EVINA STEIN

Parallel Glosses, Shared Glosses, and Gloss Clustering

Can Network-Based Approach Help Us to Understand Organic Corpora of Glosses?

Journal of Historical Network Research 9 (2023) 36–100

Keywords co-occurrence networks, medieval Latin glosses, manuscript studies, gloss parallelism, shared manuscript transmission

Abstract Glossing was an important element of medieval western manuscript culture. However, glosses are notoriously difficult to analyze because of their triviality, fluid nature, heterogeneity of origin, complex transmission histories, and anonymity. Traditional scholarly approaches such as close reading and the genealogical method often do not produce satisfactory results, especially in the case of gloss corpora that are highly organic, i.e., display the traits listed above to a significant degree. This article outlines a method for analyzing the organic corpora of glosses based on their treatment as networks. The theoretical model for the proposed method is the co-occurrence network, a network model in which relationships between entities (nodes) are established based on certain shared properties or constituent elements (edges). In the case of corpora of glosses, glossed manuscripts are assumed as nodes, and the glosses that specific manuscripts have in common constitute the edges between them. Since gloss parallelism can arise through different processes, including randomness, the article describes two strategies that reduce such noise so that the transmission of glosses can be effectively examined. The method is demonstrated on a representative corpus – the early medieval glosses to the first book of the *Etymologiae* of Isidore of Seville.

I. Introduction

Glossing (a term used in this article interchangeably with annotation) represented an important aspect of many pre-modern written cultures, including in Europe before the advent of print. In the medieval Latin-writing world, handwritten texts, copied in the writing block of the manuscripts (the black space), were commonly equipped in the margins, between the lines and columns and in other spaces left blank on the page (the white space) with enriching information – commentary, explanatory vocabulary, grammatical and stylistic remarks, translation to other languages, diagrams, cross-references, and critical remarks about the text's quality and veracity (Fig. 1).[1]

Medieval western annotation has traditionally been of interest to various scholars: linguists, who have found it a valuable source of information on the development of various European languages; philologists, who edited and analyzed the most important commentaries, glossing traditions and glossaries in manners similar to how they treated other historical and literary sources; and historians of intellectual life, who studied specific commentaries, glossing traditions and glossaries in their historical and social contexts. Only in the last decades have we seen a growing interest in medieval western annotation as a phenomenon in its own right, its use as a source for the understanding of medieval western culture more broadly, and its consideration in the global context of annotation cultures of other regions and periods. As a result of this broadening of horizons, a development that owes much to the increasing permeation of Big Data approaches to Humanities, the advent of digitization and computer-assisted methods, and the re-envisaging of annotations as data rather than as a traditional historical, literary or linguistic source, it has been recognized that medieval western annotations could contribute to research questions for which they had not been traditionally exploited.

In this article, I explore such a novel direction in research, looking at how the study of medieval western annotations could benefit from the application of a network-based approach. I hope to demonstrate how this approach can open new avenues to answer long-standing questions about glossing and provide us

Acknowledgements: Data used in this article were gathered as part of the *Innovating Knowledge* project funded by the Dutch Research Organization (NWO) under the grant agreement VENI 275-50-016 in 2018–2021. The author of this study would like to thank Siamak Taati and Sara Najem for their valuable advice about network theory, Peter Boot for helping with various aspects of data processing, and Bernhard Bauer for his comments on an earlier draft of the article.

Corresponding author: Evina Stein, Huygens Institute, Dutch Academy of Arts and Sciences (KNAW), Amsterdam. evina.steinova@gmail.com.

1 For an introduction to medieval western annotation, see Holtz, "Glosse e commenti"; Tura, "Essai sur les *marginalia*"; Schiegg, *Frühmittelalterliche Glossen*.

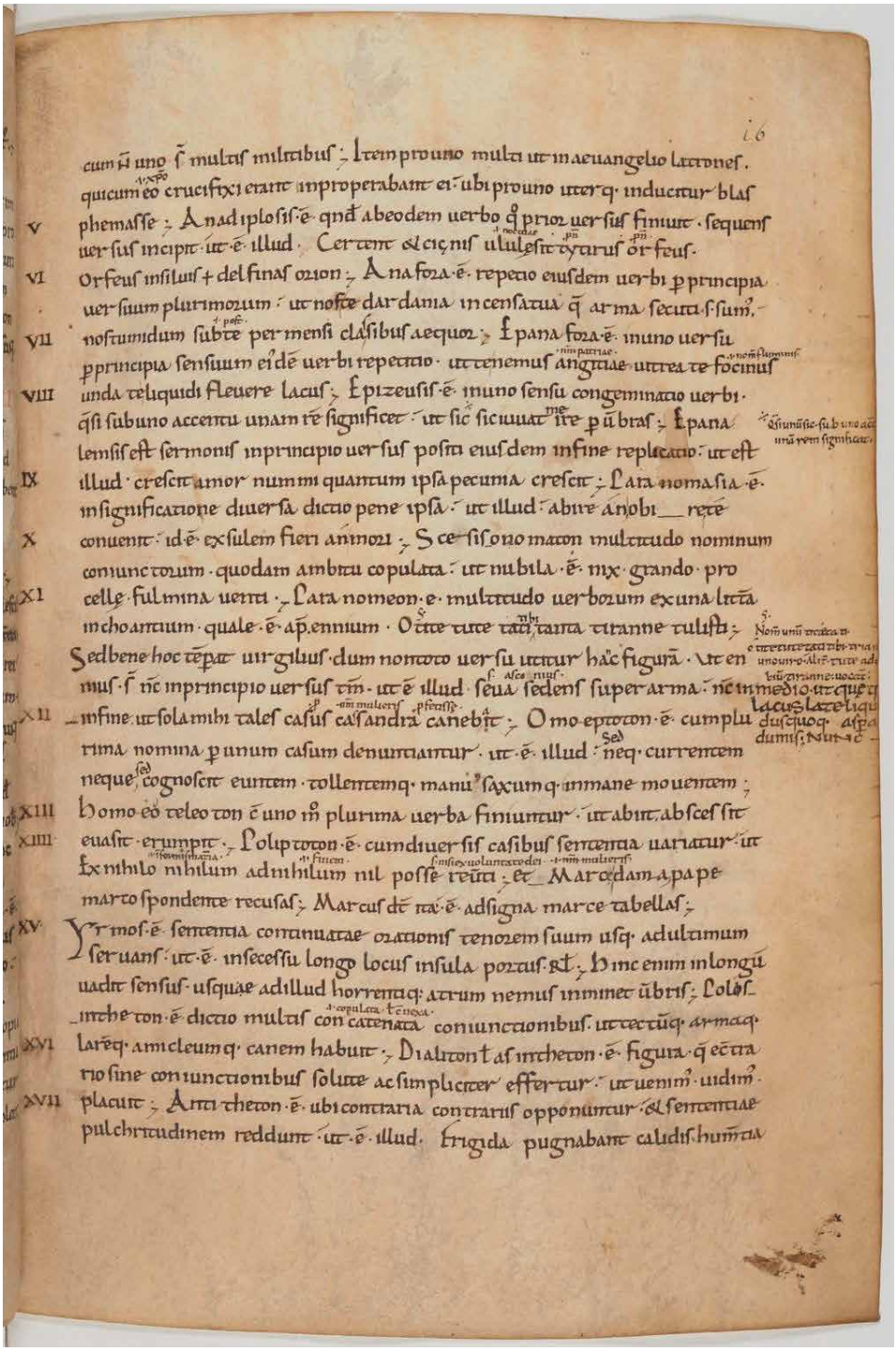

Fig. 1 An image of a glossed medieval western manuscript, Paris, Bibliothèque nationale de France, Latin 7585, fol. 16r. Source: Gallica, at https://gallica.bnf.fr/ark:/12148/btv1b10542288m/f35.item.r=%22Latin%207585%22.

with a means to overcome some of the well-known problems that scholars typically face. The new avenues relate to the how's, why's and what's of glossing: How were glosses produced and transmitted in the medieval Latin-writing world? Were they usually a result of spontaneous inspiration in response to momentary stimuli? Or were they rather handed down, exchanged, and collected?[2] In this regard, what is the significance of the same glosses recurring in different annotated manuscripts? How common is such gloss parallelism, and to what extent does it reflect transmission, as opposed to other historical processes, or chance? What does it suggest about the circulation of glosses in the medieval Latin-writing world? To tackle these questions, we can look at the patterns of gloss parallelism between manuscripts to establish their mutual relationships and examine the properties of networks constructed in this manner. The networks we can construct using this principle capture specific, intrinsic, dynamic aspects of glossing that are difficult to examine by other means, especially as extrinsic evidence about the production and circulation of glosses is often scarce or non-existent.[3]

As for the well-known problems in the study of medieval glossing, scholars need to tackle what may be termed their triviality. Glosses often amount to no more than a single phrase, word or even a syllable or a letter.[4] As a consequence, we cannot assume that multiple occurrences of a trivial gloss signal transmission, as would be the case with non-trivial glosses.[5] Moreover, the collections in which glosses typically survive are notoriously flexible and fluid, lacking the degree of coherence and sequentiality that define a typical text.[6] As a result, scholars may

2 Lapidge, "The Evidence of Latin Glosses"; Wieland, "The Glossed Manuscript"; Teeuwen, "Marginal Scholarship: Rethinking the Function of Latin Glosses in Early Medieval Manuscripts"; Teeuwen, "Writing in the Blank Space of Manuscripts."

3 This network-based approach is partially inspired by earlier attempts at employing network visualization to express the relationship between annotated manuscripts and explore gloss parallelism by Bernhard Bauer; see Bauer, "The interconnections of St Gall, Stiftsbibliothek, MS 251 with the Celtic Bede manuscripts"; Bauer, "The Celtic Parallel Glosses on Bede's 'De Natura Rerum'"; especially Bauer, "Venezia, Biblioteca Marciana, Zanetti Lat. 349. An Isolated Manuscript ?"

4 In the demonstrative corpus introduced in section 3, for example, the average length of a gloss is 2.6 words, and 45% of glosses are constituted by a single word. If opening formulas common to Latin glosses (see footnote 13) are discounted, the average length of a gloss in this corpus drops to 2.2 words, and 60% of glosses are constituted by a single word. See also Wieland, *The Latin Glosses on Arator and Prudentius in Cambridge University Library, MS Gg. 5.35,* 8; Nievergelt, "Glossen aus einem einzigen Buchstaben."

5 This observation is a variation on the well-known principle of indicative errors in genealogical textual criticism, which is explained in Chiesa, "The Genealogical Method: Principles and Practice," 79–80; Palumbo, "The Genealogical Method: Criticism and Controversy," 102–5.

6 See Teeuwen, "The Impossible Task of Editing a Ninth-Century Commentary," 197–200; Teeuwen, "Writing in the Blank Space of Manuscripts." They can be described as text colonies, using the terminology devised by the linguist Michael Hoey; Hoey, *Textual Interaction,* 74–76.

miss important connections within a corpus of glosses. Both traits of glossing pose a serious challenge in so far as we want to examine them with scholarly methods of traditional textual and historical scholarship (e.g., close reading).[7] As a result, certain types of glosses tend to be overlooked and understudied, while overconfident application of these methods may lead to inaccurate or misleading conclusions about other kinds of material. However, provided that we observe specific precautions, networks can be constructed even by relying on trivial glosses that do not form well-defined sequences. The network approach can, therefore, bypass some of the stumbling blocks of the research of medieval western glossing.

The questions articulated in this introduction are not fully resolved in this article. Instead, the present contribution aims to outline a particular methodology for analyzing medieval western glosses, demonstrate its utility on a representative corpus of glosses, and provide examples of network-driven analysis that can answer specific research questions. This article is, thus, primarily an invitation to: a) further develop a network-based approach to the study of glossing; b) apply it to different corpora of material; and c) test its usefulness. The proposed method is purposefully presented with a minimum of mathematical formalism and coding so that it is as accessible as possible to humanities scholars.

The article is divided into eight sections. The three sections following this introduction (section 1) provide the essential background for the network-based approach to annotated manuscripts. Section 2 defines concepts essential for the network-based approach to glossing and data pre-processing. Section 3 introduces a dataset used in this article to demonstrate this approach on a real-world corpus of glosses from medieval Europe. Section 4 outlines the general method used to construct a specific kind of network, namely the co-occurrence network, which can be used to harness gloss parallelism for research purposes. Sections 5 and 6 represent the analytical core of this article. In section 5, I describe and analyze several co-occurrence networks constructed from the data provided in section 3. In section 6, a selected network from section 5 is visualized and inspected in the light of extrinsic evidence. Finally, section 7 addresses the potential and limitations of the method, and section 8 presents the most important conclusions of this study.

2. Concepts and definitions

In this article, the term 'collection of annotations/glosses' is applied to manuscripts (e.g., a collection of annotations in Leiden VLF 48 or St. Gallen 904), while the term 'corpus of annotations/glosses' is used in connection to texts (e.g., a cor-

7 Other problems posed by glosses are described in O'Sullivan, "Problems in Editing Glosses: A Case Study of Carolingian Glosses on Martianus Capella."

pus of glosses to the Psalms, Virgil's *Aeneid,* or Priscian's *Institutiones*). A corpus of annotations represents all known glosses to a specific text. It is typically assembled from multiple collections of annotations found in manuscripts – its witnesses. Every gloss has two elements: a lemma (pl. lemmata), which corresponds to a specific word, phrase, or other textual unit in the black space, anchoring the gloss in a substrate (text or manuscript); and a body, in which enriching information is provided, usually in the white space. When the term gloss is used below it designates both elements, or the body of a gloss if it is explicitly distinguished from a lemma.

A gloss is the basic building block of a collection and a corpus of glosses, which may count tens, hundreds, or thousands of annotations. Although each gloss may be considered a micro-text as far as it is textually self-sufficient and can be added, removed, altered, and its position changed, glosses in a manuscript collection appear in a chain, i.e., we can state which gloss precedes or follows another and order them based on the sequence of folia and lines. However, without the support of a manuscript substrate, for example, if we compare glosses from different manuscripts or constitute a corpus, they disassemble into an unordered pool. The sequence of glosses in a collection may be relevant for certain types of research. However, the method described below is insensitive to it. The corpus and collections of glosses are, therefore, treated as pools, i.e., when a particular collection of annotations is discussed, the order in which glosses appear within it is taken into consideration only to a minimal degree.[8]

2.1 Systematic versus organic glossing

Based on the character of annotations in a collection or a corpus, it is useful to distinguish systematic from organic glossing. Systematic glossing can be defined as a programmatic annotation carried by one agent (a single individual or a group) with the intention of coherently engaging with a specific text, often meaning glossing it in its entirety, and therefore extensively.[9] Scholars frequently use

8 The 'minimal degree' applies here to glosses to identical lemmata appearing in different chapters of the annotated text, i.e., further apart than glosses to identical lemmata appearing within a single chapter. In theory, a researcher can encounter the same lemma-gloss pair in different chapters, as the text could contain the same words in multiple chapters, and these could attract the same glosses. However, these are not considered instances of gloss parallelism in this study.

9 Examples of medieval systematic glossing in the Latin-writing world include the ninth-century commentary on Martianus Capella's *De nuptiis Philologiae et Mercurii* by John the Scot, the twelfth-century commentaries of Anselm of Laon on the Gospel of John, and the coeval commentary on the Code of Justinian by the jurist Accursius. On these commentaries, see Jeauneau, "Le Commentaire érigénien sur Martianus Capella (De nuptiis, Book I)"; Rossi, *Atti del Convegno internazionale di studi Accursiani;* Andrée, "Anselm of Laon Unveiled."

terms such as commentaries, commentary traditions, and *scholia* to recognize the systematic nature of certain corpora of glosses and attribute them to specific individuals (authors) or groups (circles).[10] However, glosses were often inserted into manuscripts in a non-systematic manner, on an *ad hoc* basis, and perhaps even spontaneously, responding to the immediate needs and concerns of their makers rather than reflecting a program. As such, they do not form coherent collections nor provide a structured exposition, come from many different contexts of origin, and resulted from uncontrolled accumulation or growth not overseen by a specific agent. To distinguish these from the systematic collections of glosses and commentaries, I call them organic. A corpus or collection of annotations may possess a mixed set of traits, as it arose through both organic growth and systematic composition and compilation. While we can thus employ the designations systematic or organic for certain corpora of glosses that display very clear traits of one or the other type, it is more accurate to talk about the extent of organicity or systematicity in a corpus or a collection of glosses.

Traditionally, scholars have paid more attention to systematic than organic glosses due to the former's greater prominence among source material, their perceived higher aesthetic, literary and historical value, because these corpora better fitted the traditional notions of textuality and authorship, and due to the suitability of traditional approaches (e.g., genealogical editing and close reading) for their analysis. However, organic glossing may have been more prevalent in medieval Europe and thus more characteristic of medieval western annotation practices, particularly during certain periods. While the network-based approach may produce relevant results when applied to highly systematic corpora of glosses, these corpora tend to respond well to traditional methods, and thus the network-based approach may serve as a useful complement to these methods, although it is unlikely to be a scholar's primary option. This article is rather concerned with glossing that is organic to such a degree that its lack of coherence, heterogeneity of origin, purpose and language, multilayered character, and fluidity allow for a limited deployment of traditional methods.

2.2 Isolated, parallel and shared glosses

For the purposes of the network-based approach, a gloss corpus consists of two types of glosses. Certain glosses appear in it only once. I shall call these isolated glosses, and deal with them only marginally since they do not allow us to postulate a relationship between manuscripts. Other glosses feature in a corpus more than once since they appear in several of its manuscript witnesses. For example, the gloss *significat* is attached to the lemma *pingit* in two manuscripts of Bede's *De temporum ratione,* studied by Pierre-Yves Lambert and Bernhard Bauer

10 On this terminology, see, for example, Teeuwen, "Writing in the Blank Space of Manuscripts," 13.

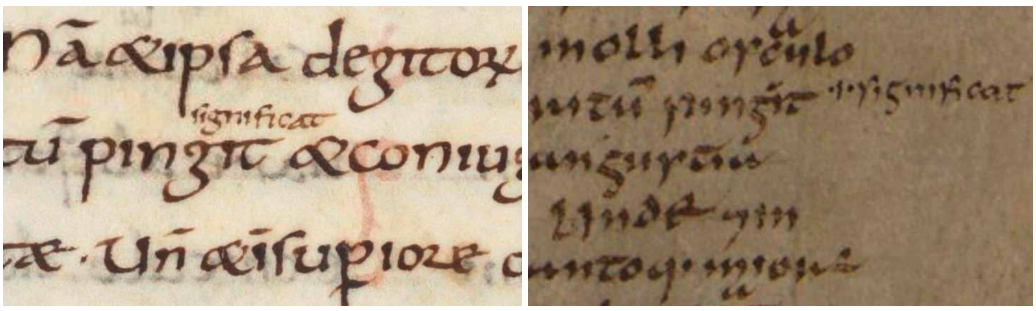

Fig. 2 A parallel gloss *pingit.significat* in two manuscripts of Bede's *De temporum ratione.* Left: Angers, Bibliothèque municipale, 477, fol. 45v (source: BVMM, at https://bvmm.irht.cnrs.fr/iiif/1097/canvas/canvas-375525/view). Right: Karlsruhe, Badische Landesbibliothek, Aug. Perg. 167, fol. 24r (source: Badische Landesbibliothek, at https://digital.blb-karlsruhe.de/id/20736).

(Fig. 2).[11] Such glosses establish a relationship between manuscripts that is at the heart of co-occurrence networks. I shall call these parallel glosses.[12]

Importantly, labelling a gloss as parallel does not provide an explanation as to why it recurs within a corpus. The term merely signals that due to the extent of their philological similarity, two or more glosses are judged as manifestations of the same philological entity.[13] Parallel glosses may be parallel due to transmission, that is, they reflect a relationship between manuscripts that implies contact between historical individuals, groups, and institutions. However, the parallelism illustrated by Fig. 2 could also result from processes other than transmission, especially if it concerns individual trivial glosses. Specifically, a trivial gloss that represents a logically derived explanation of a lemma (e.g., a synonym, a translation, or an etymology) could have been coined independently by multiple

11 Lambert, "Les commentaires celtiques a Bède le vénérable"; Lambert, "Les commentaires celtiques à Bède le vénérable"; Bauer, "The interconnections of St Gall, Stiftsbibliothek, MS 251 with the Celtic Bede manuscripts," 34.

