Book Section:
Pagé-Perron, Émilie (2018) Network Analysis for Reproducible Research on Large Administrative Cuneiform Corpora. In: CyberResearch on the Ancient Near East and Neighboring Regions. Digital Biblical Studies. Brill, pp. 194-223.

https://doi.org/10.1163/9789004375086_008

CHAPTER 6

Network Analysis for Reproducible Research on Large Administrative Cuneiform Corpora

Émilie Pagé-Perron

Introduction

Although network analysis is becoming a more commonly utilized methodology in the study of Mesopotamian texts,[1] it is usually still employed on unique archives or small corpora. This is because there are no automated tools to reliably annotate large corpora, of which the administrative genre is the least annotated. Until we devise Natural Language Processing (NLP) tools with efficient machine learning components,[2] annotation requires expert knowledge and is very time-consuming. I wish to ease this burden by rethinking the ways in which we can simplify the process of extracting and analyzing relationships among entities.

Network analysis techniques can both expand Assyriological horizons in the study of Mesopotamian social history and enable research reproducibility. Network analysis permits one to approach the data in large amounts of texts through a new lens, by inquiring quantitatively with a focus on the relationships among entities of interest.[3] Such inquiries facilitate the detection of meaningful patterns that could not be easily perceived using traditional methods. Furthermore, when combined with practices of open access, open data, and method disclosure, network analysis research can be fully reproducible.

1 My deepest thanks to the editors, readers, and reviewers of earlier drafts of this chapter. A special mention is in order for Vanessa Bigot Juloux, Amy Rebecca Gansell, Terhi Nurmikko-Fuller, and Sarah Whitcher Kanza for their insightful comments and suggestions. Of course, any errors are my responsibility alone.

2 Regarding Natural Language Processing, see also in this volume, Prosser, 322, and Svärd, Jauhiainen, Sahala, and Lindén, 246. The Machine Translation and Automated Analysis of Cuneiform Languages project is working on this task (<https://cdli-gh.github.io/mtaac/> [accessed June 12, 2017]).

3 In textual studies, network analysis is a digital methodology that focuses on the relationships among entities in the written record and enables these relationships to be studied on a larger scale than is normally feasible with traditional philological methods.

 | DOI 10.1163/9789004375086_008

Thus, a traditional philological approach can be supported by quantitative, reproducible, and fully verifiable arguments.

This chapter will show how network analysis can provide new insights into large amounts of data from cuneiform texts. I will demonstrate a method for the preparation and extraction of relevant data for network analysis, as well as present some techniques for graph visualization. A discussion will follow explaining how the methods themselves and the results they yield support research in Mesopotamian social history.

Background

Legal and administrative cuneiform documents make up the largest subset of textual material from ancient Mesopotamia. These clay tablets contain essential details about the economic, social, religious, and political practices of the ancient societies that created them. Although individually these texts are short and contain little information, when grouped and analyzed as a corpus, they can offer insights into complex social topics. Despite comprising the most numerous type of surviving cuneiform documents, administrative texts are the least-annotated genre of Mesopotamian sources and are thus not prepared adequately for network analysis. Because they also remain untranslated, they have the additional drawback of being inaccessible to specialists in adjacent disciplines who are not trained to read cuneiform.[4] Through digital processing, however, these limitations may be mitigated.

Administrative texts mostly document transactions involving people, things, actions, and places. By using graphs and, more precisely, network theory,[5] scholars can examine the relationships among entities present in the text from new perspectives. The current practice in social network analysis based on cuneiform sources uses verbs to create directed links between named entities or concepts.[6] This requires an additional layer of analysis in which the researcher identifies not only verbs that are meaningful in the context of relationships but also who performs the action and who benefits from it. This type of analysis is manageable for smaller corpora—texts in the hundreds—but

4 Pagé-Perron et al. 2017.

5 Network theory: the study of systems and patterns found in network graphs. Social network analysis uses network theory to understand social relationships.

6 In this volume, see Bigot Juloux (161, 166–181), who proposes analytical taxonomies to analyze the relation between verbs and animated entities.

when a corpus has over a thousand texts, this manual operation becomes a serious investment of time.

When scholars extract information from cuneiform texts, they usually organize it in lists or tables to help with classification. In reality, when the focus of interest is, for example, the people, places, institutions, and goods present in the text, it is the relationships among these entities that enrich our understanding of them. The network graph presents a data structure that emphasizes the relationship among entities and can be analyzed and traveled (meaning queries can follow pathways from node to node) in a computationally less expensive (i.e., rapid and efficient) way compared with processing similar inquiries using text files or relational databases such as MySQL.[7]

The use of network analysis is well established in some disciplines, such as Biology and Sociology, but it is only slowly gaining popularity among Assyriologists. The first important work of prosopographic research using network analysis was Caroline Waerzeggers' study of Marduk-rēmanni's archive; she published a network analysis method overview the same year.[8] Allon Wagner and colleagues have prepared a method paper with interesting examples, modern and ancient, to demonstrate the process of network analysis.[9] Sara Brumfield's research on Akkadian imperial policy over controlled provinces benefits from Waerzeggers' innovative use of text mining techniques and network analysis but in novel ways to answer her research questions.[10] Further contributing to the Assyriological use of network analysis, David Bamman and colleagues published a statistical model that can infer missing elements in a network, and Eduardo Escobar has recently utilized semantic network analysis to trace the identity of ancient Mesopotamian ingredients.[11]

---

7 MySQL is a relational database software that employs Structured Query Language (SQL) to fetch data. Relational databases are a type of data format that uses unique identifiers (ID) to represent data from one table in another, so that when data is queried using SQL, it is possible to fetch related information from multiples tables at once.

8 Waerzeggers 2014a; 2014b.

9 Wagner 2013.

10 Brumfield 2013. The Old Akkadian period lasted approximately from 2340 to 2200 BCE. The "Old Akkadian period" describes a period in Mesopotamian history during which this region saw the development of an important political entity that, at its apogee, extended almost from the Mediterranean in the west, to the Persian Gulf in the south, and to the Zagros mountains in the east. The Sumerian and Akkadian languages were written using the cuneiform script. Prominent Mesopotamian rulers from the Old Akkadian period include Sargon and Naram-Sin. For another text-mining method, see in this volume, Bigot Juloux, 181–187.

11 Banman et al. 2013; Escobar 2017.

Problem

The problem explored in this chapter concerns the following questions: How it is possible to manipulate information from larger cuneiform corpora with network analysis methods–especially in the case of unannotated texts? And how can this quantitative approach support the researcher in inquiries about Mesopotamian social history?

As previously established by Waerzeggers, network analysis has the advantage of providing an overview of the larger networks in which individuals of interest inscribed themselves.[12] Social network analysis's main task is to explore social relations, and the network graph data structure is the most appropriate data type to focus on relationships between entities. But there is an additional advantage to employing quantitative research methods: the possibility of generating reproducible research. As explained by Ben Marwick, a researcher can take dispositions, such as sharing data and code,[13] to increase the reproducibility of their work. In turn, papers published in this context are generally more widely read and cited.[14]

With these preceding questions in mind, this chapter provides an overview of the methods and tools used to extract and organize information from Mesopotamian cuneiform administrative texts. It then discusses the processes necessary to convert this data into network graph data, which enables the exploration of the relationships among entities. Transcription, standardization, tokenization, lemmatization, and information extraction are explained, followed by an introduction to exploratory visualization and graph manipulation algorithms. Figure 6.1 shows the flow of the tasks required to prepare graph data for network analysis.

Three major steps comprised in this method are pre-processing the source texts,[15] extracting pertinent information, and analyzing this extracted information. Using exploratory visualization and network analysis algorithms enables the detection of meaningful groups of individuals from the sources.[16]

12 Waerzeggers 2014b.

13 Sharing data and code can be achieved by offering a copy in a public repository and using a permissive license, such as the MIT license, for software and a creative commons attribution license for data. A practical aspect of public repositories is the possibility of collaboration for the maintenance and further development of the product.

14 Marwick 2016.

15 Pre-processing: the task of preparing the (textual) data before starting the annotation process or some computational analysis.

16 In computer science, an algorithm is a set of instructions that a computer can execute to solve a (mathematical) problem. Network analysis algorithms: algorithms specifically geared to network analysis.

