis as it is conducted within the social sciences is usually not being carried out. In particular, a subset of literature uses the term “network” without using methods from SNA or HNA, see for example van de Kamp [11].

Networks in early Christianity have not yet been fully investigated. Duling [2, 136] summarizes the situation: “interest in SNA by Biblical scholars has been sporadic, but steady, and is apparently growing”. First approaches can be found in Thompson [12], who examines the communication of information in the network of early Christians between the years 30 and 70 A.D. Further attempts to explore these questions with the help of social network analysis were carried out in the work of Duling [13] and Duling [14] which are entitled “The Jesus Movement and Social Network Analysis”. In general, Dulling’s work remains unfinished. Another scholar working with SNA is McClure<sup>1</sup> who draws her final observations in [17, 35]: “The results provide a unique window into the relational dynamics portrayed by the Gospels, producing a variety of insights, some which may not surprise biblical scholars but others which hopefully will inspire further consideration.” However, to sum up, a complete computable network of early Christianity according to the biblical texts is still missing.

Within narrative studies, a character is a main (or minor) actor described in the text. This is equivalent to the actor in SNA. Narrative criticism provides a more detailed view: “Characters reveal themselves in their speech (what they say and how they say it), in their actions (what they do), by their clothing (what they wear), in their gestures and posture (how they present themselves)” [18, 121] Resseguie also points out a social perspective by mentioning their position within society. Thus, it is also important to think about the constellation of characters, which means their position in a network<sup>2</sup> and their relation to the plot.

The character analysis can be separated into quantitative and qualitative questions: When is a character present (in drama: “stage presence”) and with whom does he interact? Qualitatively,

---

<sup>1</sup>She worked with a harmonized version of all gospels and was first working on support, conflict and compassion [see 15]. After that, she investigated subgroups and balance [see 16]. While the methodological approach remains somehow unclear (for example the data is changed which makes the studies incomparable), she carries no detailed discussion on her choice of methods.

<sup>2</sup>This equivalent to *Figurenkonstellation* found in Finne and Rügemeier [19, 204].

one can also ask about content (the “character speech”) or about characterizations. The first is answered by the “figure configuration” and its “configurational structure”: In the first, the person and their interactions are inferred; in the second, they are juxtaposed. While the extraction of characters as word entities is not very difficult, the accurate analysis of interactions is challenging. Therefore, current studies focus on “co-presence”. Rarely are models explored to precisely describe these interactions, and they are usually limited to actor lists or dramas, see [20] or [21]. In New Testament studies, figure constellations have been generated manually so far, see [7],[9], and [22]. However, co-presence is more than co-occurrence, which describes only those terms which explicitly occur in the same sentence.

### 3. Methodology

#### 3.1. Data

Here, we will focus on the Greek text, and English Bible translations, although this approach can be used for any other language with linguistic annotations. There are several software packages available to access Biblical texts. Some commercial software like Logos offer no or only very limited access to their API<sup>3</sup>. Thus, we did our work on the basis of the SWORD Project, which offers a full API available under GNU license<sup>4</sup>. As a basis for the Greek text, we used the SBLGNT 2.0 from Tyndale House, based on SBLGNT v.1.3 from Crosswire. This text is with some minor changes comparable to the Nestle-Aland/United Bible Societies text. The English texts are based on and ESV (English Standard Version, 2011). Beside of them, all data is available with a free license. See <http://www.crosswire.org/sword/modules/> for details of these packages. Since these texts are already annotated with linguistic information and contain Strong’s annotations referring to the original Greek term, several components of NLP-pipelines like POS-tagging, lemmatization, and NER can be omitted. There are several annotations which can be displayed in different ways.

We will use annotations for extracting information, storing and comparing them. However, while these dictionary annotations allow the processing of terms with their linguistic information, we still have no information about whether a word refers to a person, a location or other entities. To collect the training data, we could use the complete New Testament texts mentioned above. This leads to 7,957 verses in each version. There are 5,624 entries in the Strong’s dictionary. However, we will now discuss the limitation of our analysis to those parts with evaluation data.

#### 3.2. Evaluation data

To overcome the limitations mentioned above, we will proceed with a manually curated network of the Gospel of Luke and Acts. The first one comprises 99 nodes and 628 edges, the network for Acts comprises 126 nodes and 646 edges, for details we refer to [22] and [8]. We will limit our analysis to these two books for an evaluation with this manually created network. The

---

<sup>3</sup>See for example [https://wiki.logos.com/Logos\\_4\\_COM\\_API](https://wiki.logos.com/Logos_4_COM_API) and [23].