12 I borrow this term from Bauer, "Venezia, Biblioteca Marciana, Zanetti Lat. 349. An Isolated Manuscript ?," 91.

13 The issue of philological similarity would deserve theoretical reflection, which is not possible in this article. Here, it can be noted that traditional textual scholarship also operates with a notion of similarity in assessing indicative errors, variant readings, and text versions as the same or similar. When considering glosses as parallel in this study, I ignore spelling variation, manner of abbreviation, morphological form (e.g., whether the gloss accepts the case, number, etc. of the lemma or not), word order, the presence of introductory phrases characteristic to Latin glosses that do not affect the meaning of a gloss (e.g., *id est, hoc est, scilicet, sicut, quasi, vel*), textual corruptions as a result of a mechanical error in a single witness, and omission of or variation in minor elements that do not alter the meaning of a gloss (e.g., prepositions and prefixes).

annotators, since the lexicon of a language provided them with limited options, they were glossing the same text and therefore the same lemmata, and they likely received similar training and had similar resources at their disposal. Furthermore, in a scenario described in this article, gloss parallelism can be expected to occur to an extent even randomly, not mirroring any historical process but rather arising from the method itself.

Since gloss parallelism is central to the method described in this article, it is essential to distinguish parallel glosses that reflect transmission from those that are the result of what I shall call spontaneous composition and random gloss parallelism. For this reason, I introduce a specifying category: the shared gloss. A shared gloss can be defined as a subtype of parallel glosses, the similarity of which can be explained as a consequence of transmission. Distinguishing parallel glosses from shared glosses is difficult in a real-life research context, especially in organic corpora of glosses, for the transmission of which we typically lack sufficient extrinsic evidence. A researcher can, nevertheless, establish philological criteria to assess parallel glosses as shared. In this study, I use a system of three ranks, following the principle that the more particular a gloss is, the less likely it is that it arose independently multiple times.[14] Beyond a certain degree of peculiarity, a scholar can consider a gloss monogenetic, i.e., having originated only once, and therefore assume that all its manifestations in a gloss corpus are instances of transmission. By contrast, the more generic information a gloss provides, the more likely that it is polygenetic, i.e., having originated independently multiple times, and therefore cannot be assumed to have been evidence for transmission.[15] In this particularity ranking, the lowest rank, 1, is accorded to generic parallel glosses that are possibly polygenetic, and are thus treated as instances of spontaneous composition in the following sections.[16] The intermediate rank, 2, is used

14 The following general criteria are used in this study to assess to what degree a parallel gloss is shared:
a) number of identical words appearing in the same or similar sequence in a gloss (here at least four);
b) the presence of the gloss in a significant number of witnesses (here at least five);
c) the gloss is a citation from a known source;
d) the presence of idiosyncratic, unusual, or erroneous information;
e) the presence of textual errors, corruptions, or paleographic features that are indicative of copying;
f) the gloss depends on an error in the substrate text but also appears in witnesses without this error;
g) if multiple parallel glosses form logically coherent sets within the text; and
h) in the case of lemmata that attracted many different isolated glosses if gloss parallelism is observed.

15 Monogenicity and polygenicity are discussed in Trovato, "Neo-Lachmannism: A New Synthesis?"; Conti, "A Typology of Variation and Error," 243–45.

16 A common example of glosses ranked 1 are glosses that expand an obvious ellipsis in the text. In the corpus introduced in section 3, for example, the lemma *Hebraeorum lit-*

for glosses that cannot be determined by philological assessment, i.e., they may have been transmitted, but it cannot be ruled out that they emerged as a result of spontaneous composition.[17] Finally, the highest rank, 3, is assigned to glosses so particular that they can be treated as transmitted.[18] For the most part, I shall use the term parallel gloss in the following sections of this study, but if the term shared gloss is used, it refers specifically to glosses that are assumed to have been transmitted, i.e., those with rank 3. Using this ranking method, rather than classifying parallel glosses binarily as shared or not, will allow us to account for the complexity of the real-world data introduced in the following section and consider different scenarios of gloss parallelism and transmission.

2.3 Gloss sets and gloss clusters

As glosses behave as self-sufficient micro-texts, we can expect to encounter some that circulated individually (as we shall see, this is the case with some shared glosses analyzed below). In practice, however, it is more common to encounter glosses preserved in particular groups of manuscript witnesses as sets, i.e., to observe that certain parallel glosses always travel together as a unit. As with the case of parallel glosses, dissecting a gloss corpus into sets does not explain why we encounter them today in this form. Some of the gloss clustering in a corpus is due to transmission, as the sets correspond to textual units circulating in the Middle Ages.[19] However, a degree of clustering is also a natural result of gloss parallelism due to spontaneous composition and even randomness.

For this reason, I distinguish gloss sets (any batches of glosses that appear together in two or more witnesses in the corpus), from gloss clusters (sets that can

---

teras a Lege coepisse per Moysen ("The Hebrew letters [are believed] to have begun from the Law through Moses") in the third chapter of the first book of the *Etymologiae* is glossed with *dicimus* ("we claim") in two manuscripts.

17 A common example of glosses ranked 2 are synonyms and glosses that provide non-specific clarifying information about the name of a person or place, or the grammatical category of the lemma. In the corpus introduced in section 3, for example, the lemma *repertus* ("found") is glossed as *inventus* ("discovered") and the name of the mythological king Cadmus is glossed as *rex* ("a king") in the third chapter of the first book of the *Etymologiae*.

18 An excellent example of a gloss from the corpus introduced in section 3 assigned rank 3 is a gloss to the term *sicilicus,* a special orthographic sign used to mark the duplication of letters in Latin, in chapter 27 of the first book of the *Etymologiae.* This gloss reads: *et sicilicus quia in Sicilia inveniebatur primo* ("and it is called a *sicilicus* because it was first invented in Sicily"). As the name of *sicilicus* is derived from *sicilis* ("a sickle") rather than related to the island of Sicily, this imaginative etymologization, found in three manuscripts, should be consider highly peculiar and therefore monogenetic.

19 We know from the extrinsic evidence that medieval scribes usually copied glosses from manuscript to manuscript in batches; see Dionisotti, "On the Nature and Transmission of Latin Glossaries"; Godden and Jayatilaka, "Counting the Heads of the Hydra," 365.

be assumed to reflect gloss transmission in the Middle Ages). In this study, I define a gloss cluster as a set constituted by at least ten glosses, or only by glosses with rank 3. Importantly, unless we possess historical evidence that would allow us to reconstruct a specific historical unit of transmission fully, which is rarely the case, it is a scholarly reconstruction. In the form in which we can reconstruct them, even large clusters that doubtlessly reflect transmission can be affected by spontaneous composition and random parallelism, and may therefore contain glosses attached to a genuine historical core by chance. We must therefore bear in mind that clusters inform us about the general contours of transmission, i.e., they attest to it and allow us to identify manuscripts containing transmitted material, but they do not provide us with an exact picture, i.e., we cannot be sure that all parallel glosses in a cluster were transmitted. In the case of clusters containing glosses with all ranks, only the core of these clusters, constituted by glosses with rank 3, can be considered as certainly transmitted, while we must remain in doubt about the glosses with the lower ranks, 1 and 2. For this reason, the smallest and most generically-looking sets, in particular those constituted by only one or two parallel glosses with ranks 1 and 2, may be phantoms created by scholarly reconstruction.

3. Data

To demonstrate the practical utility of the network-based method, I select a single representative corpus of medieval annotations – the glosses to the first book of the *Etymologiae* of Isidore of Seville.[20] This corpus displays characteristic traits of medieval western glossing, including those that cause the most problems to scholars applying traditional methods. It is therefore ideally suited for testing the network-based approach described below.

The *Etymologiae,* produced in the first decades of the seventh century in Visigothic Spain by the bishop Isidore of Seville (d. 636), was the most important encyclopedic work of the western Middle Ages. That it survives today fully or in parts in at least 1,400 manuscripts copied between the seventh and the sixteenth centuries is a lasting testament to the popularity of this work. Many of these manuscripts are annotated. The highest intensity of annotation took place in the early Middle Ages (c. 600–1000 CE), a period in which the *Etymologiae* was the only widely available encyclopedia and served as the ultimate go-to for

20 This corpus is published online at: https://db.innovatingknowledge.nl/edition/#right-network. The underlying data can be downloaded as an Excel file from Zenodo: 10.5281/zenodo.5359401. Those wishing to use this data will note that the published dataset uses a slightly different particularity ranking scale, with four ranks and a broader clustering scheme including sets larger than five glosses among clusters as small clusters (see footnotes 23 and 27).

the Latin-writing world. Today, 74 manuscripts of this work that were annotated in the period from the eighth to the beginning of the thirteenth centuries are known, preserving slightly more than 7,000 glosses. This corpus is highly organic, being the work of many anonymous annotators separated by time, space, linguistic context, interest, and skill.[21]

The corpus glosses are unevenly distributed, both across the identified witnesses (from one to more than a thousand glosses in a manuscript) and the twenty topic-based books into which the *Etymologiae* is structured (from 42 to more than 4,000 glosses per book). Because of this disparity, only parts of this larger corpus are suitable for network analysis. More specifically, one of the twenty books, the first book dedicated to the ancient and medieval discipline of grammar (*grammatica*), preserves most of the known glosses: 4,286 (i.e., ~ 62% of the entire corpus) and is annotated in the most manuscripts (54 of the 74 witnesses, i.e., 73%). These glosses to the first book of the *Etymologiae* form the main dataset for this study.

An overview of the 54 manuscripts that preserve the 4,286 glosses analyzed below, with the latter's full shelfmarks, shortened labels referenced in this study, assumed periods and places/regions of glossing, and the numbers of all glosses and parallel glosses of different ranks, are provided in Appendix I.[22]

3.1 Parallel and shared glosses in the dataset

Of the 4,286 glosses that constitute the corpus of glosses to the first book of the *Etymologiae,* 2,554 are isolated, and 1,732 are parallel. If described as a minimum corpus of unique glosses, i.e., parallel glosses encountered in multiple witnesses are considered manifestations of the same entity, then the corpus consists of 3,279 glosses, of which 2,554 are isolated (i.e., feature in the corpus exactly once) and 725 are parallel (i.e., feature in the corpus twice or more). Tab. 1 presents the distribution of parallel glosses with various ranks in the minimum corpus based on the number of manuscripts they appear in.[23] As can be gleaned from it, most parallel glosses in the corpus adopted for this study were assigned rank 2 (417, ~ 58% of parallel glosses) and appear in two manuscripts (547, ~ 75% of parallel glosses). Nevertheless, approximately 31% of parallel glosses from this data-

21 The identified annotated manuscripts and their historical context of origin are described in detail in Steinová, "Annotation of the Etymologiae of Isidore of Seville in Its Early Medieval Context."

22 As only 47 of these manuscripts contain any parallel glosses, seven manuscripts (highlighted in grey) are included in the Appendix only for the sake of completeness.

23 In this article, parallel glosses assigned ranks 3 and 4 in the original dataset have the rank of 3.

no. of mss. in which a gloss appears	all parallel glosses	rank 1	rank 2	rank 3
in two mss.	547	72	304	171
in three mss.	110	4	75	31
in four mss.	35	3	22	10
in five mss.	24	1	12	11
in six mss.	5	0	2	3
in eight mss.	2	1	1	0
in nine mss.	1	0	1	0
in ten mss.	1	0	0	1
Total	725	81	417	227

Tab. 1 Distribution of glosses in the minimum corpus based on the extent of their co-occurrence and particularity rank.

set were assigned rank 3, and up to ten manuscripts from the corpus feature the same parallel gloss.

The high number of isolated glosses and the limited extent of gloss parallelism, rarely extending beyond three manuscripts, are not the features of the corpus, nor do they inform us about the character of medieval western glossing. They are very likely a consequence of the loss of annotated manuscripts from the Middle Ages.[24] If we had access to the corpus of all glosses to the first book of the *Etymologiae* generated in the Middle Ages, as opposed to only those that are preserved by surviving manuscripts, we would likely see that many glosses that appear isolated were in fact parallel, and some of the parallel glosses had been shared by more manuscripts than is the case in the present-day corpus. For this reason, the reconstructions of relationships between the surviving annotated manuscripts of the *Etymologiae* must be understood as representing the best achievable minimalistic result, rather than faithfully corresponding to historical reality. Moreover, the links created between manuscripts by gloss parallelism should not be, as a rule, understood to reflect direct relationships between surviving witnesses

24 In this respect, it may be compared to the bifidity of stemmata in traditional textual scholarship; see Guidi and Trovato, "Sugli stemmi bipartiti. Decimazione, asimmetria e calcolo delle probabilità." On the extent of the loss of manuscripts from the Middle Ages, see Buringh, *Medieval Manuscript Production in the Latin West,* 179–252.

or the transfer of material from one witness to another, but rather, as in a stemma, indirect relationships facilitated by lost intermediaries.[25]

3.2 Gloss sets and gloss clusters in the dataset

The 725 parallel glosses can be split into 228 sets that are unique to anything between two and ten manuscripts. Of these 228 sets, 142 (62.3%) consist of a single gloss, twenty (8.7%) of two glosses, 26 (11.5%) of two to nine glosses, and 40 (17.5%) of ten or more glosses. The forty sets of ten or more glosses can be sorted into twelve clusters of 11 to 157 glosses, labelled by letters of the alphabet to distinguish them.[26] In addition, one set of seven glosses and nine sets of one to three glosses can be recognized as clusters following the principle that all their constituent glosses have a rank of 3. The latter sets, which illustrate that glosses could circulate independently, are labelled as C and subdivided into seven micro-clusters. These twenty clusters consist, on average, of eight or nine glosses and have an average rank of 2.87 (i.e., leaning strongly towards being particular rather than generic). An overview of the twenty clusters is provided in Tab. 2. The remaining 178 sets of one to nine glosses with lower ranks of 1 and 2 are assigned the generic label X, so that they can be filtered out from the following network analysis and visualization. These unassigned sets consist, on average, of one or two glosses with an average rank of 1.96 (i.e., leaning towards generic rather than particular).[27]

25 Compare with Roelli, "Definition of Stemma and Archetype," 213.

26 The discrepancy between the number of sets with ten or more parallel glosses and the number of established clusters is due to the consideration of extrinsic evidence. For example, Paris7490 today contains only chapters 5–17 of the first book of the *Etymologiae,* and Orleans296 chapters 21–44. Nevertheless, the analysis of glossing hands, layout and ruling pattern, and context of preservation suggest that the two manuscripts are closely related, and their collections of glosses may represent two parts of a single whole (e.g., two damaged codicological units from the same glossing circle); Steinová, "Annotation of the Etymologiae of Isidore of Seville in Its Early Medieval Context," 24. For this reason, Paris7490 and Orleans296 are assigned to the same clusters as Orleans296/Paris7490, even though they are treated as separate nodes in the analytical sections of this article.

27 The original dataset used for this study distinguishes small clusters of five to ten glosses (labelled as H, R, and T–Z), which are treated as unassigned sets (X) in this article and clusters distinguished by numerals (e.g., F1 and F2), which corresponds to the distinction between parallelism with Orleans296 and Paris7490 (see the previous footnote). The distinction of unassigned sets X1 and X2, introduced purely to distinguish different sets of parallel glosses to the same lemma, is also not maintained.

label	manuscripts that share the glosses from the cluster	no. of glosses	avg. rank	no. of mss.
Clusters (13)				
A	Hamilton689, Harley3941, MontpellierH53, Paris7585, Paris7670, Paris7671, Paris11278, Reims425, Reims426, Trier100, VLO41, VLF82	14	2.97	12
B	Reims426, VLO41	18	2.11	2
D	CesenaSXXI5, VeniceII46	11	3	2
E	IRHT342, CotCalAxv, GothaI147, Paris7585, Queen320	50	3	5
F	Harley3941, Orleans296/Paris7490, Reims426	157	2.29	4
G	IRHT342, CotCalAxv, Harley3941, Laon447, Paris7585, Queen320	30	2.4	6
I	Orleans296/Paris7490, VLO41	54	1.84	3
M	Paris7670, VLO41	17	2	2
N	Orleans296/Paris7490, Paris7670	29	2.06	3
O	Harley3941, Paris7670	13	2.31	2
P	RAH25, RAH76	17	2.12	2
Q	CotCalAxv, Harley3941, Paris7585, Paris11278	7	3	4
S	Orleans296/Paris7490, Reims426	21	1.93	3
micro-clusters (7)				
C1	Arundel129, Bern101, BrusselsII4856, Clm4541, Clm6250	1	3	5
C2	Hamilton689, MilanL99sup, VatLat5763	1	3	3
C3	IRHT342, Harley3941, Paris7490, VLF82, Wolfenbuttel64	1	3	5
C4	IRHT342, Bern101, Harley3941, Paris7559, Paris7671, Schaffhausen42, VLF82	3	3	7
C5	Clm4541, Clm6250, Laon447, Schaffhausen42	1	3	4
C6	Bologna797, Paris11278, Schaffhausen42	1	3	3
C7	IRHT342, GothaI147, Harley3941, Paris7559, Paris7585, Paris7670, Paris10293, Queen320, Schaffhausen42, Wolfenbuttel64	1	3	10

Tab. 2 Overview of gloss clusters in the corpus.

3.3 Historical context of the dataset

Clues put together based on manuscript evidence situate the glossing of the first book of the *Etymologiae* within the general contours of early medieval intellectual life. In the early Middle Ages, most of the intellectual production of the Latin-writing world, including glosses, originated in monastic and cathedral scriptoria, libraries, and schools. These religious centers, at least 650 of which have been documented and of which most if not all produced and used books, formed an interconnected network stretching across the Latin-writing world.[28] The circulation of glosses happened via this network through mechanisms that entailed written and oral transmission (e.g., the exchange of annotated manuscripts and personnel, and instruction).[29] The patterns of transmission of glosses to the first book of the *Etymologiae* does map onto the historical network of intellectual centers, although due to the loss of material from the Middle Ages, we can obtain only its faint echo.[30]

The manuscript evidence suggests that the glossing of the first book of the *Etymologiae* in the Carolingian environment was driven by the integration of this text into the grammatical curriculum starting from the end of the eighth century.[31] It thus appears to have been a response to the needs of school education, serving Carolingian schoolmasters and students. Many of the surviving annotated manuscripts of the first book of the *Etymologiae* reflect this purpose: they can be described as schoolbooks or instructional manuals and were produced and annotated during the ninth century in the modern region of northern France, the heart of the Carolingian empire. However, glosses are also found in manuscripts from Brittany, England, northern Italy, the German area, and Spain. Moreover, some of the annotated manuscripts, including those that preserve the richest collections of glosses, are books that were designed to sit on a lectern in a library, suggesting that while school study may have been an important stimulus for glossing, glosses nonetheless originated in and permeated other contexts of use. Overall, the extrinsic clues create the impression of a substantial circulation of the glosses to the first book of the *Etymologiae,* rather than the prevalence of spontaneous composition.