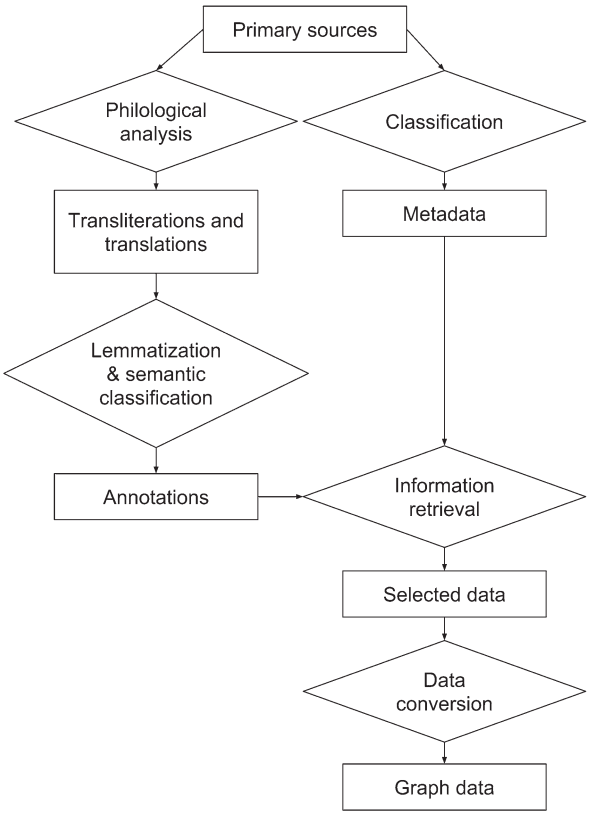

FIGURE 6.1
Information processing flowchart

A corpus of approximately 2,700 texts from the city of Adab will serve as the main dataset, and the examples used will showcase the relationships among the people who appear in the texts, based on their co-occurrence in the sources.[17] The discussion section below follows up on how these steps can be made reproducible and how they can enhance traditional approaches.

Sources Acquisition and Pre-processing

First, the text corpus must be digitized. Thankfully, there are plenty of already digitized corpora available online that can be reused for quantitative inquiries. The Cuneiform Digital Library Initiative (CDLI) and the Open Richly Annotated

17 It is possible to use this method with other types of entities. For instance, Escobar (2017) uses network analysis to better understand ancient recipes.

Cuneiform Corpus (Oracc) both offer a wide range of prepared transliterations that can be used in the following steps.[18] Let us first consider CDLI. CDLI is the largest database of cuneiform artifacts; it seeks to collect information about all Mesopotamian inscribed objects. It stores this data as metadata and images named "fatcrosses,"[19] as well as transliterations, normalizations, and translations in various languages. The full search page of the CDLI website is complex but also powerful for the advanced user.[20] The Search Aid page explains the possibilities of each search field.[21] One can, for instance, find the exact way to indicate a specific time period or place name in order to restrict a search query. Catalog metadata and textual data, when available, can also be downloaded freely from the CDLI search results page.[22] For example, searching for "Ur-Ninsun,"[23] an inhabitant of the Mesopotamian city of Adab in the Old Akkadian period (2340–2200 BCE), will bring up all of the texts in which this person, and potentially his homonyms, occurs. Transliterations can thereafter be collected via a download link.

One of CDLI's roles is to maintain a complete catalog and digital copy of all cuneiform texts. Contributions are updated periodically by CDLI staff based on new research, changes in standards, and direct contributions from schol-

18 Transliteration: transcription of cuneiform inscriptions into a romanized rendering of the cuneiform sign readings. This includes marking word boundaries and some structural markers, such as line numbering, object surfaces, etc. See Table 6.1 of this paper for an example of transliteration. For example, if one finds 𒈗 on a cuneiform tablet, the transliteration would read "lugal," which means king. For CDLI, see <https://cdli.ucla.edu> (accessed May 19, 2017), and for Oracc, see <http://oracc.org> (accessed May 19, 2017). See also in this volume, Eraslan (285), who discusses special characters specific to C(anonical)-ATF, which is used to encode transcriptions of cuneiform signs in CDLI. For a practical example of Oracc data manipulation for semantics research, see in this volume Svärd, Jauhiainen, Sahala, and Lindén, 227–229. See also in this volume, Bigot Juloux, 166, 166n73, 172.

19 Metadata is information about the text that is external to the textual information itself and can be used for classification purposes. It may describe provenience, period, genre, etc. See also in this volume, Matskevich and Sharon, 38.

20 CDLI: <http://cdli.ucla.edu/search/> (accessed May 19, 2017).

21 <http://cdli.ucla.edu/?q=cdli-search-information> (accessed May 19, 2017).

22 Most texts do not have an accompanying translation, but if the user searches with the expression "/./" in the translation field, the search engine will interpret the period as meaning any character and fetch entries that have at least one character in a translation line.

23 This search query must be entered in the transliteration field using the notation "ur-{d}nin-sun2," based on the C-ATF encoding, which will be discussed below. For additional information about C-ATF, see in this volume, Eraslan, 285–286.

ars.[24] Additionally, CDLI has backups at other sites, presently in Oxford, Berlin, and Toronto. Because of these characteristics, CDLI is the ideal place to store cuneiform editions, as it will stand the test of time. When preparing transliterations in digital form for the first time, contributing the results to a well-known database such as CDLI will enable the preservation and accessibility of a researcher's work.

Oracc presents itself as a platform hosting sub-projects that are managed by their contributing teams.[25] CDLI and Oracc both use standardized encoding schemes that are based on the same original format: ATF. ATF was created by CDLI as a stable archiving format for the long-term storage of texts.[26] It has evolved, first to adapt to the usage of CDLI, and later it branched out into Oracc-ATF. Specific characteristics of Oracc-ATF include a wider array of characters permitted in the transliteration lines and the validation of data content managed at the project level. As a result of these characteristics, transcription standards vary. Whether one is working directly from the ancient tablet or from paper publications, the ATF format is an excellent choice for encoding one's work for long-term preservation. Preparing data in a machine-actionable format is essential for its reuse.[27] The Oracc help pages provide instructions with the most complete documentation about both ATF formats.[28] Other websites offer digital transcriptions of cuneiform texts, but some of those websites fail to meet the requirements for a strict and machine-actionable encoding format that is necessary for the success of the next steps. Note that checking licenses is essential, since some sites do not permit reuse of their data.

---

24 Properly prepared transliterations and translations in the C-ATF format made using the guidelines can be sent to cdli@ucla.edu. Quick pointers can be found here: <http://cdli.ucla.edu/?q=support-cdli> (accessed May 19, 2017).

25 To contribute to Oracc, see this web page: <http://oracc.museum.upenn.edu/doc/about/contributing/index.html> (accessed May 19, 2017).

26 Koslova and Damerow 2003.

27 Machine-actionable (also called "machine-readable") data is structured in a way that it can be processed by computer software. For other encodings that are easily machine-readable, see in this volume, Matskevich and Sharon, 46; Bigot Juloux, 163–164; Nurmikko-Fuller, 339–340, 353–354.

28 For the C-ATF Primer, see <http://oracc.museum.upenn.edu/doc/help/editinginatf/primer/index.html> (accessed May 19, 2017) for Oracc-ATF. There are also quite a few other help pages that can be useful when encoding cuneiform textual data to ATF. See, for example, <http://oracc.museum.upenn.edu/doc/help/> (accessed May 19, 2017). For an example of an encoded transcription of cuneiform in CDLI, see in this volume, Eraslan, Table 9.1, 286.

Network Analysis

Network graphs are composed of nodes and links; nodes are data points representing entities.[29] Edges link nodes, representing the relationships between those nodes. This data takes the form of "triples" comprising two nodes and an edge that connects them (Fig. 6.2). In network analysis, nodes represent entities that can be of one or more types, such as people, institutions, or places. Edges between entities can be directed and weighted. Directed edges will not be discussed here, since we are only using the co-occurrence of entities in the texts to observe relationships; as such, these relationships are not directed.[30] The weight of an edge here represents the quantity of co-occurrences of two individuals in the corpus. It is important to note that because a relationship is assumed from the co-occurrence of individuals on the same tablet, the nature of this relationship is not similar to a network of Facebook users or employees of a company at their downtown building. For instance, long cuneiform ration lists enumerate people who might never have met but who were part of the same redistribution scheme. The relationship between them can be viewed as weak.[31] However, preparing a graph from this information makes it possible to observe the big picture of the social organization of the workforce, and it is especially helpful when working with a large number of sources.