<sup>4</sup>See <http://crosswire.org/sword/index.jsp>

	$ V(G) $	$ E(G) $	$ E(\overline{G}) $	Precision	Recall	$F_1$ Score
Gospel of Luke	99	628	4223	0.62	0.56	0.58
Acts	126	646	7229	0.71	0.53	0.61

**Table 1**

Number of nodes  $|V(G)|$  and edges  $|E(G)|$  in the network, and number of edges not in the network but in the complementary graph  $|E(\overline{G})|$ . We present precision, recall and  $F_1$  Score for the co-occurrence approach.

networks provide a list of expected actors, locations, and concepts. Thus, it already limits the output of the algorithmic approach, which we will discuss in the next subsection.

### 3.3. Algorithmic approach

As we have discussed earlier, there are currently no AI approaches available for stage presence of actors [9]. Thus, our initial algorithmic approach is based on the following observations: Given two actors and their Strong-number, for example, Paul (strong:G3972) and Barnabas (strong:G0921): Which verses have a co-occurrence of both actors? Given a corpus  $\mathbb{C}$  of texts with sentences  $s \in \mathbb{C}$ , we will denote this value with  $c(a, b)$  for given actors  $a$  and  $b$ :

$$c(a, b) = \begin{cases} 0 & \nexists s \in \mathbb{C} : a \in s, b \in s \\ 1 & \exists s \in \mathbb{C} : a \in s, b \in s \end{cases}$$

In the social network, we can add edges between  $a$  and  $b$  if they are co-occurrent, which means  $(a, b) \in E \Leftrightarrow c(a, b) = 0$ . Co-occurrence is a widely studied, yet not unproblematic, approach for text analysis, see [24] and [25].

The evaluation will be carried out using two approaches: (1) Which gaps can be identified when comparing the results of co-occurrence with a manually curated network as ground truth? (2) Which gaps can be identified with a detailed perspective on particular actors?

## 4. Analysis

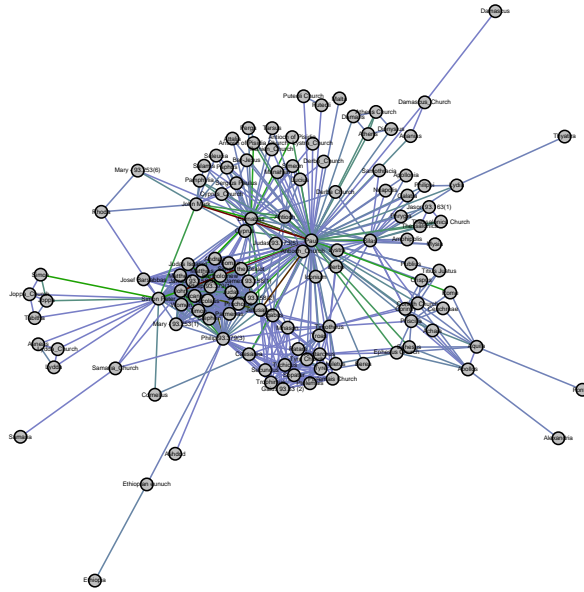
### 4.1. Social Network Comparison

First, we will analyze the quality of the co-occurrence based approach with a manually curated network of Acts. We will consider the the  $F_1$ -score which is as a weighted average of the precision and the recall:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \text{ where}$$

$$\text{Precision} = \frac{TPR}{APR}, \text{ and}$$

$$\text{Recall} = \frac{TPR}{APS}.$$



**Figure 2:** Illustration of the manually curated network representation of Acts. The edge color refers to the value  $c(a, b)$  of edges between actors  $a$  and  $b$  by co-occurrence. Blue refers to a value of zero, green to low and red to the highest values.

Here,  $TPR$  refers to true positive results,  $APR$  to all positive results, and  $APS$  are all samples that should have been identified as positive.

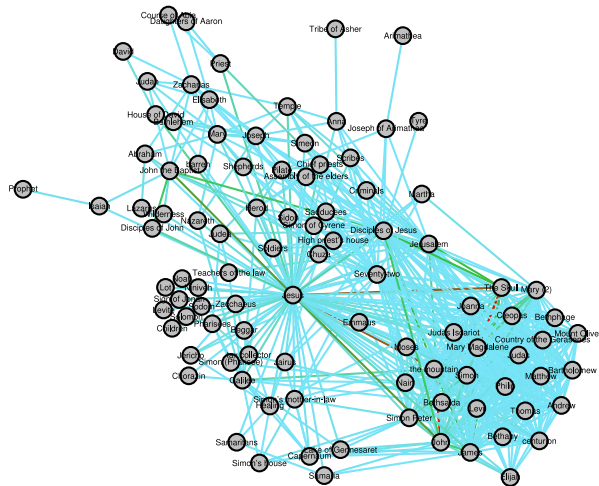
In Table 1 we present the results for both the Gospel of Luke and Acts, in Figures 2 and 3 a visual presentation of the values for the edges in the network. For Acts, we found in average 1.49 evidences for every edge. However, for 286 edges, we could not find any evidence in the text with this method. Some of these values suggest problems: For example, we see 30 co-occurrences of Paul and John – and Paul and John Mark. Both are annotated with the same name ‘John’. In addition, several actors are called James. We will continue a discussion about these well-known issues of disambiguation in the next subsection.