---

28 Ganz, "Book Production," 789; Contreni, "The Pursuit of Knowledge in Carolingian Europe," 127.

29 Teeuwen, "Marginal Scholarship: Rethinking the Function of Latin Glosses in Early Medieval Manuscripts," 30–32.

30 On these intellectual networks, see for example Moulin, "Paratextuelle Netzwerke."

31 Steinová, "Annotation of the Etymologiae of Isidore of Seville in Its Early Medieval Context," 19–29.

4. Method

4.1 A co-occurrence network as a model

The basic blueprint for networks constructed, described, and analyzed in the following sections may be called a co-occurrence network. A co-occurrence network is one in which similarity between entities, for example the sharing of properties or constituent elements, is used as a basis for establishing relationships between them.[32] Unlike social networks, the most common network model currently employed in historical network research, co-occurrence networks do not represent direct relationships facilitated by human interaction. Some of the concepts from social network analysis should, therefore, not be assumed to apply to them. Co-occurrence networks are instead suitable for exploring relationships and similarities between man-made objects, such as texts, written artifacts, or creative output (e.g., music and visual art).[33] In this regard, they resemble stemmata rather than social networks.[34]

This study is concerned with co-occurrence networks representing relationships between manuscript witnesses of a gloss corpus based on the patterns of gloss parallelism. In this model, manuscripts serve as nodes while parallel glosses supply the edges. Thus, if two manuscripts share a parallel gloss, their nodes are connected by an edge, if three manuscripts share it, all three are connected by edges, and if such a gloss appears in four manuscripts, all four are connected (Fig. 3). Because of this principle, a characteristic trait of co-occurrence networks is the presence of many locally complete sub-graphs.[35]

Since multiple glosses may be shared by two manuscripts, rather than plotting many parallel edges between two nodes, I provide each edge with a weight corresponding to the number of parallel glosses that form it. For example, the heaviest edge in the dataset chosen for this study, which connects Harley3941 and

32 A similar network model has been proposed in Valleriani et al., "The Emergence of Epistemic Communities in the Sphaera Corpus," 57–58. The co-occurrence network model can be considered a more generic version of the network of shared textual transmission developed in Fernández Riva, "Network Analysis of Medieval Manuscript Transmission. Basic Principles and Methods"; and explored in Kapitan, "Perspectives on Digital Catalogs and Textual Networks of Old Norse Literature." See also the networks of material culture explored in Peeples et al., "Analytical Challenges for the Application of Social Network Analysis in Archaeology," 65–67.

33 Compare with Brughmans, Collar, and Coward, "Network Perspectives on the Past," 11.

34 On stemmata as models and graphs, see Hoenen, "The Stemma as a Computational Model"; Roelli, "Definition of Stemma and Archetype."

35 The number of edges generated by a parallel gloss shared by N manuscripts can be calculated as $N(N-1)/2$. Thus, the parallel gloss shared by most manuscripts in this corpus, which is ten according to Tab. 1, generates 45 edges.

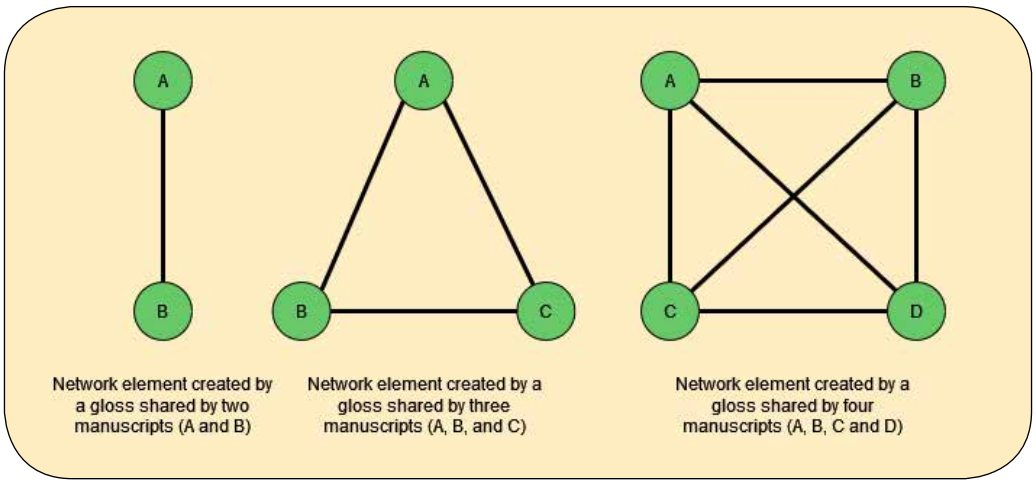

Fig. 3 The basic model for the network structure examined in this article. Illustrated here are the connections created by a parallel gloss shared by two, three, and four manuscripts. Produced with app.diagrams.net.

Orleans296, is constituted by 175 parallel glosses, and thus has a weight of 175. All network graphs described below are undirected.

4.2 Preparing the data

The data used for the construction of the co-occurrence networks was taken from a TEI-XML file containing a transcript of the glosses to the first book of the *Etymologiae,* produced in the context of preparing a digital scholarly edition of this gloss corpus.[36] As part of the encoding, each of the 4,000+ glosses in the XML file was equipped with attributes that indicate whether it was isolated or parallel, along with its particularity rank and cluster or set. This data was exported into an edge table suitable for network analysis using an XSL script.[37] The main edge table used in this study records the following information: a) labels of manuscript pairs sharing glosses (as source and target); b) cluster to which these edges belong (as cluster); c) the number of glosses of particular ranks constituting the edge (as rank 1, 2, and 3); and d) the total number of parallel glosses constituting the edge (as no. of glosses, see Tab. 3). The complete edge table has 417 rows, i.e., it corresponds to 417 edges between the 47 manuscripts containing at least one parallel gloss. It is provided in Appendix II.

36 This TEI-XML file is available at: https://github.com/HuygensING/isidore-glosses.
37 I would like to thank my colleague Peter Boot for writing this script.

Source	Target	Cluster	rank 1	rank 2	rank 3	no. of glosses
Harley3941	Orleans296	F	6	74	58	138
Orleans296	VLO41	I	3	35	0	38
Gotha1147	Paris7585	E	0	0	32	32
Harley3941	Paris7585	G	1	14	15	30
Harley3941	Orleans296	X	3	25	1	29
Harley3941	Reims426	X	2	19	3	24
Orleans296	Paris7670	N	5	18	0	23
Harley3941	Paris7490	F	0	10	9	19
Reims426	VLO41	B	1	14	3	18
RAH25	RAH76	P	0	15	2	17
Paris7670	VLO41	M	2	13	2	17
Harley3941	VLO41	X	1	12	3	16
Harley3941	Paris7670	X	1	14	1	16
Paris7585	Queen320	E	0	0	15	15
Orleans296	VLO41	X	1	13	1	15

Tab. 3 A segment of the complete edge table representing the co-occurrence network of parallel glosses in the studied corpus. Displayed are the top 15 rows ordered by the number of parallel glosses.

This edge table is complemented by a node table containing information about the 47 manuscripts containing parallel glosses to the first book of the *Etymologiae*. It was manually prepared by the author of this study and is also attached to this article in Appendix II. Its columns store the following information: a) manuscript label taken from Appendix I (as label); b) manuscript type (as type) based on whether the manuscript is a grammatical handbook containing only the first book of the *Etymologiae* (AI, 29%), a complete copy of the *Etymologiae* (BI, 67%), or a manuscript containing only excerpts from the first book of the *Etymologiae* (EXC, 4%); c) place of estimated glossing, represented as GPS coordinates (as latitude and longitude); and d) the number of parallel glosses found in the manuscript (as no. of parallel glosses, see Tab. 4). The GPS coordinates, manuscript type, and number of parallel glosses are used in the visualization plotted in section 6.

Fig. 4 represents a sample segment of a co-occurrence network of gloss parallelism between eight of the manuscripts studied here, constructed from the edge and node tables described above.

Label	type	latitude	longitude	no. of parallel glosses
Harley3941	BI	48.16667	–2.83333	305
Orleans296	AI	48.85341	2.3488	301
VLO41	AI	47.80281	2.31321	191
Paris7670	BI	48.85341	2.3488	130
Reims426	BI	49.25	4.03333	127
Paris7585	BI	51.27904	1.07992	116
Paris7490	AI	48.85341	2.3488	68
Paris7559	AI	48.85341	2.3488	42
Paris7671	AI	49.15964	5.3829	37
IRHT342	BI	48.16667	–2.83333	35
GothaI147	BI	48.16667	–2.83333	32
Paris11278	AI	43.71553	1.604	30
VLF82	BI	48.85395	2.33449	29
RAH25	BI	42.32962	–2.8722	28
RAH76	BI	40	–4	28

Tab. 4 A segment of the node table representing the nodes in the co-occurrence network of parallel glosses in the studied corpus. Displayed are the top 15 rows ordered by the number of parallel glosses.

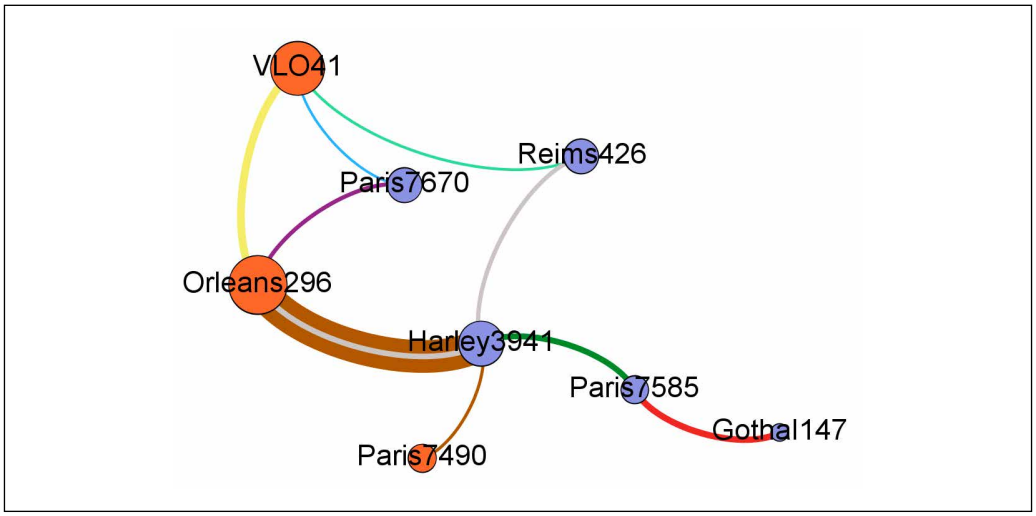

Fig. 4 A sample segment of a co-occurrence network of parallel glosses. The visualization displays manuscripts connected by edges with a weight larger than seventeen (i.e., it corresponds to the first nine lines of the edge table in Tab. 3). Edges are colored based on the cluster they represent (light green: B, dark red: E, brown: F, dark green: G, yellow: I, light blue: M, dark purple: N, and grey: X). The thickness of edges is proportional to their weight. The size of the nodes is proportional to the number of parallel glosses they contain. Parallel edges (here between Harley3941 and Orleans296) are overlaid. The visualization was created in Gephi with Yifan Hu layout.

4.3 Accounting for multiple scenarios

The combination of the particularity ranking and gloss clustering allows us to construct and examine several network scenarios based on different degrees of inclusivity of data (e.g., treating all parallel glosses as shared versus treating only glosses with rank 3 as shared), rather than having to represent the entire corpus with a single co-occurrence network. In this way, we can mitigate some of the issues stemming from the complexity of the real-world data and the limits of the scholarly reconstruction. The comparison of different scenarios yields insights into the stratigraphy of the corpus that would otherwise remain concealed. In the end, we can select one or more networks from the available scenarios that we find most suitable to answer our research questions.

The particularity ranking allows us to distinguish and isolate specific layers from the larger co-occurrence network of all parallel glosses, based on whether we consider them transmitted or not. The edge tables corresponding to individual layers can be derived from the main edge table by extracting and combining columns with glosses of specific ranks. We can work with up to six different scenarios: three for individual ranks (Rnk1, Rnk2, and Rnk3), two for a combination

of ranks (Rnk12 and Rnk23), and one representing all parallel glosses (Par). Gloss clustering helps us to distinguish those elements of the network that we can consider historical artifacts with certainty (clusters and micro-clusters) from those that may be mere noise (unassigned sets). We can use these to plot two types of networks: one in which all sets are included, irrespective of their weight and label (clustered), and one from which the sets labelled as X are removed (clustered-noX). We can also ignore the division into clusters and sets and construct co-occurrence networks, in which all glosses shared by two manuscripts establish an edge between them (unclustered). The two clustered network types differ from unclustered networks in that they are hypergraphs, i.e., two nodes can be connected by several parallel edges, since a pair of manuscripts can participate in multiple clusters.[38] The manuscript pair Harley3941-Paris7585, for example, features in clusters A, G and Q, as well as the micro-cluster C7 (see Tab. 2). Moreover, the values of the network properties of the two clustered network types can exceed the maximum values observable in unclustered networks. The edge tables for constructing clustered-noX networks can be derived from the main edge table by removing the rows assigned label X (180 rows). The edge tables for constructing the unclustered networks can be produced by contracting all rows with the same source and target (282 rows).

5. Analysis

By exploiting the particularity ranking and gloss clustering described in section 2, each of the three network types representing a different extent of gloss clustering (unclustered, clustered, and clustered-noX) can be paired with six network layers or layer combinations obtained through the particularity ranking (Rnk1, Rnk2, Rnk3, Rnk12, Rnk23, and Par). Thus, altogether we can construct eighteen co-occurrence networks from the data introduced in section 3. In this section, the main analytical part of this article, the network properties of these eighteen networks are examined to gain as complete a picture of the corpus as possible, probing its structure and dynamics by comparing various network types and layers and assessing their relative value. The following analysis does not require visualizing any of the eighteen networks.[39] Rather, the conclusions about the corpus are reached entirely from the network properties. At the end of this section, one of the eighteen networks is selected for visualization and detailed treatment in the following section.

The eighteen networks are labelled by a combination of a network layer and type in the following two sections, e.g., Rnk3-clustered-noX refers to a net-

38 Newman, *Networks,* 114–15.

39 The author of this paper, nevertheless, explored all eighteen network scenarios in Gephi to obtain some of their network properties.

Network		Rnk1	Rnk2	Rnk3	Rnk12	Rnk23	Par
Nodes	**Unclustered**	30	40	38	41	47	47
	Clustered	30	40	38	41	47	47
	clustered (no X)	7	12	35	12	35	35
Edges	**Unclustered**	87	176	160	201	262	282
	clustered	100	218	237	245	393	417
	clustered (no X)	13	23	172	24	180	180
Components	**unclustered**	2	1	2	1	1	1
	clustered	2	1	2	1	1	1
	clustered (no X)	1	2	3	2	3	3
Network diameter	**unclustered**	4	3	5	3	4	4
	clustered	4	3	5	3	4	4
	clustered (no X)	3	3	4	3	4	4
Density	**unclustered**	0.2	0.226	0.228	0.245	0.242	0.261
	clustered	0.23	0.279	0.339	0.279	0.364	0.386
	clustered (no X)	0.619	0.348	0.289	0.364	0.303	0.303
Avg. degree	**unclustered**	5.8	8.8	8.421	9.805	11.149	12
	clustered	6.667	10.9	12.53	11.951	16.42	17.745
	clustered (no X)	3.714	3.833	9.829	4	10.286	10.286
Median degree	**unclustered**	4.5	7.5	7	8	10	10
	clustered	5	8.5	6.5	9	13	13
	clustered (no X)	4	3.5	7	4	9	9
Max. degree	**unclustered**	16	27	26	30	33	35
	clustered	20	38	51	41	66	68
	clustered (no X)	6	12	34	12	36	36
Avg. edge weight	**unclustered**	1.61	4.97	3.31	5.05	5.35	5.47
	clustered	1.4	4.01	2.23	4.14	3.57	3.7
	clustered (no X)	2.31	11.56	2.52	12.33	3.88	4.05
Max. edge weight	**unclustered**	9	106	60	115	166	175
	clustered	6	74	58	80	132	138
	clustered (no X)	6	74	58	80	132	138

Tab. 5 Selected network properties of the eighteen possible networking scenarios described in section 4.

work that consists of only glosses with rank 3, belonging to one of the twenty clusters outlined in section 3.2. The network properties employed as descriptors of individual networks are: the number of nodes, edges[40] and connected components,[41] network diameter,[42] network density,[43] the average, median, and maximum degree,[44] and the average and maximum edge weight (Tab. 5).[45]

5.1 The number of nodes and edges

These two variables tell us how many annotated manuscripts contain parallel glosses of a specific rank or glosses belonging to sets and clusters (nodes) and how many instances of gloss parallelism there are within networks with certain properties (edges).[46] The number of nodes and edges gives us a glimpse of the similarity and robustness of individual networks. They reveal that Par-clustered-noX and Rnk23-clustered-noX both have 35 nodes and 180 edges and thus share network properties apart from their edge weight distribution. Moreover, Rnk3-clustered-noX (35 nodes and 172 edges) closely resembles the previous two, having an identical diameter (4) and number of components (3). As it contains more than 95% of the edges of the two less restrictive networks, it also has a similar average degree (9.829 compared to 10.286) and density (0.289 compared to 0.303). The identical properties of Par-clustered-noX and Rnk23-clustered-

40 The number of edges corresponds to the number of rows in an edge table after rows with specific ranks or cluster labels are removed or contracted (see section 4.3). The number of nodes corresponds to the number of unique manuscript labels that remain in the source and target columns of the same edge table. Both values can also be obtained via Gephi.

41 The number of connected components can be calculated using a method described in Barabási, *Network Science,* sec. 2.9. In this study, it was obtained via Gephi.

42 The network diameter can be calculated using a method described in Ibid., sec. 2.8. In this study, it was obtained via Gephi.

43 The network density was calculated from the number of edges and nodes following the method described in Newman, *Networks,* 128–30. It can also be obtained via Gephi.

44 The average degree was calculated from the number of edges and nodes following the method described in Ibid., 127–28. It can also be obtained via Gephi. The median and maximum degrees were established based on the degree distribution produced by Gephi. The degree distribution could also be obtained manually from an adjacency matrix constructed from the edge table; see Barabási, *Network Science,* sec. 2.3; Newman, *Networks,* 106–8.

45 The average edge weight corresponds to the average value of the column no. of glosses of an edge table. The maximum edge weight is equal to the highest value present in the same column in the same edge table.

46 The range of values we can expect for nodes is up to 47, the total number of annotated manuscripts containing parallel glosses. In unclustered networks, the number of edges can reach 1,081 (if all annotated manuscripts shared at least one parallel gloss with all other annotated manuscripts), while in clustered networks, the number could potentially be higher because two nodes can be connected by parallel edges. In practice, the number of edges is lower because few of the possible connections are present in real-life networks. See Barabási, *Network Science,* sec. 2.5.

noX indicate that glosses with rank 1 play a limited role within our co-occurrence networks. Indeed, Rnk1-clustered-noX has only seven nodes and thirteen edges, containing thus only 20% of the nodes and approximately 7% of edges of the largest network of the same type (Par-clustered-noX), and only 15% of nodes and 3% of the edges of the largest network that can be produced from the data (Par-clustered). This is partially because only 11% of glosses from the dataset have the lowest particularity rank. However, it also transpires from Tab. 5 that there is no manuscript connected to other manuscripts only by glosses with rank 1 (otherwise, the number of nodes of Par and Rnk23 networks would differ). Moreover, only 18 out of 282 edges of Par-unclustered and 24 out of 417 edges of Par-clustered (approximately 6% of the edges of both most inclusive networks) are constituted solely by glosses with rank 1. Rnk1 networks thus do not carry much weight on their own. Therefore, it seems safe to exclude glosses with rank 1 from further consideration in this analysis.

Rnk2-clustered-noX (12 nodes and 23 edges) and Rnk12-clustered-noX (12 nodes and 24 edges) also comprise a very small proportion of nodes and edges of the most inclusive network that can be constructed from the data available. They are thus not particularly robust when it comes to analyzing the connectivity within the corpus. However, Rnk2-clustered-noX has a very high average (11.56) and maximum (74) edge weights, an indication that glosses with rank 2 represent a significant portion of the volume of many clusters. Therefore, Rnk2-clustered-noX cannot be discarded altogether. As is shown below, it is a vital complement to Rnk3-clustered-noX, the most restrictive network that can be constructed from the data available and the network that is most relevant to understand the connections between annotated manuscripts of the first book of the *Etymologiae* that are certainly due to the transmission of glosses.

5.2 The number of connected components and network diameter

The number of connected components reveals whether all the nodes in a network are mutually interconnected (one component) or whether any parts of the network are isolated from each other (a higher number of components).[47] In this study, the latter means that certain annotated manuscripts mutually share parallel glosses, but otherwise differ from all other manuscripts in the corpus. As Tab. 5 shows, we find a single connected component in many of the network scenarios we can construct. However, once unassigned sets are removed, disconnected components appear. The extensive connectivity thus appears to be due to noise. If we furthermore limit our attention to glosses with rank 3, the network disintegrates into three components. This fragmentation points to the existence of several disconnected or weakly connected glossing communities.