Information Type

Administrative texts are typically short and formulaic,[32] with simple sentences and few actions. This type of text is thus an ideal candidate for network graph analysis in which individuals who appear in the text are chosen as entities that will become nodes in the final graph, and their co-occurrence generates links that connect them. The corpus that will be used for the examples below com-

29 Network graph: the total of all data points, the nodes, and their relationships—i.e., the edges. Some nodes can be completely disconnected from others and form independent ensembles. Together, connected and disconnected nodes form the network graph.

30 The directionality of a link usually represents who acts upon or toward whom. In these cases, the relationship is more precise than a co-occurrence in a text. For example, if A delivers something to B, the edge can be named "delivers to" and has a direction, from A to B. Linked data is a directed network graph in which all edges are named and directional.

31 Brumfield (2013) has already noted this in her work.

32 Administrative texts of the Old Akkadian period are formulaic in the sense that they mostly consist of lists of people, things, or both, with verbs that have specific, technical meanings in particular contexts. There is little deviation from the usual formula.

prises around 2,700 administrative texts, all dated to the third millennium BCE and provenienced in Adab (a city in Southern Iraq). The corpus is mostly restricted to the Old Akkadian period (2340–2200 BCE).[33] The network it creates has over 800 individuals and 7,000 relationships.

The method proposed in this chapter can be applied to any corpus of administrative texts. A good strategy is to delimit a corpus based on region and temporality and to trim the unnecessary sources from the lot. By comparison, searching for a type of product or concept and harvesting the transliterations from this search query can be problematic; texts bearing the name of important individuals that appear in other contexts will be omitted too. These high-ranking individuals are often bridges, that is, nodes that connect otherwise unconnected groups of nodes together. They are usually very important in understanding social structure using network analysis.

Information about the individuals in a corpus is traditionally stored in the form of lists or as a relational database. These data structures do not, however, facilitate the manipulation of the relationships between the entities. To mitigate this rigidity, the data must be converted to graph data based on triples (Fig. 6.2).

Such data takes the form of triples,[34] where the triple always has two nodes and an edge (individual 1, individual 2, and the relationship between them). Their edge could be weighted, since those individuals can appear concomitantly in texts more than once. In the examples here, labels are attached to the nodes, so the names of the individuals represented by these nodes can be read while manipulating the network graph.[35] For instance, if the ancient Mesopotamian textile workers Geme-Enlil and Mama-ummi appear in the same text,[36] both of them would be nodes, and the edge that links them represents the text in question. If Geme-Enlil and Mama-ummi appear together in nineteen texts, then the edge linking them will be made visibly thicker and have a value of 21.[37] The individuals and their interpersonal relationships thus form a graph.

---

33 For a preliminary network graph of individuals appearing in third-millennium BCE Adab texts, Pagé-Perron 2017, (<http://irkalla.net/adab/> [accessed May 22, 2017]).

34 For another meaning, see in this volume, a) Matskevich and Sharon, 46; b) Nurmikko-Fuller, 345–347 (especially for online-publishing purposes).

35 Label: a form of attribute used to identify nodes and edges easily instead of using only their unique identifiers.

36 Such as in the texts MS 4049 and Lippmann Collection 189 (<http://cdli.ucla.edu/search/search_results.php?SearchMode=Text&ObjectID=P472489,P253146> (accessed May 20, 2017).

37 Searching for "geme2-{d}en-lil2, ma-ma-um-mi" in the transliteration field of the CDLI search page (<http://cdli.ucla.edu/search/> [accessed June 4, 2017]) yields the result that

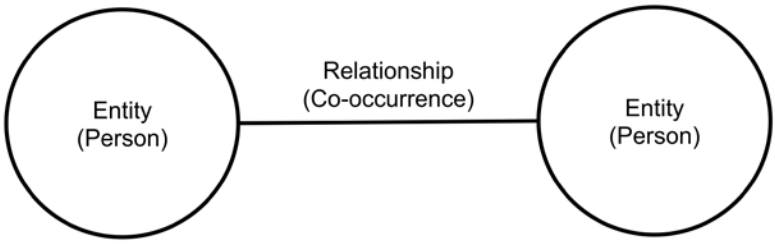

FIGURE 6.2 *A triple*

Tokenization and Lemmatization

Tokenization and lemmatization are the next steps after the homogenization of the transliterations, meaning after the sign readings, hyphenation, and structural notation are made consistent throughout the corpus. "Tokenization" is the segmentation of a text into similar units called tokens. Tokens can be sentences or signs, but, in this case, the texts are divided at the word level. They are then assigned to lemmata, which are also called "dictionary entries," in a process called "lemmatization."[38] Annotations preserve this new information about the text. Various annotation methods employ XML and TEI, popular markup languages. Oracc uses inline glossing as a solution for adding informa-

---

these two individuals appear together 21 times. See such search results here: <https://cdli.ucla.edu/search/search_results.php?SearchMode=Text&requestFrom=Search&TextSearch=geme2-%7Bd%7Den-lil2,ma-ma-um-mi> (accessed June 4, 2017). Examples of such texts are Cornell University Studies in Assyriology and Sumerology (CUSAS) 20, 066: <http://cdli.ucla.edu/P325058> (accessed June 4, 2017) and CUSAS 20, 067 (<http://cdli.ucla.edu/P323404> [accessed June 4, 2017]). Following is this last text (the June 4th version in CDLI):

```
&P323404 = CUSAS 20, 067
#atf: lang sux
@tablet
@obverse
1. [1(asz@c)] […] x
2. [1(asz@c)] ma#-ma-um-mi
3. 1(asz@c) geme2-{d}en#-lil2
4. 1(asz@c) GI4-NE-LI
5. 1(asz@c) da-[ni2]-a
6. 1(asz@c) nin-AD2-gal
@reverse
$ blank
```

38 Lemma (plural: lemmata): the headword used in a dictionary entry.

tion directly into the transliteration.[39] An alternative approach is to keep the annotations in spreadsheets or a database outside of the studied transliterations with formats such as CoNLL.[40] Niek Veldhuis designed a scraper specialized in extracting the lemmata assigned to each token of a text annotated in Oracc.[41] When used on an already annotated text, this type of tool will make it easier to gather specific types of entities: for instance, people in a text are tagged with an indicator that the word is a personal name: "`[PN]`". Disambiguation among people in a corpus with the same name is achieved by assigning a unique number to identify each one.

XML annotations, including TEI, are perhaps the most popular means of annotating texts across disciplines.[42] These types of annotations provide two major advantages: (1) ease of marking (with precision) the exact position of the annotated entity in the text and (2) flexibility of the markup,[43] making it possible to annotate multiple layers of information in one text. A good example of a successful annotated corpus available on the web using XML is the Electronic Text Corpus of Sumerian Literature (ETCSL).[44]

One of the possible ways to preserve complex relationships among elements of information is to use a relational database, in which unique ID represent recurring information in conjunction with intermediary tables that associate elements from different tables.[45] In a model to store texts, the tokens it contains, and a glossary, at least five tables are required: texts, tokens, and words, supplemented with relational tables that store the relationships among

39 Inline glossing results in annotations about a transliteration line are stored in the line beneath it.

40 Buchholz and Marsi 2006. CoNLL is a column-based file format.

41 Scraper: a type of software that selectively retrieves information from a website when a machine-actionable version of the data is not available. See Veldhuis (2017, <https://github.com/niekveldhuis/Digital-Assyriology/tree/master/Scrape-Oracc> [accessed May 19, 2017]). Tokenization: the process by which occurrences of a word, or other units (such as signs or sentences), are separated as units from the original textual data. Working from transliterations, the major separator to take into account is of course the space character.

42 XML: a form of markup language, like HTML, that encodes annotations such as the codes hidden behind the text we see when we use a word processor, such as LibreOffice or Microsoft Word. See also in this volume, Bigot Juloux, 163–164.

43 Markup: a way to annotate information using tags. HTML is a markup language. For further explanation, see in this volume, Bigot Juloux, 163, and for markup to create taxonomies, Bigot Juloux, 166. For markup in EpiDoc, see in this volume, Eraslan, 289–290).

44 <http://etcsl.orinst.ox.ac.uk/> (accessed May 19, 2017). For further information, see in this volume, Nurmikko-Fuller, 351–352.