Thus, the true positive rate is 360 and the recall 0.56. However, to calculate the precision, we should also think about those 7229 edges which have not been added to the network. For this purpose, we build the complement graph and compute the co-occurrences of both actors.  $\bar{G}$  is called the *complement* of  $G$  with  $V(\bar{G}) = V(G)$  and

$$e \in E(\bar{G}) \Leftrightarrow e \notin E(G).$$

Here, we found in average 0.16 evidences for every missing edge. However, for 6648 edges we could not find any evidence. The precision of this approach is 0.62, and the  $F_1$ -score is 0.58 and the result is similar for the Gospel of Luke, see Table 1.

However, in the Gospel of Luke we found in average 3.51 evidences for every existing edge, which highlights that the narrative shows a different structure. However, the scores underline the question, if this method is applicable at all. Thus, we will discuss some results in more detail.



**Figure 3:** Illustration of the manually curated network representation of the Gospel of Luke. The edge color refers to the value  $c(a, b)$  of edges between actors  $a$  and  $b$  by co-occurrence. Blue refers to a value of zero, green to low and red to the highest values.

## 4.2. Detailed Analysis

For a detailed analysis, we consider the actor Paul (strong:G3972). Acts states several times that he is working closely with Barnabas (strong:G0921). Indeed, we find 14 co-occurrences of both terms. However, we can identify the problem of disambiguation. In Acts 13:7 we find a different Paul tagged with the same Greek form: “Sergius Paulus, a man of intelligence, who summoned Barnabas and Saul”. Next, we can study the co-occurrences of Paul and Peter (strong:G4074) – there are none, although scholars identify a connection between both. Thus, the second issue can be identified by information which is only stated implicitly.

However, we might try to find connections identifying major groups of actors. Thus, we might either consider disciples (strong:03101) or apostles (strong:00652), Peter should belong to both. For the first group we find four, for the last group three co-occurrences. Again, these terms are not clearly defined. For example, in Acts 19:1 we see that Paul finds disciples in Ephesus which are not related to the disciples in Jerusalem. In Acts 14:14 Barnabas and Paul are called apostles, and not the initial twelve. Thus, before continuing to solve disambiguation by hierarchies of actors, we do not only need its data but also information on how these terms are used.

Yet another unresolved problem is the naming of actors in scenes. A verse-based co-occurrence will fail on actors mentioned at different occasions. Although promising results were described for dramas, see [26], it is not clear if they generalize. In addition, for biblical texts we might rely on the traditional pericopes, but again it is questionable if they precisely represent the stage occurrence of actors.

To sum up, even the co-presence of actors is difficult to extract using co-occurrence, since often we do not have detailed information about every particular actor in a narrative. Thus, applying methods to compute the figure configuration or the narrative connection between two actors remains challenging.

## 5. Discussion and Outlook

This paper has described and analyzed a first naive approach towards the automated generation of social networks on narrative texts. It comprises the automated linking of actors and other information, e.g. spatiality, from a previously defined list, and we analyzed how co-occurrence can be used to generate these networks or how it might improve the linking of data in manual exegetical work. In this article, we focused on biblical texts. Here, linguistic resources are already available, and thus we did not consider NLP methods. Even though, the performance of this naive approach turned out to be rather poor.

Our analysis of performance reveals some questions and also possible further improvement. First, we need to consider a hierarchy or taxonomy of actors and groups to tackle some challenges of implicit data. However, this will not solve all problems, since actors might be named in different places, which leads to the problem of scene-detection.

Second, we need to investigate on name disambiguation. Finally, it is worth to consider the results of this method within the framework of exegesis. Bourgeois et al. [25, 4] stated that with co-occurrence “it is impossible to extract a meaningful information”. The results of this work underline this. However, they might support scholars working on the text. Working with historical and narrative texts brings several challenges. Thus, we might also consider the improvement of tools to support the scholars and exegetes with feedback of their data to novel AI approaches. This might also help for disciplines where pre-curated texts are not available.

While our naive implementation is both working and generic, it is still very early work on an issue which needs more attention. We hope that it will also highlight the importance of more interdisciplinary research in this field.