47 Ibid., sec. 2.9; Newman, *Networks,* 133–37.

The network diameter corresponds to the longest direct path between nodes in a network component.[48] A network with a diameter of 1 is complete (as it is possible to reach every node from every other node), while in a network with the diameter of 2, all nodes in a component are connected through a single central node (called a hub) so that it is possible to reach any node from any other node through this central point. The network diameter thus provides us with a measure of connectivity within a network related to the presence of clusters and weakly connected segments. Translated into scholarly language, the larger the diameter, the more annotated manuscripts we can expect to display weak connections to other manuscripts. As in the case of connected components, identifying these weakly connected segments in a network is valuable for tracing parts of the corpus that are mutually distinct or relatively different and thus identifying glossing communities that are disconnected or poorly connected.

In our case, the smallest network diameter we observe in our co-occurrence networks is 3. This diameter, just one step beyond the network with a single central hub, appears in networks built solely or principally from glosses with rank 2 and one of the networks constituted by glosses with rank 1. It is an indication that glosses with these two lower ranks, particularly rank 2, play a role in shortening paths between nodes. By contrast, two networks constructed from glosses with rank 3, Rnk3-unclustered and Rnk3-clustered, have a large network diameter of 5, while Rnk3-clustered-noX has a smaller diameter of 4. The decrease in the diameter in Rnk3-clustered-noX compared to the other two Rnk3 networks signals that in the latter networks, particular far-flung nodes are only connected to the rest by glosses not assigned to any cluster. Once these weak, possibly phantom, connections are removed, the network becomes disconnected into several components.[49] For our purposes, Rnk3 networks show the most topographic detail and therefore merit further examination to identify weakly connected segments and potential bridges (i.e., manuscripts that connect otherwise disconnected parts of the network). Rnk2 networks, on the other hand, could be useful to inspect to determine to what extent glosses with rank 2 generate meaningful connections within our co-occurrence networks that do not feature in networks constructed from glosses with rank 3, and to what extent the increased connectivity is due to noise.

48 Barabási, *Network Science,* sec. 2.8; Newman, *Networks,* 133.

49 Thus, the number of components in Rnk3-clustered-noX increases as its network diameter decreases.

5.3 Network density

Network density informs us about how complete a network is, i.e., what proportion of the possible connections in the network, as given by the number of nodes, have been realized.[50] In this study, the network density provides us with a complementary perspective into the extent to which annotated manuscripts are connected to other annotated manuscripts containing glosses to the first book of the *Etymologiae*. A low density may suggest that manuscript annotators rarely acquired glosses from other intellectual centers (if we assume transmission, such as in the case of glosses with rank 3), or seldomly came up with glosses similar to those coined elsewhere (if we rather assume spontaneous composition), their glossing activity being one-of-a-kind. A high density, by contrast, is a significant indicator of gloss parallelism and therefore, potentially, of the extensive circulation of glosses. Typically, the density value can range from 0 (i.e., 0%, if no manuscript shares a gloss with another manuscript) to 1 (i.e., 100%, if every manuscript shares at least one parallel gloss with all other manuscripts).[51]

If we disregard networks with low robustness, the network density values in the corpus studied here range from 0.226 (Rnk2-unclustered) to 0.261 (Par-unclustered) in unclustered networks, from 0.279 (Rnk2-clustered and Rnk12-clustered) to 0.386 (Par-clustered) in clustered networks, and from 0.289 (Rnk3-clustered-noX) to 0.303 (Rnk23-clustered-noX and Par-clustered-noX) in clustered networks removing the unassigned sets. As can be seen, the gloss parallelism observable in the more robust co-occurrence networks is relatively stable, corresponding to between approximately 23% and 39% of the gloss parallelism we would see if all manuscripts shared glosses with all other manuscripts. These ratios make our co-occurrence networks relatively dense, in particular when compared to other real-world networks, such as those examined by Lászlo Barabási in his *Network Science*.[52] Some of this density is certainly due to the method described here, for example the decision to remove isolated nodes from the co-occurrence networks studied here (thus the number of nodes varies per network), and at least in part due to the presence of many locally complete sub-graphs (see section 4.1). It can also be partially attributed to noise due to spontaneous composition and random gloss parallelism. We can see this in Par-unclustered and Par-clustered, the two most inclusive co-occurrence networks, if we subject them to a small test, removing edges constituted by less than a certain number of glosses. The density of Par-unclustered (0.261) drops to 0.14 if we exclude edges

50 Barabási, *Network Science,* sec. 2.5; Newman, *Networks,* 128–30.

51 Due to the presence of parallel edges, the value could theoretically exceed 1 in a clustered network.

52 For example, the science collaboration and the citation networks provided as examples of real-world networks by Barabási have densities of 0.00035 and 0.000046, respectively. See Barabási, *Network Science,* sec. 2.2.

constituted by a single gloss, to 0.093 if we also exclude edges constituted by two glosses, and to 0.06 if we exclude edges constituted by less than six glosses. The density of Par-clustered (0.386) similarly drops to 0.168 if we remove edges constituted by a single gloss, to 0.127 if we also exclude edges constituted by two glosses, and to 0.069 if we exclude edges constituted by less than six glosses.

This rapid decrease in density could mean that the relatively high densities of our networks are due to noise rather than meaningful connections between annotated manuscripts. If this were the case, the gloss parallelism in an organic corpus of glosses could be assumed to occur mainly due to processes other than transmission. However, the most restrictive network, Rnk3-clustered-noX, which filters out potential noise in a stricter fashion than the test described above, is significantly denser than even the most lightly filtered network containing unassigned sets, and has a higher density (0.289) than even its unclustered counterpart, Rnk3-unclustered (0.228). The high densities of the clustered-noX networks can be interpreted as an indicator that transmission contributes to the high gloss parallelism observed in our co-occurrence networks to a rather significant degree. Moreover, it also tells us that by gloss clustering and removing unassigned sets, it is possible to remove some of the noise from the corpus without significantly imperiling, and even increasing, its informative value. Filtering out edges based solely on their weight, on the other hand, degrades the networks wholesale, eliminating not only noise but also valuable information. Gloss clustering and the removal of unassigned sets are thus essential data pre-processing strategies, if we want to obtain high-quality insights into organic gloss corpora, while ignoring the presence of gloss clusters and sets of low importance is likely to produce unreliable results, particularly if the study of gloss transmission is the main concern.

5.4 The average, median and maximum degree

The average degree (the average number of connections a node has with other nodes in a network) and the median degree (the number of connections that a node in the exact mid-point of the degree distribution has with other nodes in a network) inform us how well-connected individual nodes are to each other and what role different types of connections (e.g., clusters and unassigned sets) play in forging this connection.[53] Depending on the network type, the average degree indicates the average number of manuscripts that any given manuscript shares glosses with (unclustered), or the average number of sets and clusters (clustered) or clusters alone (clustered-noX) through which any manuscript is connected to

53 Ibid., sec. 2.3. In an unclustered network, the average and median degrees can range from 1 (in a network constituted by isolated manuscript pairs) to one less than the maximum number of nodes (in a complete network). In clustered networks, the average and median values can be higher because a node can be connected to another node by several parallel edges.

other manuscripts. The median degree tells us that at least half of the manuscripts in a network have the same or a higher number of connections with other manuscripts in the same network than the median value. In unclustered networks, these connections refer to manuscripts, while in clustered networks, they refer to sets and clusters (clustered) or clusters alone (clustered-noX).[54] The maximum degree reveals the largest number of manuscripts with which a manuscript from the corpus shares parallel glosses (unclustered), and the largest number of connections of any manuscripts facilitated by sets and clusters (clustered) or clusters alone (clustered-noX). Ideally, the following analysis of average, median and maximum degrees would complement the degree distribution plotted for the eighteen network scenarios described by Tab. 5. However, since this would require significant space, we shall rely on these three network properties to understand degree, the centrality measure chosen to characterize our co-occurrence networks.[55]

The values of average degree observed in unclustered networks range from 5.8 (Rnk1) to 12 (Par), meaning that, depending on the layer of the corpus examined, an annotated manuscript shares parallel glosses with, on average, between approximately six and twelve other manuscripts (20% to 26% of manuscripts in the respective networks).[56] The average degree ranges from 6.667 (Rnk1) to 17.745 (Par) for clustered networks and from 3.714 (Rnk1) to 10.286 (Rnk23 and Par) for clustered networks excluding unassigned sets, telling us that sets and clusters facilitate on average approximately seven to eighteen connections and clusters alone, approximately four to ten connections within the corpus.[57] The median degree values of unclustered networks occupy a range from 4.5 to ten manuscripts. This piece of information tells us that, depending on the layer of corpus examined, half of the manuscripts share glosses with at least four other manuscripts in the smallest network and with at least ten other manuscripts in the largest network.[58] The median degree ranges from five to thirteen connections in clustered networks and 3.5 to nine connections in clustered networks excluding unassigned sets.[59]

---

54 Thus, in Par-unclustered (median degree of 12), half of the manuscripts share glosses with twelve or more manuscripts; and in Rnk3-clustered-noX (median degree of 7), half of the manuscripts have seven or more connections to other manuscripts via clusters.

55 Alternative centrality measures used in network research are described in Newman, *Networks,* 159–77. On the relative utility of the four most common centrality measures, including degree, in historical network research, see Valeriola, "Can Historians Trust Centrality?"

56 The general corpus average is nine to ten manuscripts.

57 The general corpus average is twelve to thirteen connections if both clusters and sets are considered and seven connections if only clusters are considered.

58 The average median for the entire corpus is seven to eight manuscripts.

59 The average median for the entire corpus is nine connections between manuscripts facilitated by sets and clusters and six connections facilitated by clusters alone.

Even in the absence of an available comparison with co-occurrence networks constructed for other types of material, such as highly systematic corpora of glosses and regular texts, the observed average and median degree appear very high. In particular, it can be noted that the large majority of parallel glosses (~ 90%) are shared by only two or three manuscripts, i.e., they generate one or three edges, while only a small number of glosses (~ 4.5%) are shared by five or more manuscripts, i.e., they generate ten or more edges (see Tab. 1).[60] We could therefore expect many nodes in our co-occurrence networks to have relatively low degrees and very few nodes that have high degrees.[61] Yet, there are relatively few manuscripts with the lowest degrees of one to three in our co-occurrence networks and relatively many nodes with degrees of ten or higher. To provide examples: only seven nodes have a degree of three or lower (14.9%) in Par-unclustered (47 nodes), but 28 nodes have a degree of ten or higher (60%); meanwhile, only six nodes have a degree of three or lower (12.8%) in Par-clustered (47 nodes), but 29 nodes have a degree of ten or higher (61.7%). In the most restrictive network, Rnk3-clustered-noX (35 nodes), these ratios are slightly more balanced; nonetheless, only seven nodes have a degree of three or lower (20%), while sixteen nodes have a degree of ten or higher (45.7%).[62] The degree distribution of our co-occurrence networks does not follow the distribution of parallel glosses (or edge weights treated below), nor does it resemble the degree distribution common to many real-world networks, which follows the power law.[63]

The high average and median degree values warrant further investigation that cannot be fully carried out in this article. Intuitively, it could be assumed that this is an effect of noise, i.e., that the degree distribution is distorted by gloss parallelism due to spontaneous composition and randomness. It can be noted that the average and median degree values tend to be highest in clustered networks, whose densities suggest that they are not entirely reliable, and the lowest in clustered networks excluding unassigned sets, that is, in networks constructed with

60 Even if we considered the total number of edges a parallel gloss can generate based on the number of manuscripts it connects (rather than the number of parallel glosses), based on Tab. 1, 57% of the edges in the dataset are due to parallel glosses shared by two or three manuscripts (i.e., generating one or three edges), 29% of the edges are due to parallel glosses shared by four or five manuscripts (i.e., generating six or ten edges), and only 14% of the edges are due to parallel glosses shared by six or more manuscripts (i.e., generating 15 to 45 edges).

61 Compare with Barabási, *Network Science,* sec. 2.3, 3.5 and 4.2.

62 We encounter the highest ratios of nodes with degrees of one to three in Rnk12-clustered-noX (5 nodes, 42%) and Rnk2-clustered-noX (6 nodes, 50%). However, these networks are very small. These ratios are, therefore, less meaningful than in networks with more nodes and edges.

63 The power-law distribution means, in the most general terms, that nodes with the lowest degree should be most numerous and nodes with the highest degree should be least common in a network. Degree distribution following power law is a feature of the so-called scale-free networks treated at length in Barabási, *Network Science,* sec. 4.2.

a precaution taken against the distortive effect of noise. Unclustered networks tend to have a value lower than clustered ones and are similar to clustered-noX networks. Thus, the co-occurrence network with the most edges, Par-clustered, displays 47% more connections than Par-unclustered and 73% more connections than Par-clustered-noX. This inflation in the degree values in clustered networks is due to unassigned sets, many of which are presumably noise.

We may opt to derive our insights entirely from clustered-noX networks, considering their average and median degree values are minimally distorted by noise, or less distorted than in clustered or unclustered networks. Even so, the values of these two properties in the three most robust clustered-noX networks (Rnk3, Rnk23, and Par) are still surprisingly high. They indicate that an annotated manuscript of the first book of the *Etymologiae* contains, on average, parallel glosses from ten clusters, and at least half of such manuscripts contain parallel glosses from seven to nine clusters, even though the loss of manuscript evidence is potentially substantial. What is more, Rnk3 networks deviate from the trend described in the previous paragraph, as the average degree of Rnk3-clustered-noX (9.829) is higher than that of Rnk3-unclustered (8.79). As in the case of network density, this deviation suggests that meaningful transmission-related connections between manuscripts missing from unclustered networks are revealed in Rnk3-clustered-noX. Given the rich topography of Rnk3-clustered-noX, as suggested by the number of connected components and network diameter, it is tempting to relate the high average and median degree values to the multilayered character of the organic corpus studied in this article (a trait demonstrated in the following section).[64] The high average and median degree values of the clustered-noX networks, especially Rnk3-clustered-noX, may thus reveal the extent to which collections of annotation in manuscripts of the first book of the *Etymologiae* are amalgamating batches of glosses originating in distinct contexts (more on this in section 7).[65]

Unlike spontaneous composition and random gloss parallelism, the accumulation of glosses of heterogeneous origin in a manuscript is bound to generate hubs, i.e., manuscripts that stand out because they share glosses with an un-

64 We can engage in a thought experiment, imagining how the process of transmission of glosses together with the substrate text (e.g., copying from an annotated exemplar to its apograph) differs from the process of copying batches of glosses or individual glosses into a manuscript. While the former manner of transmission can increase the degree of a node in a co-occurrence network of parallel glosses only by one, the latter transmission process can increase the degree of a node by n, where n is the number of manuscripts that already contain the same batch. The collection of glosses thus has the potential to increase the degree of a node at a rate significantly higher than copying from an exemplar to an apograph and create hubs.

65 The corpus thus confirms the scholarly theories about the cumulative nature of early medieval glossing; O'Sullivan, "Text, Gloss, and Tradition in the Early Medieval West."

usually large number of other manuscripts, or because they are connected to other manuscripts by an unusually large number of clusters.[66] Indeed, the small difference between the average and median degrees, ranging from –0.29 to 2.8 in unclustered and clustered-noX networks, informs us that we should expect some, albeit not too many, hubs in our co-occurrence networks.[67] The maximum degree values in Tab. 5 reveal that in Par-unclustered, we encounter a manuscript that shares glosses with as many as 76% of the other manuscripts. The same manuscript shares glosses with 66% of other manuscripts in Rnk3-clustered-noX. The manuscript in question is Harley3941, which looms large among the annotated manuscripts of the first book of the *Etymologiae* due to its remarkable extent of gloss parallelism and gloss sharing – it is the most significant hub in our networks.[68] Several other nodes with high degrees compared to both average and median degree values also qualify as hubs. Rnk3-clustered-noX can be considered the most informative in this regard, given its difference between the average and median degree values (2.8). In this restrictive network, we encounter four manuscripts other than Harley3941 that are connected to a higher number of manuscripts than average in this network (29%): Paris7670 and Schaffhausen42 (50%), Paris7585 (44%), and IRHT342 (38%). These are examined against the background of extrinsic evidence in the following section.

5.5 The average and maximum edge weights

The average and maximum edge weight (the average and highest number of glosses shared between a pair of manuscripts), gives us an insight into the volume of gloss parallelism within the corpus. In our co-occurrence networks, the average edge weights range from 1.4 glosses (Rnk1-clustered) to 12.33 glosses (Rnk12-clustered-noX), while the maxima range from 6 glosses (Rnk1-clustered-noX) to 175 glosses (Par-unclustered). The edge weights are distributed in a more standard pattern than degrees, with the majority of edges consisting of very few glosses, and few edges being very heavy.[69] In the two most inclusive networks, Par-unclustered and Par-clustered, for example, edges constituted by one or two

66 Hubs are described in Barabási, *Network Science,* sec. 2.11; Newman, *Networks,* 178–80.

67 The difference between the average and the median degree values tells us to what extent the average is skewed by nodes with unusually high degrees, which are not entirely representative of the network. The more positive its value, the more outliers with high degrees (i.e., hubs) there are in a degree distribution of a given network. The more negative its value, the more outliers with low degrees there are in a degree distribution of a given network.

68 Harley3941 has the largest degree in most of the eighteen networks described in Tab. 5. The exceptions are Rnk1-clustered-noX (Reims426), Rnk2-unclustered (VLO41), Rnk2-clustered (VLO41, Orleans296 and Reims426), Rnk12-unclustered (VLO41), and Rnk12-clustered (VLO41).

69 Plotting the distribution of the edge weights on a logarithmic scale suggests that it follows the power law.

glosses comprise 65% (183 edges) and 67% (280 edges) of all network edges, respectively, while edges constituted by more than ten glosses comprise 12% (34 edges) and 8.4% (35 edges) of the network edges, respectively. This distribution matches what was observed in section 3.1, namely that most cases of gloss parallelism between manuscripts in the corpus (77.6%) are due to one or two glosses. Importantly, these lightweight edges correspond not only to unassigned sets, which could be interpreted as noise, but also to the seven micro-clusters C1–C7 and several clusters constituted by glosses with rank 3. This is why the average edge weight remains low in Rnk3-clustered-noX (2.52) but increases significantly in Rnk2-clustered-noX (11.56), which does not feature glosses that belong to unassigned clusters or those with rank 3. This network provides us with a particularly undiluted view of the volume of gloss parallelism in our co-occurrence networks, both because of the significant proportion of parallel glosses with rank 2 in the corpus (58%) and because of their role in adding weight to clusters whose contours are provided by glosses with rank 3. Indeed, if the edge tables of Rnk2-clustered-noX and Rnk3-clustered-noX are compared, it can be observed that the former contains only two edges that do not appear in the latter, i.e., most of the glosses with rank 2 appear in the same clusters as glosses with rank 3 and can be therefore considered to reflect transmission.

Just as the analysis of average and maximum degrees reveals some nodes to be hubs, the average and maximum edge weights reveal certain edges as outliers, constituted by an exceptionally high number of glosses. One edge that stands out in this regard, across all networks, is Harley3941-Orleans296, which corresponds to cluster F in clustered networks. In Par-unclustered, this edge consists of 175 glosses, while the next heaviest edge (Orleans296-VLO41) is constituted by 57 glosses (i.e., less than a third of the former's weight); in the most restrictive network, Rnk3-clustered-noX, Harley3941-Orleans296 amounts to 58 glosses, followed by Gotha1147-Paris7585 with 32 glosses (55% of the former); and in Par-clustered, the network with the most edges, this edge consists of 138 glosses, followed by Orleans296-VLO41 with 38 glosses (i.e., approximately a quarter of the former's weight). For the same reason that Harley3941 and other manuscripts with exceptionally high degrees can be considered hubs, we can identify this edge as a highway.