45 Identifier (ID): a unique number that does not have a specific meaning. Its quality resides in its uniqueness. See also in this volume, Matskevich and Sharon, 36.

tokens and texts and the relationships among words and tokens. This system provides the same advantages as annotations but without leaving traces in the transliterations. The alpha version of my database and interface are freely available on GitHub.[46]

At the moment, there is no multi-purpose tokenizer available for cuneiform transliterations, but judiciously chosen regular expressions can extract a text's vocabulary, which can then be stored in tables created to this effect (Table 6.1).[47] Only transliteration lines, which always start with a number followed by a period, are to be processed, and structure-marking elements are discarded. Special attention must be given to broken text. Square brackets and hashtags representing breaks are removed from the text as if there were no damage.[48]

TABLE 6.1 *From text to glossary*[a]

Text	**Word tokens**	**Words/lemmata**
&P329002 = CUSAS 13, 134	1(asz@c)	1(asz@c)
#atf: lang sux	sila4	sila4
@tablet	ur-{d}nin-sun2	ur-{d}nin-sun2
@obverse	muhaldim	muhaldim
1. 1(asz@c) sila4	1(asz@c)	udu
2. ur-{d}nin-sun2	udu	nam-ti-e2-mah-ta
3. muhaldim	nam-ti-e2-mah-ta	zi-ga
4. 1(asz@c) udu	udu	
5. nam-ti-e2-mah-ta#	zi-ga	
@reverse		
$ blank space		
1. udu zi-ga		

a "Archival view of P329002," <http://cdli.ucla.edu/P329002> (accessed May 19, 2017).

46 Pagé-Perron 2016a, <https://github.com/epageperron/cuneiform_mining> (accessed May 19, 2017).

47 RegExr (<http://www.regexr.com/> [accessed May 19, 2017]) is an excellent online tool that features a find-and-replace function using regular expressions. For additional information on regular expressions, see in this volume, Eraslan (293), who uses them to work on damaged signs. The Adab corpus is stored in C-ATF format.

48 Note that the Assyriological field does not adhere to the Leiden conventions on representing the condition of the text, although most encoding schemes that are used display similar conventions.

In the above table, column two gathers word tokens found in the text, and the third column assembles only the unique tokens. Repeated tokens are assigned to the same lemma, ensuring the final list is duplicate-free. This information is then inserted into the database; new tokens create new entries in the token table, and re-occurring tokens are counted. This count is added to the join table between the tokens and texts. These tokens can be semi-automatically associated with lemmata. When this process has been applied to all of the texts in the corpus, the result is a word list—that is, a list of occurrences of these words as tokens and a list of associations between texts and tokens. The lemmata themselves should be associated with a part of speech and, in the case of personal names, with the specific named entity type "personal name."

Data Conversion

After the text has been processed and its vocabulary extracted and stored, the precise data to be represented in the network graph has to be filtered out. This is done by querying the MySQL database tables prepared in the preceding section and saving the results in comma separated value (CSV) text files.[49] The management tool phpMyAdmin can help prepare the queries by showing the results to the user before exporting the data to files.[50] Two files are required: one that lists the entities to study and one that lists the relationships among those entities based on the texts in which they appear (Table 6.2). In the case at hand, only personal names are extracted. The list of individuals can be used as is, and each person will be represented as a node in the graph.

Next, the join data must be converted to network graph triples, which generates an edge between all of the pairs of individuals occurring in the same text. This operation must be carried out for each text, with care taken to remove duplicates, or else to have a duplicate of each combination, so as not skew the edge weights.[51] This new list of pairs is the edge list for the network graph. Creating the combinations can be done using online tools, but working with over 50 nodes is made easier by using a combination algorithm to process all

49 CSV: a format that organizes data as a table, with new lines forming rows and commas delimiting columns. See also in this volume, Monroe, 274.

50 <https://www.phpmyadmin.net/> (accessed May 19, 2017).

51 If connections between "a" and "b" (i.e., "a, b") and between "b" and "a" ("b, a") are both present in an edge list, two edges will be created between these entities, or a weight of 2 will be attributed to a single edge between them.

TABLE 6.2 *Excerpt from a MySQL table storing information on the occurrence of tokens in the corpus*

Identification (ID) number	Tablet ID number	Token ID number	Occurrences[a]
1892083	652	3	0
1892085	652	4	0
1892086	652	20	0
1892101	652	21	0
1892092	652	312	0
1892095	652	868	0

a The occurrences field is specifically used to note cases in which a token appears more than once in a text. As such, when the field is set to "1," it means that there is one extra occurrence present. Because in this example the token IDs belong to individuals, the value is set to "0." It is rare that the same individual appears twice in the same text.

the relationships (Table 6.3).[52] At this stage, the data is now expressed as a

52 The Text Mechanic Combination Generator Tool is a good example of such a tool. (<http://textmechanic.com/text-tools/combination-permutation-tools/combination-generator/> [accessed May 19, 2017]). A short PHP script adapted to converting a join table to an edge list, thus generating the required combinations, can be found here: <https://gist.github.com/epageperron/9f631e7898975653b7dc8085867637 87> (accessed May 19, 2017). The code can be read like this:

```
<?php
$join_data = array(); // create an array to hold the join data
$fp = fopen('words_tablets.tab','r'); //read the join data file
exported from MySQL
// while reading the data line by line as tab separated values,
add data as a row in the array precedingly prepared
while (($line = fgetcsv($fp, 0, "\t")) !== FALSE) if ($line)
$join_data[] = $line;
fclose($fp); // close the file after reading
$fp = fopen('edges.csv', 'w'); // create the edges file
fwrite($fp, 'Source, Target, Type, id, Label'.PHP_EOL); //
create the headers for the data
$i=0; // set a counter
$pnos = array(); // set a new array
foreach ($join_data as $jd) { // for each row of the join data
array
```

network graph, and it is ready to be fed into a network graph visualization software.

TABLE 6.3 *Excerpt from an edge list*

Source[a]	Target	Type	Id	Weight
697	1000	Undirected	1	2.0
914	1290	Undirected	2	1.0
1577	1572	Undirected	3	1.0
1065	1062	Undirected	4	1.0
1426	1424	Undirected	5	1.0

a "Source" and "target" are word IDs for personal names. Since the graph is undirected, which ID comes first does not have any bearing on the results.

```
    $pnos[] = $jd[0]; // add the tablet id to the pnos array
}
$uniquePnos = array_unique($pnos); // pick only unique tablets
foreach ($uniquePnos as $upno){ // going through tablet ids one
by one
$to_combine=''; //instantiate a variable
foreach ($join_data as $jd){ // for each line of the join_data
array
if ($upno==$jd[0]) { // if the unique tablet id appears in the
first column of that line
$to_combine[]=$jd[1]; // add the content of column 2 in the
array "to combine"
}}
while ($item = array_shift($to_combine)) { // while combining
all individuals appearing on each text
foreach ($to_combine as $val) {
fwrite($fp,'"'.$item.'","'.$val.'","undirected","'.$i++
.'","'.   $jd[2].'"'.PHP_EOL); // create a row to
represent an edge, the relationships between each
individual
}}}
fclose($fp); // close the file
?>
```

Among the software available on the market, Gephi, with its intuitive interface, is a good entry-level tool for producing a first network graph. Importing data into Gephi is straightforward, and it is easy to style nodes and edges depending on the user's needs. Gephi can also convert graph data to numerous formats; this comes in handy when one is using more than one tool to manipulate the data. Gephi also integrates the sigma.js plugin,[53] which can prepare a website on the fly for displaying an interactive network graph. Finally, Gephi includes a useful feature for isolating ego-networks that provide information about individuals' relationships.[54] Cytoscape is a more powerful software than Gephi.[55] It is used not only by digital humanists but also by biologists and other scientists. Cytoscape offers an extended number of features, especially for statistical data analysis. Complemented by modules such as NetworkX and igraph, the programming language Python can further facilitate exploration of the data in a network graph.[56]

To illustrate the possibilities of exploratory visualization and algorithmic analysis, the next section delineates processes for revealing patterns in the data. The question being asked is: can groups of individuals described in the secondary literature be identified in the dataset using exploratory visualization and algorithmic analysis?

Identifying Individuals of Interest and Group Cores

Anybody visualizing graph data practices exploratory visualization, but it is often overlooked in discussions of research methods. Exploratory visualization exploits our natural ability to recognize patterns visually. Investigating these

53 Sigma.js is available as a Gephi module (<http://sigmajs.org/> [accessed May 19, 2017]; <https://gephi.org/plugins/#/plugin/sigmaexporter> [accessed May 19, 2017]).

54 Ego-network: a network graph of one individual, all of its neighbors, and optionally all of the neighbors of these neighbors. In this context, a neighbor refers to a node that is directly connected with the entity being investigated.