## References

- [1] A. Collar, *Religious Networks in the Roman Empire*, University Press, Cambridge, 2013.
- [2] D. C. Duling, Paul’s aegean network: The strength of strong ties, *Biblical Theology Bulletin* 43 (2013) 135–154.
- [3] I. Czachesz, Women, Charity and Mobility in Early Christianity: Weak Links and the Historical Transformation of Religions, in: I. Czachesz, T. Biró (Eds.), *Changing Minds. Religion and Cognition Through the Ages*, Peeters, Leuven, 2011, pp. 129–154.
- [4] L. White, *Semeia 56: Social Networks in the Early Christian Environment*, Society of Biblical Literature, Atlanta, 1992.
- [5] P. Van Nuffelen, Religious Networks, *The Classical Review* 65 (2015) 224–226.
- [6] C. Rollinger, Prolegomena. problems and perspectives of historical network research and ancient history, *Journal of Historical Network Research* (2020) 1–35.
- [7] A. Cornils, *Vom Geist Gottes erzählen: Analysen zur Apostelgeschichte*, Francke, Tübingen, 2006.
- [8] J. Dörpinghaus, Die soziale Netzwerkanalyse: Neue Perspektiven für die Auslegung biblischer Texte?, *Biblich erneuerte Theologie* (2021).
- [9] J. Dörpinghaus, Computergestützte Verfahren für die Narrative Exegese, *Biblich erneuerte Theologie* (2022).



- [10] M. Reitmayer, C. Marx, Netzwerkansätze in der Geschichtswissenschaft, in: C. Stegbauer, R. Häußling (Eds.), *Handbuch Netzwerkforschung*, VS Verlag für Sozialwissenschaften, Wiesbaden, 2010, pp. 869–880.
- [11] J. van de Kamp, Übersetzungen von Erbauungsliteratur und die Rolle von Netzwerken am Ende des 17. Jahrhunderts, *Beiträge zur historischen Theologie*, Mohr Siebeck, Tübingen, 2020.
- [12] M. B. Thompson, The Holy Internet: Communication Between Churches in the First Christian Generation, in: R. Bauckham (Ed.), *Gospels for All Christians*, Bloomsbury Academic, London, 1998, pp. 49–70.
- [13] D. C. Duling, The Jesus Movement and Social Network Analysis (Part I: The Spatial Network), *Biblical Theology Bulletin* 29 (1999) 156–175.
- [14] D. C. Duling, The Jesus Movement and Social Network Analysis (Part II. The Social Network), *Biblical Theology Bulletin: A Journal of Bible and Theology* 30 (2000) 3–14.
- [15] J. M. McClure, Introducing Jesus’s social network: Support, conflict, and compassion, *Interdisciplinary Journal of Research on Religion* (2016).
- [16] J. M. McClure, The structure of Jesus’s social network: Subgroups, blockmodeling, and balance., *Interdisciplinary Journal of Research on Religion* (2018).
- [17] J. M. McClure, Jesus’s social network and the four gospels: Exploring the relational dynamics of the gospels using social network analysis, *Biblical Theology Bulletin* 50 (2020) 35–53.
- [18] J. Resseguie, *Narrative Criticism of the New Testament: An Introduction*, Baker Publishing Group, Grand Rapids, 2005.
- [19] S. Finnern, J. Rüggeheimer, *Methoden der neutestamentlichen Exegese : eine Einführung für Studium und Lehre*, UTB für Wissenschaft : Uni-Taschenbücher, UTB GmbH, 2016.
- [20] D. Elson, N. Dames, K. McKeown, Extracting social networks from literary fiction, in: *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 138–147.
- [21] N. Wiedmer, J. Pagel, N. Reiter, Romeo, Freund des Mercutio: Semi-Automatische Extraktion von Beziehungen zwischen dramatischen Figuren., in: *DHd*, 2020.
- [22] J. Dörpinghaus, Soziale Netzwerke im frühen Christentum nach der Darstellung in Apg 1-12, 2020. Available at <http://uir.unisa.ac.za/handle/10500/26609>.
- [23] J. Dörpinghaus, C. Düing, Automated creation of parallel bible corpora with cross-lingual semantic concordance, in: *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)*, IEEE, 2021, pp. 111–114.
- [24] W. Martinez, Au-delà de la cooccurrence binaire... poly-cooccurrences et trames de cooccurrence, *Corpus* 11 (2012). URL: <https://doi.org/10.4000/corpus.2262>.
- [25] N. Bourgeois, M. Cottrell, S. Lamassé, M. Olteanu, Search for meaning through the study of co-occurrences in texts, in: *International work-conference on artificial neural networks*, Springer, 2015, pp. 578–591.
- [26] J. Pagel, N. Sihag, N. Reiter, Predicting structural elements in German drama, *Proceedings* <http://ceur-ws.org> ISSN 1613 (2021) 0073.