5.6 General trends in the co-occurrence networks of parallel glosses to the first book of the Etymologiae

After examining all co-occurrence networks constructed from the data introduced in section 3, we can identify several trends that characterize the corpus of glosses to the first book of the *Etymologiae*. First, we have seen that these co-occurrence networks are relatively dense and tend to have very high average and median degrees. This is somewhat surprising, given the historical context of the generation and circulation of organic glosses. Medieval scribes and masters may have been eager to acquire glosses, but they had, in practice, limited access to the totality of

glosses circulating in the Latin-writing world. Crucially, the degree of gloss parallelism among the annotated manuscripts is unusually high even in networks constructed by applying the most restrictive criteria that are intended to curb any potential inflation of connections due to spontaneous composition and other noise. It cannot thus be attributed to medieval annotators frequently generating glosses similar to those independently composed by others. It rather seems that a substantial gloss parallelism is an intrinsic quality of the corpus studied here. If we ask ourselves what real-world property the high density, average and median degree may correspond to, they may be a characteristic of the multilayered character of the corpus, telling us that manuscripts of the first book of the *Etymologiae* attracted glosses of heterogeneous origin. As such, these network properties may serve as an indicator of the high organicity of this corpus.

Second, while some of the gloss parallelism observable in the corpus studied here is due to spontaneous composition, generating glosses that appear identical but are not different manifestations of the same transmitted items or even due to random similarity, two properties of our co-occurrence networks suggest that it is mostly due to transmission. First, the fact that the most restrictive network, Rnk3-clustered-noX, does not look and behave like a network we could obtain by applying filters to unclustered networks, having significantly more edges. In addition, glosses with rank 2 do not form new edges in clustered networks excluding unassigned sets (i.e., eliminating noise), but rather add volume to edges that we can reconstruct as reflecting transmitted gloss clusters based on glosses with rank 3. The properties of the co-occurrence networks constructed in this section thus confirm that the circulation of the glosses to the first book of the *Etymologiae* in the early Middle Ages was extensive, as proposed based on the extrinsic evidence in section 3.3.

Third, Rnk3-clustered-noX, the network most geared towards investigating transmission patterns, is relatively topographically rich, featuring isolated components, weakly connected segments, and hubs. Based on the network analysis performed in this section, we cannot yet tell to what extent the observed disconnectedness is due to the extent of manuscript loss, and to what extent it reflects the relative mutual isolation of certain glossing communities. For this purpose, we need to conduct a qualitative examination in section 6. Another network element we could expect to see in the visualization in the following section are bridges, i.e., nodes that appear at the interconnection of otherwise unconnected or poorly connected segments.[70] These network elements are a reflection of human activity, which scholars may be particularly interested in identifying via network analysis in order to study them in greater detail with traditional methods.

70 Bridges are discussed in Barabási, *Network Science,* 2.9.

Finally, our analysis revealed that not all co-occurrence networks that can be constructed from the data described in section 3 have the same degree of robustness and quality for analytical purposes. Glosses with rank 1, for example, can be excluded from consideration without reducing the accuracy of the analysis of the corpus examined in this article, both because they are not particularly numerous within this corpus and because it can be shown that they do not generate any significant connections. We have also seen that clustered networks, that is, network scenarios in which sets and clusters are recognized but unassigned sets are not removed, suffer from an 'edge bloat' that can be interpreted mostly as noise and are therefore not particularly useful (apart, perhaps, from an analysis of this noise and its sources). For this reason, it also does not seem sufficient to treat all parallel glosses as shared. Overall, low-key gloss parallelism due to one or two glosses with the lower ranks 1 and 2 has a distortive effect on the quality of the co-occurrence networks we can construct from the dataset provided. It generates many weak connections between nodes that should not be trusted. However, since many connections facilitated by glosses with the highest particularity rank 3 are also due to one or two glosses (e.g., micro-clusters C1–C7), constructing unclustered networks and filtering out low-weight edges does not appear to be a good strategy to obtain high-quality data.

In seeking to answer the questions articulated in the introduction of this study, in particular the question of the extent and shape of the transmission of organic glosses, it seems most profitable to focus on co-occurrence networks that account for gloss clusters but exclude unassigned sets and glosses with the lowest rank, 1. In the following section, therefore, Rnk23-clustered-noX is visualized.[71] With this configuration, we benefit from a single network scenario that fleshes out both a) the general contours of connectivity within the gloss corpus studied here, namely how far and where the circulation of glosses extended to, and where it may have been weak or non-existent; and b) the intensity of this connectivity, i.e., where the exchange may have been most significant.

6. Visualization

In this section, Rnk23-clustered-noX is visualized and interpreted with the support of extrinsic evidence (i.e., paleographic, philological, linguistic, and historical information) in light of the network analysis performed in the previous section. This visualization was produced with Gephi by following these steps:

71 It could have been even more profitable to visualize Rnk3-clustered-noX and Rnk2-clustered-noX separately to obtain a subtler picture. However, as the space offered by this article is limited, the choice went to Rnk23-clustered-noX on the basis that only two edges in Rnk2-clustered-noX do not feature in Rnk3-clustered-noX (on these, see below) and thus there is a good overlap between the two layers of the corpus.

1) The node table and the appropriate edge table were loaded into Gephi. Nodes from the node table that do not feature in the edge table (i.e., have a degree of 0) were removed so as not to be displayed as isolated nodes.
2) Node color was adjusted to represent the manuscript type (orange: grammatical handbooks containing only the first book of the *Etymologiae;* purple: library books containing the entire *Etymologiae;* green: manuscripts containing excerpts from the first book of the *Etymologiae*). Node size was set to correspond to the number of parallel glosses.
3) The color of the edges was adjusted to correspond to clusters (A: dark blue; B and O: light green; C1–C7: orange; D and P: turquoise; E: dark red; F: brown; G: dark green; I: yellow; M: light blue; N: dark purple; Q: pink; and S: black).[72]
4) The Yifan Hu layout was applied to the network because of its suitability for small undirected weighted networks and good cluster detection. In addition, Geo Layout using the GPS coordinates in the node table was used to produce visualizations sensitive to the regional localization of annotated manuscripts.[73] The position of nodes was further adjusted with the Label Adjust algorithm to give the network graph a cleaner look.
5) The network graph was manually adjusted to accentuate particular connections and make the visualizations more compact, e.g., nodes were moved apart to prevent edge overlap and make segmentation more visible, and isolated components were moved closer to the main component to decrease the size of the visualization.

Fig. 5 contains two visualizations of Rnk23-clustered-noX: in Fig. 5a, the layout created with the Yifan Hu algorithm reveals the clustering within the network based on gloss parallelism between annotated manuscripts; in Fig. 5b, the geographical relationships between manuscripts featuring shared glosses is mapped with the Geo Layout algorithm.

72 In several cases, two unrelated clusters that concern unrelated manuscripts were assigned the same color. This was done to reduce the color palette used in visualizations and make them more readable.

73 As we typically cannot pinpoint the location of the glossing of a manuscript more precisely than to a region, the position of the nodes in the geo-located visualization should be considered approximate at best. In most cases, it does not inform us about the relation between specific places, such as early medieval monasteries. However, it is precise enough to allow us to consider the circulation of shared glosses in six regions of the early medieval Latin-writing world mentioned in section 3.3: France, the German area, Brittany, England, northern Italy, and Spain.

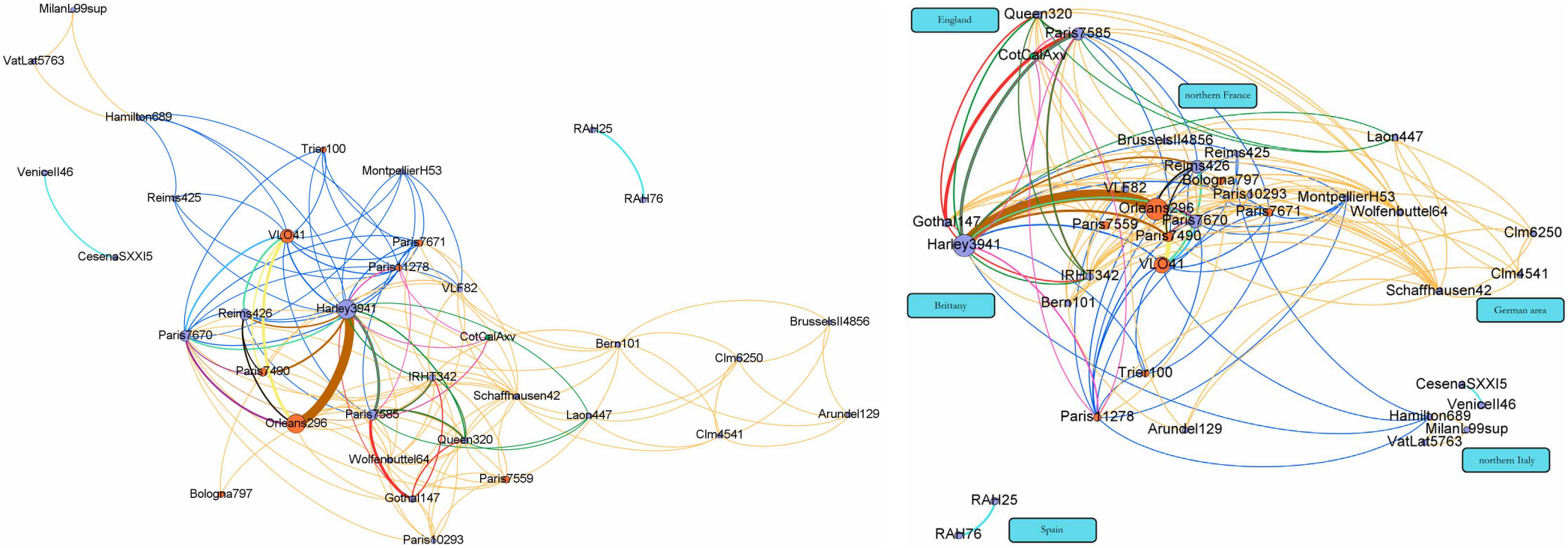

Fig. 5 Rnk23-clustered-noX projected with Yifan Hu (left, 5a) and Geo Layout (right, 5b) algorithms. Clusters displayed, in the order of the proportion of displayed edges include: C1–C7 (orange, 51.7%), A (blue, 23.3%), E (red, 6.1%), G (dark green, 6.1%), F (brown, 3.3%), Q (pink, 3.3%), I (yellow, 1.1%), N (dark purple, 1.1%), S (black, 1.1%), B (light green, 0.6%), D (turquoise, 0.6%), M (light blue, 0.6%), N (dark purple, 0.6%), O (light green, 0.6%), and P (turquoise, 0.6%). Node color: library book containing the entire encyclopedia (purple, 26 manuscripts), grammatical handbooks (orange, 6 manuscripts), and excerpts (green, 1 manuscript). Node size corresponds to the number of parallel glosses present in the manuscript.

6.1 Components and the regional glossing patterns

Rnk23-clustered-noX is dominated by a large component in which we find the majority of manuscripts (31 out of 35 nodes), flanked by two small, isolated components constituted by manuscript pairs CesenaSXXI5-VeniceII46 and RAH25-RAH76. Fig. 5b reveals that manuscripts in the large component were annotated in different parts of France, Brittany, England, and the German area. Its densest core corresponds to the area of northern France, where the most important Carolingian intellectual centers were situated.[74] The two isolated components reflect glossing in northern Italy (CesenaSXXI5-VeniceII46) and Spain (RAH25-RAH76). This figure also shows a trio of interconnected manuscripts (Hamilton689-MilanL99sup-VatLat5763) attached weakly to the large component, also glossed in northern Italy (they appear in the upper left corner of Fig. 5a, at the periphery of the main component). Were it not for a single parallel gloss from cluster A in Hamilton689 (blue), these three manuscripts would be separated from the large component.

An examination of the extrinsic evidence shows that the manuscripts in the two isolated components are closely related philologically: VeniceII46 is a direct copy of CesenaSXXI5, and RAH25 and RAH76 are either parent and offspring, or two siblings.[75] Edges connecting these manuscript pairs thus correspond to the copying of glosses from an exemplar to its apographs, a transmission pattern not observed elsewhere in the corpus. Given this connectivity pattern, it does not seem likely that the isolation of the two small components in Rnk23-clustered-noX is solely due to the loss of connection to the large component as a result of the disappearance of manuscripts. Rather, it seems indicative of distinct attitudes to glossing the *Etymologiae* and limited contact between the glossing communities in Spain, northern Italy, and other regions of the medieval Latin-writing world.

While this may not be particularly visible in Fig. 5b, the poor connectivity to the Carolingian glossing communities is characteristic not only of northern Italy and Spain but also of the German area represented in Rnk23-clustered-noX by Clm4541, Clm6250, Laon447, Schaffhausen42, and Wolfenbuttel64. These manuscripts appear close to each other in the base of the segment extending on the right from Fig. 5a, connected by several micro-clusters both mutually and with other manuscripts, mainly from France. Were it not for the micro-clusters, the German area would vanish from this network graph almost entirely.[76] Comparing the 35 nodes displayed in this visualization with the 54 annotated manuscripts

74 Contreni, "The Carolingian Renaissance," 721.
75 Bellettini, "Il codice del sec. IX di Cesena, Malatestiano S. XXI.5," 75–91; Steinová, "Annotation of the Etymologiae of Isidore of Seville in Its Early Medieval Context," 38.
76 The exception is Laon447, which contains glosses from cluster G (see below).

of the first book of the *Etymologiae* in Appendix I, we can note that many manuscripts from northern Italy and the German area are absent from Rnk23-clustered-noX. Even in light of the possibly substantial loss of annotated manuscripts, we can conclude that while the first book of the *Etymologiae* was annotated in all major regions of the early medieval Latin-writing world, the transmission of glosses to the first book of the *Etymologiae* was principally restricted to three regions – France, Brittany, and England – and was most intense in the Carolingian heartland in northern France.

6.2 Three layers of the main component

The general inspection of Fig. 5 revealed certain qualitative differences between specific regions in terms of the nature of the glosses circulating within them (e.g., transmission from an exemplar to an apograph in northern Italy and Spain, as opposed to the prevalence of the micro-clusters in the German area). To take a further step in this visually-supported analysis, we can dissect Rnk23-clustered-noX into layers corresponding to specific clusters and cluster groupings. By plotting them separately, we can better appreciate that these layers show limited overlap, have different network properties, feature glosses with distinct philological profiles, and correspond to different manuscript contexts. They thus appear to reflect distinct historical circumstances of transmission and regional trends.

We can recognize three layers in Rnk23-clustered-noX. The most prominent of these is the layer of glosses assigned to the micro-clusters C (Fig. 6a, 51.7% of edges of Rnk23-clustered-noX, three parallel edges with cluster A, seven parallel edges with other clusters). Manuscripts containing these glosses were annotated in all the regions mentioned above, apart from Spain. Most contain no other glosses to the *Etymologiae,* although some (Schaffhausen42) attracted glosses from multiple micro-clusters. Looking at the manuscript context of their transmission, we can understand why these glosses were transmitted in isolation rather than as parts of clusters and are attested in regions in which glosses to the *Etymologiae* do not otherwise seem to have circulated widely. Most appear fossilized or semi-fossilized in the main text, meaning that they passed or survived in the process of passing from the white to the black space of a manuscript.[77] Indeed, some of the connections visible in Fig. 6a are due to the fact that medieval scribes did not discern these fossilized and semi-fossilized glosses as glosses, but copied them as a part of the main text, as corrections, or as variant readings. The fossilization indicates their older age relative to the annotated manuscripts that preserve them, which mostly date from the ninth century. Since some of their witnesses are early (e.g., BrusselsII4856 copied and annotated at the end of the

77 On gloss fossilization, see the introduction of Steinová and Boot, "The Glosses to the First Book of the Etymologiae of Isidore of Seville"; Stagni, "Nell'officina di Paolo Diacono?"

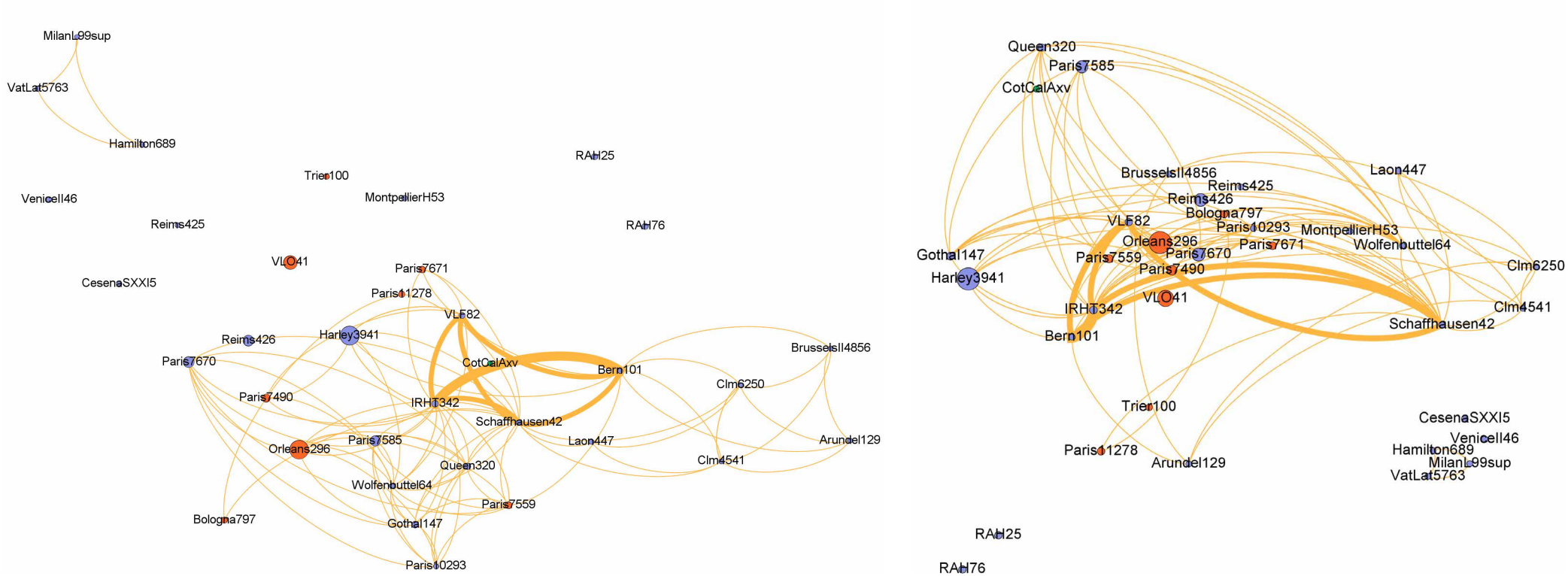

Fig. 6a The layer of Rnk-23-clustered-noX network corresponding to micro-clusters C1-C7 (orange). Layouts: Yifan Hu (left) and Geo Layout (right).

eISSN: 2535-8863
DOI: 10.25517/jhnr.v9i1.198

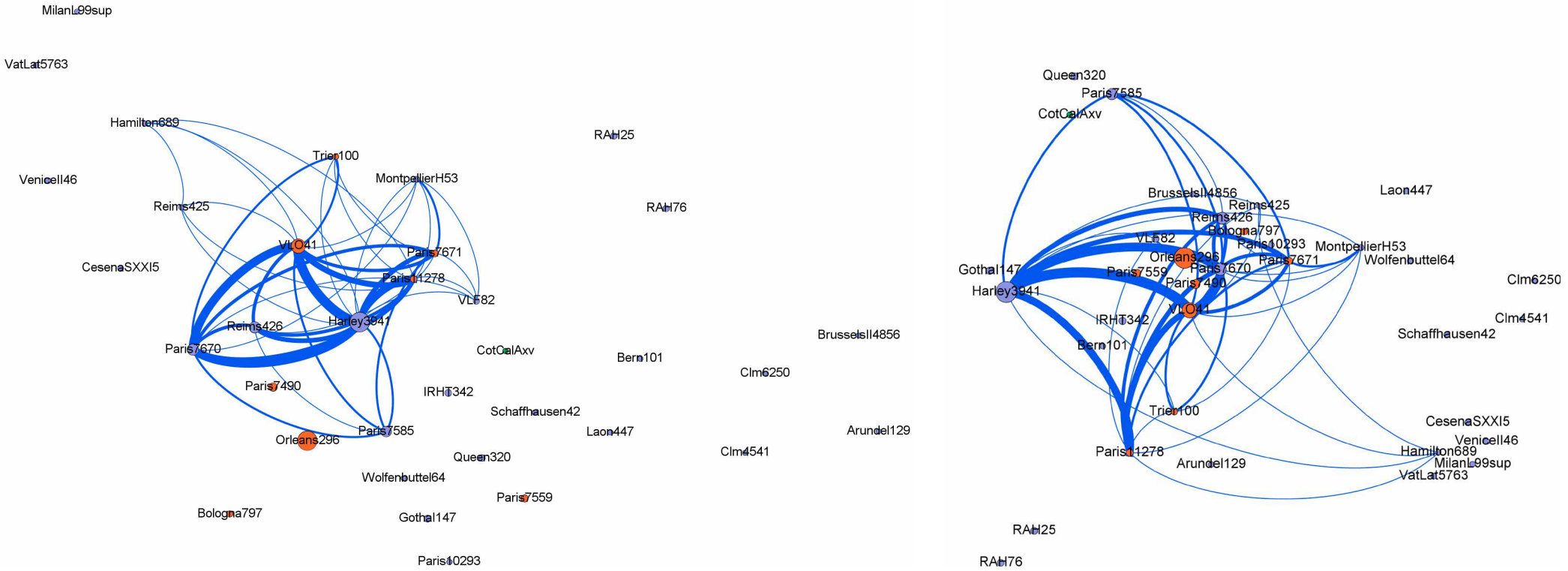

Fig. 6b The layer of Rnk-23-clustered-noX network corresponding to cluster A (blue). Layouts: Yifan Hu (left) and Geo Layout (right).