55 Both Gephi and Cytoscape are free, open source, and easy to install.

56 A graphic user interface development environment such as Spyder (<https://pythonhosted.org/spyder/index.html> [accessed May 19, 2017]) can be used. For more information about the programming language Python, visit the official website: <https://www.python.org/> (accessed May 19, 2017). NetworkX documentation can be found here: <http://networkx.readthedocs.io/en/networkx-1.11/> (accessed May 19, 2017), and igraph can be found here: <http://igraph.org/python/#docs> (accessed May 19, 2017). Note that igraph is also available for the programming language R: <http://igraph.org/r/#docs> (accessed May 19, 2017). A module, add-on, or plug-in are all ways to name a software part that can be added to a stand-alone computer program or to a website to increase functionality.

patterns can lead to clues about the organization of the network and help to refine research questions and stimulate new hypotheses. When using a corpus in which the texts, such as sub-archives discussing only one type of activity, are homogenous, network graph visualization is an efficient means for giving an overview of the relationships without requiring extensive manipulation of the graph. For example, a network graph of the fishermen associated with the temple of Bau of Girsu in the Early Dynastic period (c. 2500–2340 BCE) explicitly shows the work arrangements of these individuals. An interactive version of this graph is available online.[57] The graph shows that four major groups of fishermen and their foremen regularly performed jobs together, but the composition of the teams and the foreman in charge could vary slightly. We thus can see four tight groups but with some crossing relationships. The supervisors appear in a centralized position, acting as hubs between the different groups of fishermen.[58]

The more extensive network graph of individuals attested in third-millennium BCE Adab texts comprises over 600 nodes and 5,000 edges (Fig. 6.3).[59] When we look at a full network graph of the Adab corpus, two groups are most visible: textile workers from the Mama-ummi archive and individuals mentioned in the daily bread-and-beer temple rations of the E-tur temple and other institutions.[60] The individuals in these groups share

57 For an interactive version of the graph by Pagé-Perron (2015), see <http://irkalla.net/fishermen> (accessed May 19, 2017).

58 Hub: a node with an important number of connections compared to the average number of connections of nodes in a said network graph.

59 To navigate the Adab network graph, load the sigma.js interactive visualization by visiting <http://irkalla.net/adab> (accessed May 22, 2017) (Pagé-Perron 2017). Search for an individual using the search field on the left, then click on the individual of your choice in the list that appears below. For instance, search for "geme2-" and click on geme2-{d}en-lil2. On the right pane will appear all individuals connected to her. Her ego-network will be shown in the main visualization space. Clicking on a node will bring up that person's ego-network. Manipulation of a graph is facilitated by stand-alone software, such as Gephi and Cytoscape. As such, when the Adab graph is set in its final stage, a link to download the graph data will be available at the same URL. In the meantime, inquiries should be sent by e-mail to Émilie Pagé-Perron—epageperron@gmail.com.

60 The Adab corpus dates to the third millennium BCE and comprises all texts from Adab. The texts include both those from official excavations and those that are lacking provenance but have been attributed to Adab based on factors such as the shape of the tablet, prosopography, and the institutions and rulers mentioned. The large majority of the texts are administrative in nature. The Mama-ummi archive is a group of texts recording transactions related to textile workers. See Maiocchi (2016) for an overview.

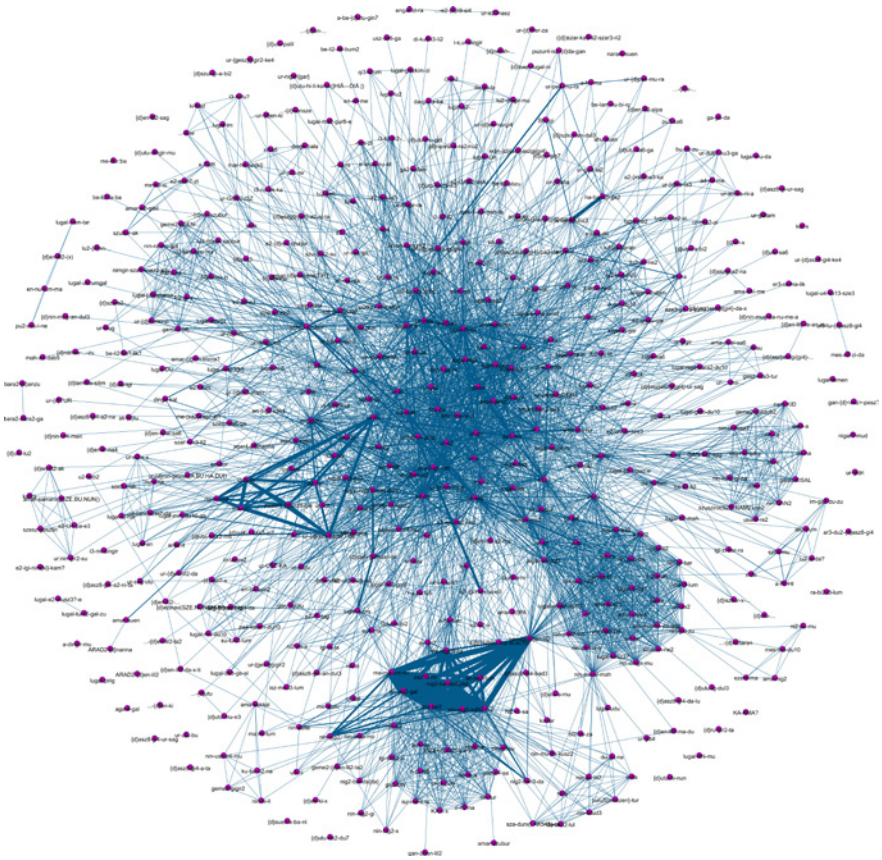

FIGURE 6.3 *The Adab network graph*

edges that have a high weight (i.e., they are thicker) because they co-occur frequently.[61]

In a network with few edge overlaps, most groups can be visually identified. In more complex graphs, extensive overlapping of edges makes it difficult to isolate specific triples. Thus, other means are necessary to indentify groups in the graph.

A possible way to highlight groups in a network graph is to remove edges that have a lower weight. Graphs become simpler and thus clearer when relationships that do not re-occur often are omitted. Nodes that become isolated

61 Weighted relationships can be highlighted in different ways. In Gephi, one can easily make the edges thicker based on the numeric value of their weight. In Cytoscape, a practical way to display the frequency or weight of an edge is either by applying a color range or by changing the thickness of the edge, with the variation depending on the weight value.

can be discarded in the operation. If most individuals in one text do not occur again in others, their contribution to understanding the makeup of groups is minimal. Removing the edges that connect these infrequent individuals reduces the number of edges in the full graph. This highlights the remaining relationships and helps to identify people who frequently appear together, and it also clarifies the role of individuals who act as bridges between separate clusters.

In the Adab corpus, bridges are high-ranking individuals, or they represent homonyms (different individuals who have the same name and who should be represented with two or more separate nodes). A graph-partitioning technique based on the identification of these individuals consists of removing entities that act as a sole or faster point of access between groups in a graph, thus separating these groups from each other. It may be evident from the visualization which individuals serve as bridges, but using a quantitative method to remove them is more efficient, since the treshold used will be numerical and not soley visual. In Cytoscape, it is possible to remove such nodes from a graph after applying the analysis setting and selecting the nodes with the largest edge betweenness,[62] hiding them from the display view. Bridges consist of nodes that usually have a high edge betweeness. Removing bridges fractures the graph and reveals communities by isolating groups of nodes that were connected. Coupled with removing low-weight edges as described above, this method clarifies the picture and highlights important ensembles of individuals. It is particularly efficient for revealing the stable core of archives where individuals co-occur frequently. In the Adab corpus, this equates to specialized workers who less often have hierarchical responsibilities. Individuals with more responsibilities are more likely to be bridges, as they will appear in multiple groups where they manage workers and perform other activities, such as bringing goods to the central authority or receiving rations to redistribute. When the bridges of a graph are removed, the formed groups are tight communities, but they are also missing some of their leaders (the bridges that were removed). It warrants searching for such individuals and their roles to get a full portrait of affairs. With the Adab dataset, groups highlighted with this method equate exactly with the core of the archives that have been discussed in the secondary literature.[63] Examples are: the Mama-ummi archive, the bread-and-

62 Edge betweeness: in network analysis, a mathematical centrality measure of relationships between network data points.

63 See: Zhi 1988; Pomponio et al. 2006; Maiocchi 2009; Maiocchi and Visicato 2012; Visicato and Westenholz 2012; Bartash 2013; Molina 2014; Maiocchi 2015; Pomponio and Visicato 2015.

beer lists to temples (E-tur, E-dam), the middle Sargonic herders, and individuals occurring in Early Dynastic ration lists.[64]

Quantitatively Identifying Meaningful Groups of Individuals

A component is a connected network graph. Any two or more unconnected sections of a single graph are thus different components. These separated groups can be manually identified, but using an algorithm to perform calculations guarantees an exact count. Components can be calculated in Cytoscape or using the "connected_components" function from the NetworkX Python module.[65] This function will traverse the graph and note the unconnected parts of the network, generating lists of entities present in each component. In the Adab corpus, most nodes are connected. This is because the choice was made to conflate entities bearing the same name into one node. These individuals should be subsequently divided into the adequate number of nodes when disambiguated.[66] Since the corpus is diverse and covers about 500 years of history, this high connectedness tells about the important level of homonymy in the corpus. Hence, the identification of groups with the verification of components should be complemented with other methods. In a graph in which the nodes have been fully disambiguated,[67] one would expect to find

64 Examples would be the texts CUSAS 11, 356 and CUSAS 11, 212. (<http://cdli.ucla.edu/search/search_results.php?SearchMode=Text&ObjectID=P412171,P323060> [accessed May 20, 2017]).