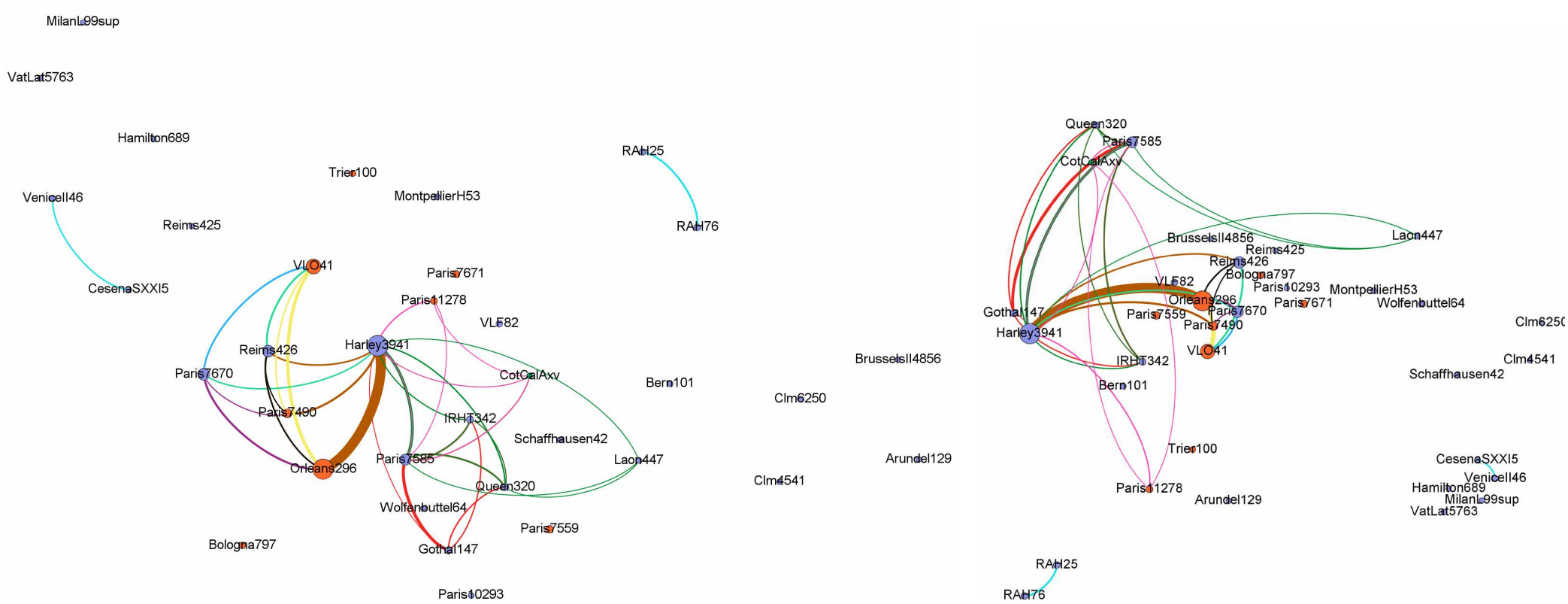

Fig. 6c The layer of Rnk-23-clustered-noX network corresponding to clusters other than A and C1-C7 (other colors). Layouts: Yifan Hu (left) and Geo Layout (right).

eighth century in Corbie, and MilanL99sup copied and annotated in the second half of the eighth century in Bobbio), and given their wide diffusion range, which also presupposes a long period of transmission prior to the earliest attestation, the micro-clusters look like a remnant of pre-Carolingian glossing. We know very little about glossing before the year 800, but we can presuppose it to have taken place given the popularity that the *Etymologiae* had already enjoyed in the Latin-writing world from the seventh century.[78] These glosses must have been more numerous and had a similar transmission dynamics as other glosses analyzed in this study (e.g., transmission in batches rather than in insolation and independently from the substrate text). However, only a small number of them survived due to fossilization. Once embedded into the black space, moreover, the patterns of co-occurrence based on the micro-clusters mirror the transmission patterns of the substrate text (the *Etymologiae*), rather than conforming to what can be expected of glosses. The extremely fragmentary state in which these witnesses of pre-Carolingian engagement with the *Etymologiae* reach us means we cannot reconstruct their context of origin or the direction of their diffusion.[79]

The second layer of glosses, corresponding to cluster A (Fig. 6b, 23.3% of the edges of Rnk23-clustered-noX, three parallel glosses with micro-clusters and six parallel edged with other clusters), display similarities with the layer constituted by micro-clusters C1–C7. Here, too, we are looking at many manuscripts that share a small number of glosses and sometimes do not transmit any other annotations. However, glosses belonging to cluster A appear consistently in the white space. Looking at their manuscript context, we can note that they represent a coherent set of annotations to the first three chapters of the first book of the *Etymologiae,* i.e., the very beginning of Isidore's encyclopedia.[80] They also inhabit a different geographical range, principally occurring in manuscripts from France, an indication that they may have originated in this region. Given their pattern of occurrence in Rnk23-clustered-noX, they may also be a remnant of an entity distinct in its age and character from other clusters, such as a larger body of glosses to the *Etymologiae,* of which only the opening sections remain due to the hazards of transmission. However, this entity is probably not as old as the glosses belonging to micro-clusters C1–C7, and perhaps not older than the early ninth century, since glosses from cluster A are not fossilized.

78 Bischoff, "Die europäische Verbreitung"; Ryan, "Isidore amongst the Islands."

79 Nevertheless, extrinsic clues indicate that some micro-clusters may have originated in the insular environment. Micro-cluster C4, for example, consists of citations from pseudo-Isidorean *De vitiis et virtutibus* that survive fully only in an Irish manuscript; see Schindel, *Die lateinischen Figurenlehren des 5. bis 7. Jahrhunderts und Donats Vergilkommentar;* Schindel, "Zur Datierung des Basler Figurentraktats (cod. lat. F III 15d)."

80 Steinová and Boot, "The Glosses to the First Book of the Etymologiae of Isidore of Seville." https://db.innovatingknowledge.nl/edition/#left-II.

Finally, we can establish a separate layer consisting of the other major clusters (Fig. 6c, 21.51% of the edges of Rnk23-clustered-noX, seven parallel glosses with micro-clusters and six parallel glosses with cluster A). While different clusters appear in this layer, they share certain commonalities. For example, manuscripts in this layer share glosses with a smaller number of manuscripts than in most other layers (thus, the average degree of this layer is 2.571, in contrast to 5.31 for layer C1–C7, and 10.286 for Rnk23-clustered-noX as a whole), although these connections are stronger (thus the average edge weight of this layer is 11.16, in contrast to 2.31 of layer A, 1.08 of C1–C7, and 3.88 of Rnk23-clustered-noX as a whole). Except for Laon447, which was annotated by a group of Irish and Carolingian scholars in Mainz,[81] all manuscripts belonging to this layer were annotated in France (particularly in the north), England, or Brittany.

We can further recognize two regional segments of this layer. First, we find manuscripts annotated in England and Brittany (CotCalAxv, GothaI147, Harley-3941, Paris7585, Queens320), as well as manuscripts annotated on the continent in insular-influenced milieus (IRHT342, Laon447) concentrated to the right of Harley3941 in Fig. 5a. These manuscripts are interconnected by three clusters, each common to at least four manuscripts: E (red, 50 glosses), G (dark green, 30 glosses), and Q (pink, 7 glosses). E and G are absent from manuscripts annotated in France, while Q appears in Paris11278, a manuscript annotated in southern France or northern Italy, which should be perhaps considered to reflect an insular influence on the grounds that it contains glosses from Q. All of the manuscripts mentioned above, apart from CotCalAxv and Paris11278, are codices of the complete *Etymologiae* into which glosses were copied. These insular clusters represent glosses that circulated in the early Middle Ages only or predominantly in areas under the insular influence, and are in all likelihood of insular origin.[82]

The parts of this layer corresponding to the main component positioned to the left of Harley3941 in Fig. 5a include, aside from this codex, manuscripts annotated in northern France, the area of most intense Carolingian intellectual activity. They display specific features when compared with the insular segment of this and other layers. First, the manuscripts belonging to this Carolingian Frankish segment feature clusters common to only one to three other manuscripts, but share glosses with many manuscripts in this manner.[83] Furthermore, except for the highway cluster F (brown, 157 glosses, 73 of which are rank 3), none of the edges visible in this part of the layer are constituted by more than five glosses with

81 Calloni, "Allegorizzare le 'Etymologiae': l'irlandese Probo e gli estratti esegetici del codice Laon BM 447."

82 Compare with Steinová, "Annotation of the Etymologiae of Isidore of Seville in Its Early Medieval Context," 29–37.

83 Orleans296/Paris7490 contains glosses from four clusters and Paris7670, Reims426, and VLO41 from three clusters.

rank 3, although glosses with rank 2 add significant weight to all of them. Clusters M (light blue, 17 glosses) and S (black, 21 glosses) contain only two glosses with rank 3, but M includes thirteen glosses and S fifteen glosses with rank 2, and cluster B (light green, 18 glosses) is built from three glosses with rank 3 but fourteen glosses with rank 2. By contrast, the insular clusters are constituted by more glosses with rank 3 than with rank 2 (G, with fifteen glosses with rank 3 and fourteen with rank 2), or only by glosses with rank 3 (E and Q). Finally, the only edges in Rnk23-clustered-noX that are constituted by no glosses with the highest particularity rank 3, corresponding to clusters I (yellow, 54 glosses) and N (dark purple, 29 glosses), appear in this segment.

Interpreting the last two clusters in their network context is difficult. On the one hand, we cannot dispel the possibility that they are phantoms, rather than historical entities since they are constituted only by glosses with rank 2; on the other hand, they are the second (I) and fifth (N) heaviest clusters in the dataset, constituted by more glosses than eight clusters whose genuineness as transmitted units is not in doubt. Cluster I, moreover, displays a peculiarly scattered pattern of gloss distribution in its two witnesses (Orleans296/Paris7490 and VLO41). In contrast to clusters constituted by many glosses with rank 3, E, F, and G, in which glosses appear concentrated in specific chapters of the first book of the *Etymologiae,*[84] the glosses in cluster I are spread across Orleans296/Paris7490 and VLO41 rather uniformly. As a result, they are interspersed by both isolated and other parallel glosses to such an extent that cluster I is invisible to the human eye, unlike clusters E, F, and G, which can be partially detected via close reading. To a lesser extent, the scattered pattern of gloss distribution also characterizes cluster N and other clusters from the Carolingian Frankish segment.

This distribution pattern, as opposed to the appearance of glosses in blocks, could be expected to arise as a result of either spontaneous composition or random parallelism. Whether we should assume that neither I nor N are genuine clusters, as is perhaps the case with parts of clusters B, M, O, and S, depends on how plausible we find the idea that spontaneous composition or random parallelism could generate phantom sets constituted by a large number of glosses, in particular more than ten glosses, the demarcation assumed for clusters in section 3.2 (see section 7). In the absence of relevant data, we can rely on traditional philological reasoning that tells us it is unlikely that cluster I, with its 54 glosses, and cluster N, with its 29 glosses, could be phantoms in their entirety. Instead, we can seek the explanation for the specific features of these two clusters, and others, in the historical processes that gave rise to them and yet that seem distinct

84 83% of glosses from cluster F appear in chapters 36–44 in Harley3941 and Orleans296, 64% of glosses from cluster E appear in chapters 37–40 in Gotha1147 and Paris7585, and 50% of glosses belonging to cluster G can be found in chapters 32–37 in Harley3941 and Paris7585.

from those described in this article so far. For example, it can be pointed out that manuscripts sharing these two clusters are grammatical handbooks rather than library books (nodes colored orange) and display paleographic features consistent with the use in or design for school use.[85] Since we have learned in section 3.3 that instruction seems to have been an important stimulus for the production of glosses to the first book of the *Etymologiae* in northern France, the peculiarities of clusters I and N, and of the Carolingian Frankish segment more broadly, may reflect the transmission of glosses in the context of instruction (e.g., involving oral transmission or selective collection with the aim of reuse for teaching).

6.3 Hubs in the network

The cluster-detection algorithm applied in Fig. 5a makes it evident that one manuscript sits at the intersection of all three layers described in section 6.2: Harley3941. As was established in the previous section, this codex stands out among manuscripts containing glosses to the first book of the *Etymologiae* due to its extensive gloss parallelism and should be considered a hub. Fig. 6c reveals that Harley3941 also acts as a bridge between otherwise disconnected insular and Carolingian Frankish segments of the large component in Rnk23-clustered-noX. Paleographic and philological evidence corroborate network analysis and visualization.[86] Harley3941 is a manuscript of the entire *Etymologiae* that was produced and annotated at the end of the ninth or the beginning of the tenth century in Brittany. Glosses from several different clusters can be shown to have been copied into it at the time of its production and during the following century, including a batch added by a single hand that corresponds to cluster G.[87] The latter act of copying suggests that this Breton codex was used to collect glosses of diverse origin, including material known only from the Carolingian environment or the insular world. The network properties of Harley3941 can be matched to its real-world status as what may be termed a depository manuscript, i.e., a codex dedicated to the accumulation of glosses for preservation and potential reuse.

While Harley3941 is the most evident hub in Rnk23-clustered-noX, four other manuscripts were flagged as potential hubs in section 5.4: IRHT342, Paris7585, Paris7670, and Schaffhausen42. Of these, Schaffhausen42, a codex produced in the second quarter of the ninth century in Mainz and annotated in the second half of the same century in St. Gallen, can be excluded from the list. It holds a prominent place in the layer of Rnk23-clustered-noX constituted by micro-clusters, as it contains the highest number of parallel glosses belonging to these micro-clusters (6) and is connected to the highest number of manuscripts trans-

85 Steinová, "Annotation of the Etymologiae of Isidore of Seville in Its Early Medieval Context," 13–15.
86 Ibid., 31–33.
87 Ibid., 17.

mitting them (18), but is otherwise not particularly central to the entire network. By contrast, Paris7585 and Paris 7670 feature in all three layers distinguished in Fig. 6, and IRHT342 appears in two of the three layers, not containing glosses from cluster A. Paleographic and philological evidence also identify IRHT342, copied in the tenth and annotated in the following century and a half in an unknown location, but showing a clear affinity to the insular world in its collection of glosses, and Paris7585, produced in France and annotated in Canterbury in the second half of the tenth century, as depository manuscripts.[88] Both assemble glosses from two insular clusters, E and G, which are transmitted separately in older manuscripts, most notably in the Breton GothaII47 (the most important witness of E) and Harley3941 (the most important witness of G). Furthermore, like Harley3941, these two manuscripts were annotated during or shortly after their production, and the glosses they contain can be shown to have circulated at least a century before the copying of the manuscripts; they are library books rather than schoolbooks, and they represent rare examples of manuscripts that attracted glosses not only to the first but to all books of the *Etymologiae*. The case of Paris7670 is more intriguing, as we lack paleographic and philological evidence that would classify it as a depository manuscript. Nonetheless, it has the properties of such a manuscript (e.g., it is a library book). Its network properties may be an indicator that it should also be considered a depository manuscript.

6.4 Gloss parallelism and geographic distribution

We can conclude the visual inspection of Rnk23-clustered-noX by visually comparing Figures 5a and 5b, noting which nodes from specific layers and segments are pulled together by the Yifan Hu algorithm even if they do not represent geographically close manuscripts, and which nodes are pushed apart even if they represent geographically close manuscripts. Such a discrepancy between geographic proximity and gloss parallelism is particularly notable in the case of Orleans296 and VLO41, the two manuscripts with the highest number of parallel glosses after Harley3941 (301 and 191, respectively) and the two most densely annotated surviving manuscripts of the first book of the *Etymologiae* (768 and 682 glosses, respectively). While these two manuscripts are connected by cluster I, the second heaviest edge in Rnk23-clustered-noX, the Yifan Hu algorithm places them on opposite sides of the Carolingian Frankish segment of the main component because of their otherwise distinct connectivity to other parts of the network. Nevertheless, both manuscripts have ties to the same location: Fleury, a monastery in central France. VLO41 was annotated there at the end of the ninth or during the early tenth century. Orleans296 was present in Fleury from the tenth century at the latest, and probably earlier.[89]

88 Ibid., 31–33 and 35.
89 Ibid., 50–54.

Another manuscript pair displaying a similar discrepancy is GothaI147 and Harley3941. These two codices, produced and annotated in Brittany, share only a single gloss with rank 3, although they contain many other glosses with this rank shared with manuscripts in England and France. In other cases, such as the Spanish RAH25 and RAH76, the northern Italian CesenaSXXI5 and VeniceII46, which we have established reflect the transfer of glosses from an exemplar to an apograph, we can observe a correlation between gloss parallelism and geographic proximity. This is also true for CotCalAxv, Paris7585, and Queens320, which seem to have been annotated in Canterbury[90] and share glosses from clusters E and G. A visual examination of Rnk23-clustered-noX is insufficient to reach a definitive conclusion about the relationship between geography and gloss parallelism. To explore this dimension of co-occurrence networks, we have to deploy a different strategy (see section 7).

6.5 The network visualization in perspective

In conclusion to this section, we can reflect on the utility of the network visualization for analyzing the corpus of glosses to the first book of the *Etymologiae*. While network visualization does not replace proper network analysis, this section has hopefully demonstrated how useful it is to complement the latter with the former, in particular as an exploratory technique that could direct scholars to avenues for further investigation[91], and how much can be gained from interrogating network visualization against the backdrop of the available extrinsic evidence. Firstly, the visualization exercise suggested that certain network properties can be a good match for extrinsic properties. To name but two examples, the detection of hubs could help us identify manuscripts used for the collection and preservation of glosses; and as specific network patterns seem to have been generated by different historical processes of gloss generation and transmission, their detection and analysis could provide us with crucial insights into how glosses were produced and circulated in medieval Latin-writing Europe. In addition, the visualization brought home how multilayered the corpus of glosses to the first book of the *Etymologiae* is, strengthening the observations made in section 5.4 based on the analysis of average and median degrees. Finally, it allowed for the inclusion of the chronological and geographical dimensions of the data, which were not directly tapped within the network analysis. Especially insofar as a more rigorous analysis of the geographical relationships between the witnesses of a gloss corpus is not feasible within a specific research project, the geo-sensitive visualizations can supply scholars with preliminary observations about regional trends and patterns.