65 The appropriate code looks as follows:

```
import networkx as nx #import the NetworkX module to
use as nx
G = nx.read_gml('/path/file.gml') #read the graph data from a
gml format file and put into "G"
print(list(nx.connected_components(G))) #print on
screen a list of the found connected components
```

The Graph Modeling Language (.gml) file can be obtained by exporting the graph from Gephi.

66 An alternative method takes the opposite approach in creating one component for each group of people in one transaction and, while performing disambiguation, merging the multiple nodes representing one individual. For a description of this method, see Broux (2017). Anderson (Bamman, Anderson, and Smith 2013) uses a similar method with the Old Assyrian corpus.

67 Although disambiguation can be facilitated using a statistical approach, in the case of individuals with very few occurrences, there can be few indicators to help with the process.

well-defined, medium-sized components, as opposed to what one finds in the central area of the graph above (Fig. 6.3).[68]

The clique is an important concept in network graph analysis, especially when working with administrative records for which the source data generates undirected links among all of the people appearing in the same text. Each tablet therefore forms a clique, that is, a fully connected subgraph where all nodes are interrelated. Cliques are not exclusive: they can overlap and are formed by any group of nodes that have a link between each of them. For example, in a group where a, b, and c are all related, there are also three cliques of two nodes: (a, b), (a, c), and (b, c).

Maximal cliques are the largest cliques identifiable in a network. If a text mentions "a" and "b," and another text mentions "a," "b," and "c," then "a, b, c" forms a maximal clique, overlapping with the "a-b" clique. As such, maximal cliques equate to the texts with the largest number of individuals who co-occur in other sources.

To find these groups in a graph, one can use the Python NetworkX function "find_cliques".[69] Calculating maximal cliques will generate several groups in a number smaller than the total number of texts but larger than the number of components of the graph. These resulting groups are formed by individuals that occur together at least once, but often more frequently. There are 1,430 maximal cliques in the Adab administrative corpus, about half of the total number of texts in the corpus. Here is an excerpted list of nodes of maximal cliques generated using a Python algorithm for the Adab corpus:

```
['lu2-lil-la', 'ad-da', 'ur-gu', 'za3-mu', 'ur-
{d}{sze3}szer7-da', 'ur-ga2'],
['lu2-lil-la', 'ad-da', 'ur-gu', 'za3-mu', 'ur-
{d}{sze3}szer7-da', 'ur-nu'],
['lu2-lil-la', 'ad-da', 'ur-gu', 'za3-mu', 'di-
{d}Utu', 'ur-e2-igi-nim', 'ur-nu'],
['lu2-lil-la', 'ad-da', 'ur-gu', 'za3-mu',
```

68 Future research in network analysis applied to cuneiform corpora could examine how advanced clustering methods can further investigations of groups of people and provide semi-automated and automated methods for disambiguation.

69 `import networkx as nxG = nx.read_gml('/path/file.gml')`
`print(len(list(nx.find_cliques(G)))) #print on screen a count of the listing of found maximal cliques`
`print(list(nx.find_cliques(G)))`
This will give a count of maximal cliques.

```
    'lugal-ezem', 'lugal-KA', 'ur-ga2'],
    ['lu2-lil-la', 'ad-da', 'ur-gu', 'za3-mu',
    'lugal-ezem', 'lugal-KA', 'ur-e2-igi-nim', 'ur-
    nu'],
    ['lu2-lil-la', 'ad-da', 'ur-gu', 'za3-mu',
    'lugal-ezem', 'lugal-KA', 'ur-…'],
    ['lu2-lil-la', 'szu-na', 'ur-{d}{sze3}szer7-da',
'ur-ga2']
```

[70]

As shown in this excerpt of the maximal cliques from the Adab corpus, one can visually detect the variety of arrangements of individuals that appear intermittently. Meaningful groups can be formed by associating some of the maximal cliques with each other. [71]

TABLE 6.4 *Sample of entity frequency in maximal cliques*

People	Appearance frequency	Times part of a max. cliques
Mesag	24	23
Mama-ummi	47	27
Ur-Ninsun	17	29

There is some degree of variation in the frequency of appearance of an individual in the whole corpus vis-à-vis its appearance in maximal cliques, as exemplified in Table 6.4. An individual appearing more often in maximal cliques than in the whole corpus is one who connects with diverse individuals. This can be explained by two phenomena: either the individual requires disambiguation, or he or she is a higher-ranking person who performs activities in different contexts. For example, "Ur-Ninsun" appears in 17 texts; a group of eight of these texts discusses expenditures of sheep dispensed to Ur-Ninsun the cook. In two other texts, he is attested as merchant and as gardener, co-occurring with different people. Based on this information, one might decide to

70 Each list of people forming a maximal clique is enclosed in square brackets. Some of the texts that include maximal cliques are as follows: Adab 0800 + 1011; CUSAS 11, 050; CUSAS 11, 084; CUSAS 11, 129; CUSAS 13, 008; CUSAS 19, 118; CUSAS 19, 179; TCBI 1, 207. (<http://cdli.ucla.edu/search/search_results.php?SearchMode=Text&ObjectID=P217531,P326098,P325845,P324963,P323569,P323191,P328938,P382459> [accessed June 4, 2017]).

71 See below for an elaboration on this topic in the *k*-plex section.

disambiguate "Ur-Ninsun" into two distinct individuals. A high frequency of occurrence in maximal cliques could also result from a high-ranking individual having influence in different spheres of activity.

A solution compensating for this high-frequency effect in maximal cliques and further clarifying meaningful groups is to relax the maximal clique rules to include more individuals. This relaxation should allow individuals to join the maximal clique members even though they are not fully connected with all other individuals. This procedure forms larger groups and reduces the quantity of groups. Consider these two maximal cliques from the Adab corpus:

```
[lu2-lil-la, ad-da, ur-gu, za3-mu,
ur{d}{sze}szer7-da, ur-ga2] [lu2-lil-la, ad-da,
ur-gu, za3-mu, ur{d}{sze}szer7-da, ur-nu]
```

"Ur-ga" and "Ur-nu" each appear in one list—but among the exact same co-occurring individuals. In a network graph, the individuals in these two maximal cliques are linked to one another, with the noted exceptions of Ur-ga and Ur-nu. Merging these maximal cliques creates what is called a 1-plex, that is, a maximal clique relaxed to accept individuals who are connected to all other individuals minus one connexion.

A *k*-plex is a relaxed maximal clique in which *k* represents the number of connections that can be missing between nodes when the nodes would still form a group, the *k*-plex.[72] In a *k*-plex, most nodes are interconnected, but with a tolerance of *k*-number of absent edges. In other words, an entity must be tied to all but *k* other entities in the group. A 2-plex would be a relaxed maximal clique in which some entities can be connected to all other nodes except two. Calculating *k*-plexes is a complex mathematical problem for which there is no Python implementation at present. Since the number of groups has already been reduced drastically, *k*-plexes can be formed manually. In the case of the Adab corpus, varying the relaxation variable yields results set between the full graph and the maximal cliques, and thus between one and 1,430 groups.

Discussion

How can using network analysis help us better understand the social organization and dynamics of ancient Mesopotamians? Although Assyriologists are generally not acquainted with quantitative analysis methods, these techniques

72 For more information on this mathematical problem, see Balasundaram et al. (2011).

are making a remarkable entry into the field. A compelling example is Paul Delnero's work concerning verbal prefixes whose frequency, used in conjunction with syntactic analysis, supports his assessments of their meaning and usage.[73] This is the first reason why using quantitative analyses is helpful for Assyriological studies: preparing data in a machine-actionable way enables researchers to explore a corpus using a large array of alternative and complementary approaches, from simple statistical models to machine learning algorithms. The results of such methods, including network analysis, yield exact results that can be counted, compared, and used as strong evidence to build an argument. For instance, the cohesion of a group can be said to be stronger in a 1-plex than in a 2-plex, and an individual with a high edge betweenness may either be reputed to be a bridge or require disambiguation.

Looking at larger sets of data, quantitative approaches also offer new means to systematically store information in a manner that retains factors that are not already known to be meaningful. When this data has been collected, network analysis offers an efficient approach to observing entities or individuals and the groups they form in their larger context.[74] One should note that network visualization is also a powerful communication tool. Information remains mostly inaccessible when hidden in tables, but a network graph has an important visual impact that can be used for teaching or to facilitate information dissemination.

Network analysis and visualization are particularly useful for comparative research questions, such as how the hierarchical structures of two different work groups compare and what factors seem to influence the differences. Research on economic, historic, and social questions is enhanced by exploring the relationship among different entities present in the texts but also by connecting entities outside of their restricted groups, between archives. Mama-ummi the textile-worker supervisor is a good example of this: she appears frequently in conjunction with textile workers in the context of the production of textiles and reception of wool. However, since she is a high-ranking manager, she also appears in texts discussing grain allocations.[75] By preparing an ego-network of Mama-ummi, including her direct connections and the connections of her connections, it is possible to situate her within a larger network and investigate the different circles in which she was involved.

---

73 Delnero 2009; 2012.

74 Waerzeggers 2014b.

75 For example: Lippmann Coll 211 (<http://cdli.ucla.edu/P472511> [accessed June 14 2017]) and Lippmann Coll 209 (<http://cdli.ucla.edu/P472509> [accessed June 14 2017]).

Sometimes, disambiguation by disconnecting homonyms is necessary, and it has the noted advantage of detecting individuals who span archives and periods—a task that is much more difficult using other methods. For instance, when the occurrence of the same personal name spans periods or archives, it is necessary to determine whether the personal name represents a single individual. For example, "ama-kesz3"[76] and "a-tu" appear together only once in the Adab corpus, but they also appear multiple times with other individuals who interconnect in texts classified into both the Early Dynastic IIIb and Old Akkadian periods. The analysis of the larger context of their networks, however, reveals that they probably are the same individuals.[77]

Digital, quantitative techniques, as demonstrated in the examples from the Adab corpus case study presented here, can help to detect unseen but meaningful patterns, especially where there are numerous sources to examine. They can also provide quantitative data and a visually compelling display to support or supplement arguments based on traditional analyses.[78]

Increasing the reproducibility of the research is possible by contributing transliterations to a digital library,[79] sharing the code used to produce some results, and/or explaining the steps taken using particular algorithms, coding languages, and software to analyze information. Using or offering open data, along with opting to use open-source software,[80] further supports reproducibility. Moreover, rather than only sharing research results, opening access to the whole methodology employed, as I have done here, increases the possi-

76 Note that the normalized form of this name is "Ama-Keš."

77 CUSAS 11, 052; CUSAS 11, 238; CUSAS 11, 285 (<https://cdli.ucla.edu/search/search_results.php?SearchMode=Text&ObjectID=P322838,P322862,P328956> [accessed March 21, 2017]). Note that the distinction between the Early Dynastic IIIb and Old Akkadian periods is contentious for some texts from Adab, and some authors assign some texts to one or the other period, although the texts are from the same archive. This is due to an overlapping period where Meskigalla was in power. It is thus possible that the same individual appears across periods, but it is also possible that the period assigned to a tablet is simply not refined enough, causing it to appear as though one individual spanned multiple periods.

78 Quantitative data: countable and measurable specimens on which mathematical operations, such as statistical analysis, can be performed.

79 In cyber-research contexts, "library" has two distinct usages: 1) a "digital library" is an online collection of digital objects that can be surrogates of actual objects or parts of objects, or born digital objects (i.e., Cuneiform Digital Library Initiative) 2) a "software library" is a bundle of code that has a specific focus of application and that can be reused by many while developing software (i.e., Python library).

80 For those pursuing digital projects such as this, it is advisable, if possible, to use free and open-source software, such as Gephi, Cytoscape, MySQL (or MariaDB), PHP, and Python, in order to democratize access to digital-research methods and results.

bility for others to validate and reproduce the research steps for their own purposes.[81] Reproducible research saves labor, allowing more analyzes to be conducted, since new investigation do not need to start from the ground up.[82] Information processed during research should be shared as Open Data; this data can then be reused by specialists and non-specialists alike. In his recent article, Marwick explains that enhancing the reproducibility of research also increases the value of a contribution,[83] since reproducibility facilitates replicability—a desirable attribute of studies we call "scientific." Furthermore, it helps to preserve one's research, and, as such, the research will better endure the test of time.

Final Words

From raw data to graph data, manipulating information extracted from cuneiform corpora, algorithmic analysis, and exploratory visualization can be instrumental in discovering new patterns in datasets, including those that have already been studied for a long time. By using network analysis to identify meaningful groups of people, it is possible to perform tasks, such as disambiguation, comparisons of group structure, investigations of meta-interactions between groups, and the exploration of ego-networks. Many compelling hypotheses and interpretations can result from this type of research. For example, of particular interest to this project are questions related to work organization and structure and to the mobility of individuals within this structure, as is shown to be pertinent to the Girsu Early Dynastic fishermen archive and the Adab Mama-Ummi textile workers archive.[84] In both cases, statistical and network analysis support the hypothesis that individuals in positions of power do not simply rest atop a power pyramid: other factors influence their appearance in the texts, and they do not always receive the largest quantity of rations or bring back or produce the most goods. They may also alternate with other central individuals in their respective work groups.

The case study of the Adab corpus presents not only results but also a foundation for further research. A next step in this methodology would be to refine

81 See Marwick's 2017 article, which lays out a comprehensive model for reproducible science that can be consulted for further insight into this topic.

82 The best example of an ancient world project of this sort would be the Perseus Digital Library and its extensions (<http://www.perseus.tufts.edu/hopper/> [accessed May 19, 2017]).

83 Marwick (2017) lays out a comprehensive model for reproducible science that can be consulted for further insight into this topic.

84 Pagé-Perron 2016b; Maiocchi 2016.

the investigation results by giving additional attributes to the nodes based on the metadata collected from the texts,[85] such as sub-period, date,[86] and known archive or find spot. Also, the tools discussed in this paper could be explored further. For instance, the Python modules NetworkX and igraph offer many other algorithms that can help resolve other research questions. More complex programming models could be explored, such as MapReduce programming,[87] which can handle both larger datasets and more insense computations to process graph data. This would be a good avenue to investigate in order to implement an algorithm computing *k*-plexes. Using information based on the verbs present in the texts, furthermore, it would be possible to build a directed network graph and use other, more refined, tools to investigative algorithms geared to the manipulation of directed edges. Because of the way the network is built, it is easy to identify individuals in positions of power and to trace the changes in their influence over time to some degree. Another interesting path to pursue would be to introduce other types of entities, such as institutions, into the graph in order to enrich the network and the interpretations we could draw through its analysis.

By using the techniques discussed in this chapter, we also come closer to the possibility of reproducing one's research. The Assyriological reader is invited to consider preparing machine-actionable text editions that could extend the use of the published information for digital research and quantitative inquiries, such as in the exciting domain of network analysis. Future research in network analysis applied to cuneiform corpora should examine how advanced clustering methods can further this investigation of groups of people and provide semi-automated and automated methods for disambiguation.[88] Clustering methods have proven useful in social network analysis for many decades. A classic example is Wayne Zachary's 1977 analysis of a karate club: he was able to predict where the network would break when the club split apart.[89] Although we do not have evidence of any ancient Mesopotamian karate clubs, dynamic

85 Attributes take the form of extra columns of information in the nodes and edges lists from which data can be filtered, such as by time period or transaction verb.

86 The date should be encoded in a manner that makes it easy to use to form a timeline for all the texts that have such a date; for example, kings can be numbered in order of reign, year placed before the month, month noted as a numeral value, etc.