---

90 Ibid., 31–32 and 35–36.

91 Compare with Fernández Riva, "Network Analysis of Medieval Manuscript Transmission. Basic Principles and Methods," 38; Lemercier and Zalc, *Quantitative Methods in the Humanities: An Introduction,* 129–36.

7. Avenues for further research and the limits of the method

The utility of the network-based approach outlined in this article is not restricted to the analysis of the network properties of co-occurrence networks (section 5), nor the visualization and interpretation of these networks in light of available scholarly evidence (section 6). It can be developed further to directly address specific research questions, as some of the network properties and models can serve as relevant proxies for historical processes or circumstances we wish to study. Unfortunately, the scope of this article does not allow us to develop such applications. Nevertheless, some potential uses of the network-based approach for the study of glosses and avenues for expanding and refining this method in the future can be sketched here.

- **Organicity of gloss corpora:** As we have seen in section 5.4, the degree of co-occurrence networks of gloss parallelism appears to be tied to the multilayered character of the corpus of glosses to the first book of the *Etymologiae,* and thus its organicity. It would be expedient to further test the utility of degree as a quantitative measure of the organicity of gloss corpora by comparing the degree distributions of co-occurrence networks constructed from different types of material (corpora of glosses with different expected levels of organicity/systematicity and standard textual traditions). If it turns out to be a good proxy for organicity, it could provide a basis for assessing gloss corpora quantitatively, as opposed to giving them a label based purely on qualitative assessment, and allow for a comparison of different gloss corpora.
- **The extent of spontaneous composition and random gloss parallelism:** It would be useful to model gloss parallelism due to spontaneous composition and randomness in order to better understand how co-occurrence networks due to transmission differ from those due to spontaneous composition and randomness, and what may be the extent and type of distortion that we should expect to observe in a co-occurrence network corresponding to real-world data. In both cases, one model that comes to mind is a type of random network model, a random intersection graph following hypergeometric distribution.[92] It is vital to further develop this or other random network models

92 On random network models in general, see Barabási, *Network Science,* chap. 3. Hypergeometric distribution corresponds to a situation in which objects (glosses) taken from a certain pool (e.g., Latin lexicon or its parts) are assigned to containers (manuscripts), either entirely randomly (random gloss parallelism) or according to specific criteria with an element of chance (spontaneous composition). Depending on the size of the pool and the number of objects assigned to a container, we should observe that the same objects are assigned to different containers and thus form a basis for the construction of co-occurrence networks at a certain rate with different probabilities, i.e., that co-occurrence networks constructed based on the co-occurrence of the same objects in different containers tend to have properties within certain ranges, and display properties in other ranges with

that could approximate random gloss parallelism and test their utility for modeling the historical process of spontaneous composition.

- **Identifying and distinguishing different transmission processes:** In the two analytical sections of this article, sections 5 and 6, it was noted that certain network patterns observed in our co-occurrence networks seem to reflect different types of transmission (e.g., copying from an exemplar to an apograph, transmission of fossilized glosses within the main text, and the collection of glosses in a depository manuscript), and that different transmission processes can be expected to generate different network properties and elements. For this reason, it could be productive to develop network models that simulate different types of transmission, in particular, genealogical transmission typically represented as a stemma and the collection of glosses in a depository manuscript, just as we can establish a network model to approximate spontaneous composition and random gloss parallelism.[93] In this way, network analysis could help to distinguish different transmission processes from one another in cases when extrinsic evidence is lacking.
- **Geographical distance as a factor in gloss parallelism:** The relationship between geographical distance and gloss parallelism could be explored more rigorously than through a visual comparison of network graphs. Taking the data from the corpus explored in this article, we could, for example, compare the distance between any two manuscripts containing parallel glosses in km (derivable from their GPS coordinates) with the extent of their gloss parallelism (represented by the number of parallel glosses or glosses of certain ranks they share) and plot them against each other.[94]
- **Gloss-hopping:** In this study, the network-based approach was used to examine the internal dynamics and structure of a single gloss corpus. In this respect, we did not stray from the traditional scholarly paradigm, which treats glosses to each text separately, i.e., as unique and distinct from those to other texts. It is still uncommon for scholars to acknowledge that such boundaries may be due to our modern scholarly perceptions and editorial needs rather than to medieval annotation practices.[95] As the network-based approach treats corpora and collections of glosses as pools, it can be used to trace gloss parallelism

---

a negligible probability. On hypergeometric distribution, see Pitman, *Probability,* 127. The random intersection graph model was developed in Singer, "Random Intersection Graphs."

93 See Hoenen, "The Stemma as a Computational Model," 229–30.

94 Since, as was explained in section 4.2, the GPS coordinates are only used to approximate the region of the glossing of manuscripts, we can hope to uncover only very general trends, for example that glosses in manuscripts from a certain region tend to be more similar than glosses in manuscripts coming from different regions. Alternatively, we could restrict ourselves to using manuscript pairs for which we know the precise location of origin (e.g., Canterbury for Paris7585) to make this experiment more precise.

95 For a rare example of the awareness of this issue, see Teeuwen, "The Impossible Task of Editing a Ninth-Century Commentary," 200–202.

> across the boundaries of text-defined corpora, transcending the compartmentalization of glosses by text to assess to what extent glossing was text-bound, and whether other boundaries may be more relevant for the understanding of medieval reality (e.g., because of the role of memorization of glosses as self-sufficient units). As long as we can formulate criteria for postulating gloss parallelism across different languages, we can similarly use the network-based approach to study trans-linguistic gloss parallelism, a phenomenon noted by scholars of vernacular glossing.[96]

After discussing the potential utility of the network-based approach to glossing, it is fitting to offer remarks on its general limits and specific avenues for improvement. First, it is hard to assess how robust the conclusions that can be obtained via this approach are in light of the loss of historical material on which they are built, a problem common to historical network research.[97] The fragmentary survival of annotated manuscripts should make us particularly cautious about interpreting the absence of evidence (e.g., in the form of isolated components or weakly connected segments in a network) as the evidence of absence without sufficient support of extrinsic evidence, and to keep in mind that the observed results represent minimalistic conclusions (e.g., gloss parallelism and the extent of transmission can always be assumed to have been higher than observed).

We also need to remember that the co-occurrence networks explore gloss parallelism, rather than gloss transmission or the social networks that facilitated this transmission. While gloss parallelism may reflect the historical process of transmission, obtaining information relevant for establishing transmission networks from co-occurrence networks may not be possible, as the information they record is, by rule, not rich enough for this purpose and is represented in a manner not compatible with, for example, constructing a stemma as a particular type of transmission graph.[98] Furthermore, as was explained in section 2.2, gloss parallelism can have different causes, and distinguishing parallel from shared glosses must be done based on criteria that are external to network analysis and requires a substantial degree of domain knowledge. Even if additional modes of data preprocessing other than particularity ranking and gloss clustering were applied to an organic corpus of glosses, it is unlikely that we could filter out all of the noise from the co-occurrence networks that can be constructed from this data. On the contrary, the more restrictive the criteria, the more likely it is that we will also lose relevant information. It remains to be seen whether the method can be fur-

---

96 Moran, "Language Interaction in the St Gall Priscian Glosses," 134–39; Lambert, "L'étude des gloses"; Bauer, "Different Types of Language Contact in the Early Medieval Celtic Glosses."

97 Knappett, "Networks in Archaeology," 28–29.

98 The presence of complete graph elements in the co-occurrence network, in particular, is irreconcilable with a transmission network model.

ther improved to target noise more efficiently without diminishing the quality of the data.

As for some of the blind spots of the method as described in this article, as we have seen in section 6.2, it may be valuable to pay attention to the gloss distribution within collections of annotation, as a substantially diffused distribution pattern is consistent with spontaneous composition and random gloss parallelism, and may therefore provide an argument for assuming gloss parallelism due to processes other than transmission. The method could be further developed to account for the relative position of glosses within a collection of annotations and to incorporate information about the paleography of the glosses to better represent the layered nature of certain collections of annotations. The current method also does not work with the temporal aspect of gloss parallelism, even though such information is available and could be used to create network graphs that account for this property, in the same way that the geographical aspect of gloss parallelism was explored above.[99] Finally, edges in the co-occurrence networks constructed and examined in this article were made undirected. However, it should be possible to incorporate directionality into network analysis and visualization, provided it is made clear that it does not represent the direction of transmission of material from one specific manuscript to another, but rather a general direction of transmission of material.

8. Conclusion

This article outlines a network-based approach that allows us to study organic corpora of glosses in their complexity. In contrast to traditional scholarship, which emphasizes particular forms of sequential textuality and transmission facilitated by copying from an exemplar to an apograph, the network-based approach allows us to operate on the subtler level of individual glosses and to take gloss parallelism, rather than transmission (or a specific type of transmission), as a point of departure. The chief advantage of the network-based method is that it allows researchers to work with historical material in the form in which it came down to us, without having to either adopt preliminary assumptions about that corpus (e.g., that all instances of philological similarity within the corpus are due to transmission or that glosses were transmitted as standard texts), or to discard some information insofar as it cannot be fitted into the narrow criteria imposed by traditional methods. At the same time, by adopting the strategies described in section 2 of this article (particularity ranking and gloss clustering), the network-based approach can account for transmission as a specific historical process of interest and larger textual units than glosses (clusters). As a result, we can fully

99 On temporality in historical network research, see Knappett, "Networks in Archaeology," 67–70.

map the internal structure of a corpus of glosses, keeping its multilayered character and heterogeneity in the picture and not sacrificing certain elements of the corpus just because they cannot be considered gloss traditions, families, or commentaries. A network can even serve as an editorial model, and a network graph can provide an alternative visualization strategy to a stemma.[100]

As for the specific conclusions that can be drawn about the glossing of the first book of the *Etymologiae* following the analysis carried out in sections 5 and 6, the most significant properties of this corpus seem to be its heterogeneity and regionality. As far as the surviving evidence can be assumed as being broadly representative of the character and circumstances of the glossing of the *Etymologiae* in the early Middle Ages, there appears not to have been any dominant gloss family or tradition transmitted by a large number of witnesses (although at least one identified cluster appearing in two manuscripts, F, stands out due to the large number of glosses it contains). Rather, we discerned thirteen different clusters of glosses that seem to reflect distinct glossing efforts undertaken in the ninth and the tenth centuries, and seven micro-clusters, which are probably remnants of glossing predating the Carolingian period.

Most clusters seem to have circulated regionally, such as E and G in the insular world, A and I in France, P in Spain, and D in northern Italy. Only the micro-clusters are diffused much more widely across the early medieval Latin-writing world, which is likely due to their substantial age and fossilized state. The intensity of the glossing seems to have been highest in northern France and lowest in the German area, from which no gloss cluster originating in the ninth and the tenth centuries survives. Importantly, the regionality of the glossing of the first book of the *Etymologiae* is also matched by transmission patterns. In both northern Italy and Spain, we can only evidence the transmission of glosses by copying from an annotated exemplar to its apographs; in the insular world and Brittany, the presence of three hubs (IRHT342, Harley3941, Paris7585) revealed in sections 5.4 and 6.3 is consistent with a preference for collecting glosses into a manuscript that serves as their depositories; and in northern France, the transmission of glosses may have been driven by instructional needs, as it has a different network pattern. A specific place in the landscape of the glossing of the first book of the *Etymologiae* should be accorded to Brittany, for which glosses circulating both in the insular world and in northern France can be shown to have been available and collected. Given the age and character of Harley3941, we should assume that Brittany benefited from glossing taking place in both northern France and the insular world in the previous 150 years.

Some of the methodological and theoretical points made in this article that deserve emphasizing are:

100 Steinová and Boot, "Editing Glosses as Networks: Exploring the Explorative Edition."

- **As gloss parallelism cannot always be attributed to transmission, it is necessary to engage in data pre-processing to reduce noise before carrying out network analysis.** The distortive effect of spontaneous composition and random gloss parallelism is demonstrated in section 5.4. It could be useful to model how much gloss parallelism should be taken as a baseline due to spontaneous composition and randomness in a co-occurrence network constructed following the principles outlined in this article. Given the information quality of data used for constructing Rnk23-clustered-noX in sections 5 and 6, the particularity ranking and gloss clustering outlined in section 2 seem to be efficient strategies for noise elimination.
- **While trivial parallel glosses cannot be considered shared by default, many of them were likely transmitted.** The corpus studied in this article, as many organic corpora of glosses, is constituted mostly by glosses too trivial to be treated as transmitted by default (i.e., assigned particularity ranks 1 and 2 in section 3.1). Gloss clustering can help determine whether they may have been part of a transmitted package of glosses. Even if we cannot claim that every trivial gloss in a given cluster must have been transmitted, they can be considered broadly indicative of transmission as long as they: a) feature in clusters constituted primarily or also by glosses particular enough to be considered transmitted (i.e., assigned particularity rank 3 in section 3.1); or b) appear in a cluster in volumes too large to be explainable by spontaneous composition and randomness alone; or c) we possess concrete extrinsic evidence that substantiates their transmission (as in the case of cluster G in Harley3941).
- **In highly organic corpora, glosses can be assumed to have circulated on their own or in very small units.** This has been shown in sections 5.6 and 6.2, particularly on micro-clusters C1–C7. Overall, the examination of the corpus of glosses to the first book of the *Etymologiae* has shown that glosses could be transmitted in relatively small units, i.e., in the range of five to ten glosses. This is probably partially an effect of the loss of evidence, as identifiable clusters may correspond to what had once been larger batches of glosses. Nevertheless, some glosses to the first book of the *Etymologiae* were evidently transmitted in the early medieval Latin-writing Europe in very small units or isolation (e.g., clusters A and Q). Therefore, we should be wary of overfocusing on large or easily detectable clusters just because they are more prominent or visible to the human eye. This should also make us wonder what the circulation of glosses in small units or isolation reveals about the character of glossing and gloss transmission in the Middle Ages and the potential 'attrition' of gloss clusters due to manuscript loss.
- **Some gloss clusters are poorly visible or invisible using traditional methods.** While several gloss clusters in the corpus used as a demonstrative case in this article could be detected and partially described via close reading (e.g., E, F, and G), some clusters are likely to escape traditional methods because of their relatively small size, low particularity, small number of witnesses, and dispersed gloss distribution in manuscripts. These include cases that involve such large volumes of parallel glosses (clusters I and N) that they can-

not be dismissed as phantoms conjured by spontaneous composition or random processes. The network-based approach described in this article may be particularly valuable for identifying 'invisible' gloss clusters. It is noteworthy that within the corpus studied in this article, these 'invisible' clusters predominate in the region of northern France, in which the extrinsic evidence suggests that glosses to the first book of the *Etymologiae* circulated in an instructional context. The two features may be interconnected, indicating that traditional methods may be blind to transmission processes that are particularly interesting to study.

- **Co-occurrence networks with a high concentration of nodes with high degrees may provide evidence for the process of the accumulation of glosses.** In section 5.4, it was shown that the co-occurrence networks constructed from data introduced in section 3 have relatively high average and median degrees, and in section 6.3 that nodes that appear as hubs in the constructed co-occurrence networks correspond to manuscripts that bear extrinsic signs of having been designed or used for collecting glosses. In one case (Paris7670), recognition of hubs may have even identified a depository manuscript, for which we lack extrinsic clues. It was proposed that the two properties may relate to the multilayered character of the corpus. If this hypothesis can be substantiated, the identification of hubs could be a quick way to trace annotated manuscripts in which glosses were collected, and the degree distribution could help us to establish to what extent the accumulation of glosses played a role in the evolution of a gloss corpus.
- **Different transmission processes may generate different network patterns.** In sections 5.4 and 6.1–2, we distinguished four network patterns that may be associated with different transmission processes. First, the copying of glosses from an annotated exemplar to its apograph (a process resembling the transmission of standard texts) corresponded to relatively thick edges between a small number of manuscripts that are isolated from other manuscripts due to their non-cumulative character. Second, the transmission of glosses within the main text due to their fossilization generated large, complete graph components with edges of minimal thickness. Third, the collection of glosses and copying of batches of glosses from one manuscript to another generated hubs. Finally, the transmission of glosses in an instructional context was connected with a network pattern in which many manuscript pairs or triplets are mutually connected with relatively light edges. These and perhaps other network patterns need to be examined further to ascertain whether they can be used to detect specific transmission processes with network analysis.

The network-based approach to glossing has much to offer. In the future, the network-based approach described and demonstrated in this article will hopefully be extended to new corpora of glosses, and thus its utility tested. In the process, the validity of the conclusions articulated here shall be ascertained. The more corpora that are probed with network-based methods, the more we are likely learn about the historical processes of generation and transmission of glosses,

and the better we will understand the character of glossing in Latin-writing medieval Europe and beyond.

9. References

Andrée, Alexander. "Anselm of Laon Unveiled: The *Glosae svper Iohannem* and the Origins of the *Glossa ordinaria* on the Bible." *Mediaeval Studies* 73 (2011): 217–60.

Barabási, Albert-László. *Network Science.* Cambridge: Cambridge University Press, 2017. http://www.networksciencebook.com/.

Bauer, Bernhard. "Different Types of Language Contact in the Early Medieval Celtic Glosses." *Proceedings of the Harvard Celtic Colloquium* 37 (2017): 33–46.

— "The Celtic Parallel Glosses on Bede's 'De Natura Rerum.'" *Peritia* 30 (2019): 31–52.

— "The interconnections of St Gall, Stiftsbibliothek, MS 251 with the Celtic Bede manuscripts." *Keltische Forschungen* 8 (2019): 31–48.

— "Venezia, Biblioteca Marciana, Zanetti Lat. 349. An Isolated Manuscript? A (Network) Analysis of Parallel Glosses on Orosius' Historiae Adversus Paganos." *Études celtiques* 45 (2019): 91–106.

Bellettini, Anna. "Il codice del sec. IX di Cesena, Malatestiano S. XXI.5: le *Etymologiae* di Isidoro, testi minori e glosse di età ottoniana." *Italia medioevale e umanistica* 45 (2004): 49–114.

Bischoff, Bernhard. "Die europäische Verbreitung der Werke Isidors von Sevilla." In *Isidoriana. Collección de estudios sobre Isidoro de Sevilla,* edited by Manuel Cecilio Díaz y Díaz, 317–44. León: Centro de estudios San Isidoro, 1961.

Brughmans, Tom, Anna Collar, and Fiona Coward. "Network Perspectives on the Past: Tackling the Challenges: Introduction." In *The Connected Past: Challenges to Network Studies in Archaeology and History,* edited by Tom Brughmans, Anna Collar, and Fiona Coward, 3–20. Oxford: Oxford University Press, 2016. https://eprints.soton.ac.uk/432433/.

Buringh, Eltjo. *Medieval Manuscript Production in the Latin West: Explorations with a Global Database.* Leiden: Brill, 2011.

Calloni, Carlo Giovanni. "Allegorizzare le 'Etymologiae': l'irlandese Probo e gli estratti esegetici del codice Laon BM 447." *Filologia Mediolatina* 29 (2022): 113–48.

Chiesa, Paolo. "The Genealogical Method: Principles and Practice." In *Handbook of Stemmatology: History, Methodology, Digital Approaches,* edited by Philipp Roelli, 74–87. Berlin: De Gruyter, 2020.

Conti, Aidan. "A Typology of Variation and Error." In *Handbook of Stemmatology: History, Methodology, Digital Approaches,* edited by Philipp Roelli, 242–52. Berlin: De Gruyter, 2020.

Contreni, John J. "The Carolingian Renaissance: Education and Literary Culture." In *The New Cambridge Medieval History 2: c. 700–c. 900,* edited by Rosamond McKitterick, 709–57. Cambridge: Cambridge University Press, 1995.

— "The Pursuit of Knowledge in Carolingian Europe." In *The Gentle Voices of Teachers: Aspects of Learning in the Carolingian Age,* edited by Richard Sullivan, 106–41. Columbus, OH: Ohio University Press, 1995.

Dionisotti, Anna Carlotta. "On the Nature and Transmission of Latin Glossaries." In *Les manuscrits des lexiques et glossaires de l'antiquité tardive à la fin du moyen âge: Actes du colloque international (Erice, 23–30 septembre 1994),* edited by J. Hamesse, 202–52. Leuven: Fédération internationale des instituts d'études médiévales, 1996.

Fernández Riva, Gustavo. "Network Analysis of Medieval Manuscript Transmission. Basic Principles and Methods." *Journal of Historical Network Research* 3 (2019): 30–49.

Ganz, David. "Book Production in the Carolingian Empire and the Spread of Caroline Minuscule." In *The New Cambridge Medieval History 2: c. 700–c. 900,* edited by Rosamond McKitterick, 786–808. Cambridge: Cambridge University Press, 1995.

Godden, Malcolm R., and Rohini Jayatilaka. "Counting the Heads of the Hydra: The Development of the Early Medieval Commentary on Boethius's Consolation of Philosophy." In *Carolingian Scholarship and Martianus Capella: Ninth-Century Commentary Traditions on* De Nuptiis *in Context,* edited by Mariken Teeuwen and Sinéad O'Sullivan, 363–76. Cultural Encounters in Late Antiquity and the Middle Ages 12. Turnhout: Brepols, 2011.

Guidi, Vincenzo, and Paolo Trovato. "Sugli stemmi bipartiti. Decimazione, asimmetria e calcolo delle probabilità." *Filologia italiana* 1 (2004): 9–48.

Hoenen, Armin. "The Stemma as a Computational Model." In *Handbook of Stemmatology: History, Methodology, Digital Approaches,* edited by Philipp Roelli, 226–41. Berlin: De Gruyter, 2020.

Hoey, Michael. *Textual Interaction: An Introduction to Written Discourse Analysis.* London: Routledge, 2001.

Holtz, Louis. "Glosse e commenti." In *Lo spazio letterario del Medioevo 1: Il Medioevo latino, 3: La ricezione del testo,* edited by Guglielmo Cavallo, Claudio Leonardi, and Enrico Menestò, 59–111. Rome: Salerno, 1995.

Jeauneau, Édouard. "Le Commentaire érigénien sur Martianus Capella (De nuptiis, Book I)." In *Quatre thèmes érigéniens,* 101–66. Montréal: Institut d'études médiévales Albert-le-Grand, 1978.

Kapitan, Katarzyna Anna. "Perspectives on Digital Catalogs and Textual Networks of Old Norse Literature." *Manuscript Studies* 6 (2021): 74–97.

Knappett, Carl. "Networks in Archaeology: Between Scientific Method and Humanistic Metaphor." In *The Connected Past: Challenges to Network Studies in Archaeology and History,* edited by Tom Brughmans, Anna Collar, and Fiona Coward, 21–33. Oxford: Oxford University Press, 2016. https://doi.org/10.1093/9780198748519.003.0007.

Lambert, Pierre-Yves. "Les commentaires celtiques a Bède le vénérable." *Études celtiques* 20 (1983): 119–43.
— "Les commentaires celtiques à Bède le vénérable." *Études celtiques* 21 (1984): 185–206.
— "L'étude des gloses: méthodes et instruments." *Britannia Monastica* 19 (2017): 45–82.
Lapidge, Michael. "The Study of Latin Texts in Late Anglo-Saxon England: The Evidence of Latin Glosses." In *Latin and the Vernacular Languages in Early Medieval Britain,* edited by Nicholas Brooks, 99–140. Leicester: Leicester University Press, 1982.
Lemercier, Claire, and Claire Zalc. *Quantitative Methods in the Humanities: An Introduction.* Translated by Arthur Goldhammer. Charlottesville, VA: University of Virginia Press, 2019.
Moran, Pádraic. "Language Interaction in the St Gall Priscian Glosses." *Peritia* 26 (2015): 113–42. https://doi.org/10.1484/J.PERIT.5.108317.
Moulin, Claudine. "Paratextuelle Netzwerke: Kulturwissenschaftliche Erschließung und soziale Dimensionen der althochdeutschen Glossenüberlieferung." In *Verwandtschaft, Freundschaft, Bruderschaft. Soziale Lebens- und Kommunikationsformen im Mittelalter,* edited by Gerhard Krieger, 56–82. Akten des Symposiums des Mediävistenverband 12. Berlin: De Gruyter, 2009.
Newman, Mark E. J. *Networks.* 2nd ed., Oxford: Oxford University Press, 2019.
Nievergelt, Andreas. "Glossen aus einem einzigen Buchstaben." In *The Annotated Book in the Early Middle Ages. Practices of Reading and Writing,* edited by Mariken Teeuwen and Irene van Renswoude, 285–304. Utrecht studies in medieval literacy 38. Turnhout: Brepols, 2018.
O'Sullivan, Sinéad. "Problems in Editing Glosses: A Case Study of Carolingian Glosses on Martianus Capella." In *The Arts of Editing Medieval Greek and Latin: A Casebook,* edited by Elisabet Göransson, Gunilla Iversen, Barbara Crostini, Brian M. Jensen, Erika Kihlman, Eva Odelman, and Denis Searby, 290–310. Studies and Texts 203. Toronto: Pontifical Institute of Mediaeval Studies, 2016.
— "Text, Gloss, and Tradition in the Early Medieval West: Expanding into a World of Learning." In *Teaching and Learning in Medieval Europe: Essays in Honour of Gernot R. Wieland,* edited by Greti Dinkova-Bruun and Tristan G. Major, 3–24. The Journal of Medieval Latin Publications 11. Turnhout: Brepols, 2017.
Palumbo, Giovanni. "The Genealogical Method: Criticism and Controversy." In *Handbook of Stemmatology: History, Methodology, Digital Approaches,* edited by Philipp Roelli, 88–109. Berlin: De Gruyter, 2020.

Peeples, Matthew A., Barbara J. Mills, Randall W. Haas Jr, Jeffery J. Clark, and John M. Roberts Jr. "Analytical Challenges for the Application of Social Network Analysis in Archaeology." In *The Connected Past: Challenges to Network Studies in Archaeology and History,* edited by Tom Brughmans, Anna Collar, and Fiona Coward, 57–84. Oxford: Oxford University Press, 2016. https://doi.org/10.1093/9780198748519.003.0010.

Pitman, Jim. *Probability.* New York: Springer, 1993.

Roelli, Philipp. "Definition of Stemma and Archetype." In *Handbook of Stemmatology: History, Methodology, Digital Approaches,* edited by Philipp Roelli, 209–25. Berlin: De Gruyter, 2020.

Rossi, Guido, ed. *Atti del Convegno internazionale di studi Accursiani.* Milano: Giuffrè, 1968.

Ryan, Martin J. "Isidore amongst the Islands: The Reception and Use of Isidore of Seville in Britain and Ireland in the Early Middle Ages." In *A Companion to Isidore of Seville,* edited by Jamie Wood and Andrew Fear, 424–56. Brill's Companions to the Christian Tradition 87. Leiden: Brill, 2020.

Schiegg, Markus. *Frühmittelalterliche Glossen: Ein Beitrag zur Funktionalität und Kontextualität mittelalterlicher Schriftlichkeit.* Germanistiche Bibliothek 52. Heidelberg: Winter, 2015.

Schindel, Ulrich. *Die lateinischen Figurenlehren des 5. bis 7. Jahrhunderts und Donats Vergilkommentar.* Abhandlungen der Akademie der Wissenschaften in Göttingen. Philologisch-Historische Klasse 91. Göttingen: Vandenhoeck & Ruprecht, 1975.

— "Zur Datierung des Basler Figurentraktats (cod. lat. F III 15d)." *Göttinger Forum für Altertumswissenschaft* 2 (1999): 161–78.

Singer, Karen B. "Random Intersection Graphs." Ph.D., The Johns Hopkins University, 1995. https://www.proquest.com/docview/304306614/abstract/B7DC2EB06B3D4803PQ/1.

Stagni, Ernesto. "Nell'officina di Paolo Diacono? Prime indagini su Isidoro e Cassiodoro nel Par. lat. 7530." *Litterae Caelestes* 4 (2012): 9–105.

Steinová, Evina. "Annotation of the Etymologiae of Isidore of Seville in Its Early Medieval Context." *Archivum Latinitatis Medii Aevi* 78 (2020): 5–81.

Steinová, Evina, and Peter Boot. "Editing Glosses as Networks: Exploring the Explorative Edition." *A Companion to Digital Editing Methods* 1 (2022): forthcoming.

— "The Glosses to the First Book of the Etymologiae of Isidore of Seville: A Digital Scholarly Edition." Amsterdam: Huygens Institute, 2021. https://db.innovatingknowledge.nl/edition/#left-home.

Teeuwen, Mariken. "Marginal Scholarship: Rethinking the Function of Latin Glosses in Early Medieval Manuscripts." In *Rethinking and Recontextualizing Closses. New Perspectives in the Study of Late Anglo-Saxon Glossography,* edited by Patrizia Lendinara, Loredana Lazzari, and Claudia Di Sciacca, 19–37. Textes et Études Du Moyen Age 54. Turnhout: Brepols, 2011.

— "The Impossible Task of Editing a Ninth-Century Commentary: The Case of Martianus Capella." *Variants: The Journal of the European Society for Textual Scholarship* 6 (2007): 191–208.
— "Writing in the Blank Space of Manuscripts: Evidence from the Ninth Century." In *Ars Edendi Lecture Series,* vol. IV, edited by Barbara Crostini, Gunilla Iversen, and Brian M. Jensen, 1–25. Stockholm: Stockholm University Press, 2016.
Trovato, Paolo. "Neo-Lachmannism: A New Synthesis?" In *Handbook of Stemmatology: History, Methodology, Digital Approaches,* edited by Philipp Roelli, 109–38. Berlin: De Gruyter, 2020.
Tura, Adolfo. "Essai sur les *marginalia* en tant que pratique et documents." In *Scientia in margine: études sur les Marginalia dans les manuscrits scientifiques du Moyen Âge à la Renaissance,* edited by Danielle Jacquart and Charles S. F. Burnett, 261–387. Sciences historiques et philologiques. Hautes études médiévales et modernes 88. Genève: Droz, 2005.
de Valeriola, Sébastien. "Can Historians Trust Centrality? Historical Network Analysis and Centrality Metrics Robustness." *Journal of Historical Network Research* 6, no. 1 (2021): 85–125. https://doi.org/10.25517/jhnr.v6i1.105.
Valleriani, Matteo, Florian Kräutli, Maryam Zamani, Alejandro Tejedor, Christoph Sander, Malte Vogl, Sabine Bertram, Gesa Funke, and Holger Kantz. "The Emergence of Epistemic Communities in the Sphaera Corpus: Mechanisms of Knowledge Evolution." *Journal of Historical Network Research* 3 (2019): 50–91. https://doi.org/10.25517/jhnr.v3i1.63.
Wieland, Gernot R. "The Glossed Manuscript: Classbook or Library Book?" *Anglo-Saxon England* 14 (1985): 153–73.
— *The Latin Glosses on Arator and Prudentius in Cambridge University Library, MS Gg. 5.35.* Studies and Texts 61. Toronto: Pontifical Institute of Mediaeval Studies, 1983.

Appendix I: Overview of the manuscripts containing glosses to the first book of the *Etymologiae*

Full Manuscript shelfmark	Shortened label	Date of glossing	Place and region of glossing	All glosses	Parallel glosses	Rank 1	Rank 2	Rank 3
Orléans, Bibliothèque municipale, MS 296 (pp. 1–32)	Orleans296	9th c., 1/2	Paris or Fleury (northern France)	768	294	29	201	64
Leiden, Universiteitsbibliotheek, Voss. Lat. O 41	VLO41	10th c.	Fleury (northern France)	682	190	25	136	29

Full Manuscript shelfmark	Shortened label	Date of glossing	Place and region of glossing	All glosses	Parallel glosses	Rank 1	Rank 2	Rank 3
London, British Library, Harley 3941	Harley3941	9th/10th c. and 10th c.	unknown (Brittany)	535	309	20	169	120
Paris, Bibliothèque nationale de France, Lat. 7670	Paris7670	9th c.	Paris (northern France)	353	126	19	79	28
Reims, Bibliothèque municipale, MS 426 (fols. 1–117)	Reims426	9th c.	Reims (northern France)	345	127	14	93	20
Paris, Bibliothèque nationale de France, Lat. 7490	Paris7490	9th c.	Paris or Fleury (northern France)	241	65	12	36	17
Paris, Bibliothèque nationale de France, Lat. 7585	Paris7585	10th c., 2/2	Canterbury (England)	225	129	7	46	76
Paris, Bibliothèque nationale de France, Lat. 7671	Paris7671	9th c.	unknown (northern France)	135	37	3	23	11
Paris, Bibliothèque nationale de France, Lat. 7559	Paris7559	9th c.	Paris (northern France)	116	42	11	22	9
Trier, Bibliothek des Bischöflichen Priesterseminars, MS 100 (fols. 1r–16)	Trier100	9th c.	unknown (France)	74	14	1	8	5
Leiden, Universiteitsbibliotheek, Voss. Lat. F 82	VLF82	9th c.	Paris (northern France)	71	27	1	18	8
Oxford, Bodleian Library, Junius 25 (fols. 134–151)	Junius25	9th c.	Murbach (German area)	60	17	2	15	0
Bologna, Biblioteca Universitaria, MS 797	Bologna797	9th c.	Area of Reims (northern France)	55	24	4	15	5

Full Manuscript shelfmark	Shortened label	Date of glossing	Place and region of glossing	All glosses	Parallel glosses	Rank 1	Rank 2	Rank 3
Paris, Bibliothèque nationale de France, Lat. 11278	Paris11278	9th c., 1/2	unknown (southern France?/ northern Italy?)	48	28	3	10	15
Institut de recherche et d'histoire des textes, BVMM, Collections privées, digitisation of CP 342	IRHT342	12th c.	unknown (France?)	47	35	3	10	22
Montecassino, Archivio dell'Abbazia, MS 320 (pp. 5–398)	Montecassino320	beginning of the 10th c.	unknown (Italy)	43	8	2	6	0
Gotha, Forschungsbibliothek, Membr. I 147	GothaI147	9th c., 2/4	unknown (Brittany)	42	34	0	1	33
Madrid, Real Academia de la Historia, MS 76	RAH76	c. 946	San Millán de la Cogolla? (northern Spain)	38	28	1	25	2
Oxford, Queen's College, MS 320	Queen320	end of the 11th c./ beginning of the 12th c.	Canterbury (England)	38	28	0	6	22
Montpellier, Bibliothèque interuniversitaire, H 53 (fols. 5–265)	MontpellierH53	10th/ 11th c.	unknown (eastern France)	33	14	4	5	5
Munich, Bayerische Staatsbibliothek, Clm 6411	Clm6411	9th c., 1/4	Passau? (German area)	30	8	2	6	0
Chartres, Bibliothèque municipale, MS 16	Chartres16	11th c.	unknown (France)	29	8	3	4	1
Madrid, Real Academia de la Historia, MS 25	RAH25	c. 954	San Pedro de Cardeña (northern Spain)	29	28	1	25	2

Full Manuscript shelfmark	Shortened label	Date of glossing	Place and region of glossing	All glosses	Parallel glosses	Rank 1	Rank 2	Rank 3
Vatican, Biblioteca Apostolica Vaticana, Barb. Lat. 477 (fols. 4–123)	BarbLat 447_4	Beginning of the 11th c.	unknown (France)	29	3	2	1	0
Vatican, Biblioteca Apostolica Vaticana, Pal. Lat. 1746	PalLat1746	9th c.	Lorsch (German area)	27	6	4	2	0
Paris, Bibliothèque nationale de France, Lat. 7583	Paris7583	9th c., 2/2	unknown (northern France)	25	6	0	3	3
Cesena, Biblioteca Malatestiana, S.XXI.5	Cesena SXXI5	9th c. and 10th/11th c.	unknown (northern Italy)	20	15	1	3	11
Munich, Bayerische Staatsbibliothek, Clm 6250	Clm6250	9th c. and 10th/11th c.	Freising (German area)	15	9	1	3	5
Vatican, Biblioteca Apostolica Vaticana, Vat. Lat. 5763 (fols. 3–80)	VatLat5763	9th c.	Bobbio? (northern Italy)	15	5	1	3	1
Bern, Burgerbibliothek, MS 101	Bern101	9th c., 1–2/3	Loire area (France)	13	9	3	2	4
Paris, Bibliothèque nationale de France, n.a.l. 2633 (fols. 18–19)	ParisNAL2633	9th c., 4/4	unknown (France)	12	2	0	1	1
Venice, Biblioteca Marciana, II 46	VeniceII46	11th/12th c.	unknown (northern Italy)	11	11	0	0	11
London, British Library, Cotton Caligula A xv (fols. 3–38, 42–64, 73–117)	CotCalAxv	12th c.	Canterbury (England)	10	7	0	1	6
Munich, Bayerische Staatsbibliothek, Clm 4541	Clm4541	9th c., 3/3 and 11th c., 2/2	Benediktbeuern (German area)	10	8	1	3	4

Full Manuscript shelfmark	Shortened label	Date of glossing	Place and region of glossing	All glosses	Parallel glosses	Rank 1	Rank 2	Rank 3
Schaffhausen, Stadtbibliothek, Min. 42	Schaffhausen42	9th c., 2/2	Mainz and St. Gallen (German area)	8	6	0	1	5
Oxford, Bodleian Library, Auct. T.2.20 (fols. 2–124)	AuctT2.20	9th c., 3/4	Auxerre (northern France)	7	2	0	2	0
Laon, Bibliothèque Suzanne Martinet, MS 447	Laon447	9th c., 2/4	Mainz (German area)	7	5	1	2	2
London, British Library, Arundel 129	Arundel129	unknown	unknown	6	1	0	0	1
Wolfenbüttel, Herzog August Bibliothek, Weiss. 64	Wolfenbuttel64	9th c.	unknown (France?)	6	4	0	1	3
Berlin, Staatsbibliothek, Ham. 689	Ham689	11th c.	unknown (northern Italy)	4	2	0	0	2
Leiden, Universiteitsbibliotheek, BPL 122	BPL122	9th c., 4/4	Lyon (southern France)	3	0	0	0	0
Paris, Bibliothèque nationale de France, Lat. 10293	Paris10293	9th c., 3/4	Reims (northern France)	3	2	1	0	1
Milan, Biblioteca Ambrosiana, L 99 sup.	MilanL99sup	8th c., 2/2	Bobbio (northern Italy)	3	2	0	1	1
London, British Library, Harley 3099	Harley3099	12th c., 2/3	Munsterbilsen (German area)	2	1	0	1	0
Leiden, Universiteitsbibliotheek, Voss. Lat. O 15	VLO15	11th c., 1/2	Limoges (Southern France)	2	0	0	0	0
Paris, Bibliothèque nationale de France, Lat. 7588	Paris7588	unknown	unknown	2	2	0	2	0
Vatican, Biblioteca Apostolica Vaticana, Reg. Lat. 1953	RegLat 1953	9th c., 1/4	Orléans (northern France)	2	0	0	0	0

Full Manuscript shelfmark	Shortened label	Date of glossing	Place and region of glossing	All glosses	Parallel glosses	Rank 1	Rank 2	Rank 3
Vatican, Biblioteca Apostolica Vaticana, Barb. Lat. 477 (fol. 3)	Barb-Lat447_3	beginning of the 11th c.	unknown (France)	1	0	0	0	0
Bern, Burgerbibliothek, MS 611 (fols. 42–93)	Bern611	8th c., 1/2	Bourges (southern France)	1	0	0	0	0
Brussels, Koninklijke Bibliotheek, II 4856	Brussels-II4856	end of the 8th c.	Corbie (northern France)	1	1	0	0	1
Cologne, Dombibliothek, MS 123 (fols. 76–80)	Cologne123	9th c., 4/4	unknown (eastern France)	1	1	0	1	0
London, British Library, Harley 2713 (fols. 1–34)	Harley-2713	9th c., 4/4	unknown (northern France)	1	0	0	0	0
London, British Library, Harley 5977 (fol. 71)	Harley-5977_71	unknown	unknown	1	0	0	0	0
Reims, Bibliothèque municipale, MS 425	Reims425	mid-9th c.	Reims (northern France)	1	1	0	0	1
Total:	–	–	–	4,286	1,732	182	993	557

For Appendix II (Edge) and Appendix III (Node) please see https://zenodo.org/record/8146577.