87 A MapReduce model handles running parallel and distributed computing tasks on a cluster of machines.

88 Clustering methods utilize a statistical approach to group together nodes that resemble each other. Those methods tell us about similarities in the dataset, but the researcher has to interpret the meaning of the groups. Pearce's 2017 Berkeley Prosopography Service might offer solutions along these lines.

89 Zachary 1977.

social networks permeated ancient society, as they do today. Network analysis, therefore, can and should be adapted to the study of Mesopotamian entities, while an array of other digital methodologies can also be employed to tame and interpret larger cuneiform corpora, the scale and complexity of which can obscure meaningful information when treated solely by traditional, qualitative techniques.[90] Especially by ensuring the reproducibility of research by using Open Data and open source software, cuneiform scholars will be able to build upon each other's work and also make it accessible to queries and projects pursued by researchers in adjacent fields who do not read cuneiform. Together we can push the limits of research in our field, while also stimulating new dialogues by opening knowledge of the field to other disciplines.

References

Balasundaram, Balabhaskar, Sergiy Butenko, and Illya V. Hicks. 2011. "Clique relaxations in social network analysis: The maximum k-plex problem." *Operations Research* 59 (1): 133–142. <http://pubsonline.informs.org/doi/abs/10.1287/opre.1100.0851>.

Bamman, David, Adam Anderson, and Noah A. Smith. 2013. "Inferring Social Rank in an Old Assyrian Trade Network." *arXiv.org* cs.CY: 1–6. <https://arxiv.org/abs/1303.2873>.

Bartash, Vitali. 2013. *Miscellaneous Early Dynastic and Sargonic Texts in the Cornell University Collections.* CUSAS 23. Bethesda, MD: CDL Press.

Broux, Yanne. 2017. "Identifying individuals through network visualizations (Perfecting prosopographies)." *Historical Dataninjas*. <http://historicaldataninjas.com/identifying-individuals-network-visualizations>.

Brumfield, Sara. 2013. "Imperial Methods: Using Text Mining and Social Network Analysis to Detect Regional Strategies in the Akkadian Empire." PhD diss. University of California, Los Angeles.

Buchholz, Sabine, and Erwin Marsi. 2006. "CoNLL-X Shared Task on Multilingual Dependency Parsing." In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 149–164. CoNLL-X '06. New York: The Association for Computational Linguistics.

Delnero, Paul A. 2010. "The Sumerian Verbal Prefixes im-ma- and im-mi-." In *Language in the Ancient Near East - Proceedings of the 53e Rencontre Assyriologique Internationale, Moscow-Saint Petersburg, 23-28 July 2007*, edited by Leonid Kogan, Natalja Koslova, Sergej Loesov, and Sergej Tishchenko, 535–561. Babel und Bibel 4 (2). Winona Lake, IN: Eisenbrauns.

90 For instructional information on network analysis and other cyber-procedures, see Weingart 2011, Padilla and Locke 2014, and Posner and Lincoln 2016.

Delnero, Paul A. 2012. "The Sumerian Verbal Prefixes mu-ni- and mi-ni-." In *Altorientalische Studien zu Ehren von Pascal Attinger*, edited by Catherine Mittermayer and Sabine Ecklin, 139–164. OBO 256. Fribourg: Academic Press.

Escobar, Eduardo. 2017. "Cuneiform Technical Recipes as Semantic Network." Paper presented at the AOS annual meeting, Los Angeles.

Koslova, Natalia and Peter Damerow. 2003. "From Cuneiform Archives to Digital Libraries: The Hermitage Museum Joins the Cuneiform Digital Library Initiative." *Proceedings of the 5th Russian Conference on Digital Libraries RCDL 2003*. <http://etana.org/node/6359>.

Maiocchi, Massimo. 2009. *Classical Sargonic Tablets Chiefly from Adab in the Cornell University Collections*. CUSAS 13. Bethesda, MD: CDL Press.

Maiocchi, Massimo. 2016. "Women and Production in Sargonic Adab." In *The Role of Women in Work and Society in the Ancient Near East*, edited by Brigitte Lion and Cécile Michel, 90–111. Studies in Ancient Near Eastern Records 13. Berlin: De Gruyter.

Maiocchi, Massimo, and Giuseppe Visicato. 2012. *Classical Sargonic Tablets Chiefly from Adab in the Cornell University collections*. CUSAS 19. Bethesda, MD: CDL Press.

Marwick, Ben. 2017. "Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation." *JAMT* 24 (2): 424–450.

Molina, Manuel. 2014. *Sargonic Cuneiform Tablets in the Real Academia de la Historia: The Carl L. Lippmann Collection*. Catálogo del Gabinete de Antigüedades 1. Antigüedades Epigrafía 1. Madrid: Real Academia de la Historia, Ministerio de Cultura de la República de Iraq.

Pagé-Perron, Émilie. 2012. "L'industrie de la pêche à Lagaš: Analyse d'un corpus de tablettes cunéiformes provenant des archives du temple de Baba." MA thesis, University of Geneva.

Pagé-Perron, Émilie. 2015. *Network Graph of the Early Dynastic Lagash Fishermen* <http://irkalla.net/fishermen>.

Pagé-Perron, Émilie. 2016a. "Cuneiform_mining." Last modified April 1st, 2018. <https://github.com/epageperron/cuneiform_mining>.

Pagé-Perron, Émilie. 2016b. "The Fishermen of Early Dynastic Lagash." Paper presented at the AOS annual meeting, Boston.

Pagé-Perron, Émilie. 2017. *Preliminary Adab Network Graph*. <http://irkalla.net/adab/>.

Pagé-Perron, Émilie, Maria Sukhareva, Ilya Khait, and Christian Chiarcos. 2017. "Machine translation and automated analysis of Sumerian." *Association for Computational Linguistics Anthology.* <http://aclweb.org/anthology/W/W17/W17-2202.pdf>.

Pearce, Laurie. 2017. "They all have the same name! Using Berkeley Prosopography Services to Tame the Hellenistic Uruk Onomasticon." Paper presented at the AOS annual meeting, Los Angeles.

Pomponio, Francesco, and Giuseppe Visicato. 2015. *Middle Sargonic Tablets Chiefly from Adab in the Cornell University Collections*. CUSAS 20. Bethesda, MD: CDL Press.

Pomponio, Francesco, Giuseppe Visicato, Aage Westenholz, Odoardo Bulgarelli, and Marten Stol. 2006. *Le tavolette cuneiformi di Adab delle collezioni della Banca d'Italia; Tavolette cuneiformi di varia provenienza*. Rome: Banca d'Italia.

Veldhuis, Niek. 2017. "Scrape Oracc repository." <https://github.com/niekveldhuis/Digital-Assyriology/tree/master/Scrape-Oracc>.

Visicato, Giuseppe, and Aage Westenholz. 2010. *Early Dynastic and Early Sargonic Tablets from Adab in the Cornell University Collections*. CUSAS 11. Bethesda, MD: CDL Press.

Waerzeggers, Caroline. 2014a. *Marduk-Rēmanni: Local Networks and Imperial Politics in Achaemenid Babylonia.* Leuven: Peeters.

Waerzeggers, Caroline. 2014b. "Social Network Analysis of Cuneiform Archives: A New Approach." In *Documentary Sources in Ancient Near Eastern and Greco-Roman Economic History: Methodology and Practice,* edited by Heather D. Baker and Michael Jursa, 207–233. Oxford: Oxbow.

Wagner, Allon, Yuval Levavi, Siram Kedar, Kathleen Abraham, Yoram Cohen, and Ran Zadok. 2013. "Quantitative Social Network Analysis (SNA) and the Study of Cuneiform Archives: A Test-case based on the Murašû Archive." *Akkadica* 134: 117–134.

Zachary, Wayne W. 1977. "An Information Flow Model for Conflict and Fission in Small Groups." *JAR* 33 (4): 452–473.

Zhi, Yang. 1989. *Sargonic Inscriptions from Adab*. Changchun: The Institute for the History of Ancient Civilizations.

Tutorials and Learning Material

Padilla, Thomas, and Brandon Locke. 2014. *Introduction to Network Analysis.* <http://thomaspadilla.org/na2014/>.

Posner, Miriam, and Matthew Lincoln. 2016. *Creating Network Graphs with Cytoscape.* <http://doi.org/10.5281/zenodo.56245>.

Weingart, Scott B. 2011. "Demystifying Networks, Parts I & II." *Journal of Digital Humanities*. <http://journalofdigitalhumanities.org/1-1/demystifying-networks-by-scott-weingart/>.