Towards a Better Understanding of *Tarājim*: Creating Topological Networks for Arabic biographical Dictionaries

Hadeel Saadany[1], Emad Mohamed[2], Raheem Sarwar[3]

[1]Centre for Translation Studies, University of Surrey, United Kingdom

[2]School of Management, University of Bradford, United Kingdom

[3]Faculty of Business and Law, Manchester Metropolitan University, United Kingdom

Corresponding author: Raheem Sarwar , R.Sarwar@mmu.ac.uk

Abstract

Biographical writing is one of the earliest and most extensive forms of Arabic literature. Some scholars tend to assume that classical Arabic biographies, widely known as *Tarāǧim*, arose in conjunction with the study of the reliability of the Hadith transmitters (the reciters of the Prophet Mohammad's sayings) which lead to a proliferation of biographical material collected and used to assess the transmitter's trustworthiness Hourani [2013], Perlmann [1964]. However, a scrutiny of the well-known classical Arabic biographical dictionaries such as Siyaru 'A'lāmi an-Nubalā' 'The Lives of the Noble Figures' for Adh-Dhahabī shows that they extend their entries to other classes of persons important to the development of particular fields such as Islamic jurisprudents, rulers, poets, philosophers or physicians. The main contribution of Arabic biographical dictionaries is the cumulative value of the thousands of life histories which construct a picture of the Islamic society in different eras. An Arabic biographical dictionary, therefore, is predominantly used by scholars to look up an eminent person's achievements and historical background. In this project, however, we explore Arabic biographies as a prosopography, rather than a biography in the strict sense. We introduce a novel method for a better understanding of Arabic biographical dictionaries by creating a network of relations among different persons. We utilise Natural Language Processing (NLP) tools to create a topological network from the unstructured data of 45,500 biographical entries collected from different dictionaries. We aim to illustrate how network analysis leveraged by NLP tools can provide scholars with innovative methods for discovering complex constellation of

relations between prominent and non-prominent figures spanning over several eras and from different fields of knowledge. We also use graph visualisation as a means to effectively communicate and explore such complex constellations. Each network visualisation is purposefully designed to be as simple and robust as possible to offer scholars a way to move relatively fluidly between the large scale of biographical entries and to easily interpret the minute ties between persons of different walks of life. We make both our data and code publicly available for researchers to replicate the experiment. It can be found at: https://github.com/sadanyh/Relational-Network-for-Arabic-Tarajem

.

Keywords

Arabic Biographies; Natural Language Processing; Topological Networks

I INTRODUCTION

With more than one million unique entries in Arabic biographical dictionaries, this written tradition constitutes one of the most significant sources of knowledge in human history Bulliet [1970]. However, the existence of this massive biographical material in the Arabic literary tradition should not be confused with the Western notion of a person's biography or bio which usually involves not only facts about a person's education, work and relations but rather a portrayal of a person's experience and lessons on these life events. Arabic biographies, on the other hand, originated to document the life of the Prophet and his companions and, most importantly the life of the transmitters of Hadith (sayings of the Prophet) Young et al. [1990]. This latter type is commonly known as 'Ilmu ar-riğāl 'science of (trustworthy) men' where the biographical material is collected to assess the reliability of the transmitter of a Hadith. Eventually, the genre extended to include other categories of persons important to other fields such as legal scholars, doctors, Sufi masters, Quran reciters and exegetes, philologists, poets, etc. A flourishing of this genre was witnessed in the Abbasid period (661-1258 CE) where biographies were written with the intention of showing how the history of the Muslim community was "essentially that of the unbroken transmission of truth and high Islamic culture" Hourani [2013].

Arabic biographies, despite the diversity of theme, have their own unique structure and purpose.

The basic characteristics of an Arabic biographical entry are a special interest in the descent of each individual as well as an emphasis on the outer events, rather than the mental development of a person's life Young et al. [1990]. Although Arabic has no single term for biography, the most widely used term Tarǧamah (pl. Tarāǧim) refers to a short biographical notice that typically starts with a genealogical record of the individual to validate the biographical authenticity, then the exact date and place of birth and death, if available. The content of the biography itself differs extensively from one biographical book to another. There are, for example, laudatory biographies or hagiographies, widely known as Manāqib (virtues, feats, exploits), which are intended to present a portrait of a morally admirable person, together with a recital of his outstanding actions and achievements (e.g. Siyaru 'A'lāmi an-Nubalā' 'The Lives of the Noble Figures' by Adh-Dhahabī in 1374 A.D). There are also dictionaries devoted to eminent persons resident in a particular city or country which usually contain a topographical and cultural description of that place (e.g. Tāryáu Baġdād , 'History of Baghdad' by Abū Bakr Aámad ibn Alī in 1071 A.D). In modern times, the Arabic biographical genre is still present. For example, we have twentieth century biographies such as 'The Arab Scientists Biography' Awad [1986], 'The Modern Arab Scholars Biography' Al-Alawnaa [2011] and 'The Prominent Innovators' Abdul-Fattah [2010]. However most of the modern Arabic biographical dictionaries are a collection of classical and medieval biographical dictionaries with additional entries for prominent politicians, scientists or literary men and women in modern times. The peculiar structure of classical Arabic biographies is still adopted in entries of modern figures but with less focus on a person's pedigree.

The information contained in Arabic biographies is considered by scholars as "the greatest untapped source of information on the medieval Middle East" Bulliet [1970]. In this study, therefore, we utilise the Arabic biographical material as a vital source of registered data, irrespective of their literary value. We seek to provide the means to identify relationships between individuals across a vast number of famous Arabic biographical dictionaries that might otherwise require extensive time and effort by researchers to manually determine. We compile a dataset of biographical entries from famous classical biographies such as سِيَّر أَعْلام النُّبَلاءِ (Al-Dhahabī in 1374 A.D) and أَسَدُ الغابَةِ فِي مَعْرِفَةِ الصَّحابَةِ (Ibnul-atheer 1305 A.D) by scraping an online

Encyclopedia of Arabic biographies. We illustrate the diversity of insight and utility of analysis that researchers can achieve by applying a variety of NLP tools to unearth relevant historical relations in a bafflingly mass material of biographical entries. Moreover, we demonstrate how graphically visualising these cross-connections in as simple and robust form as possible can help scholars interpret ties in terms of both the geographical and temporal attributes of the data. The contributions of our research, therefore, can be summarised as follows:

1. *create* a large dataset of 45,500 biographical entries compiled from different biographical dictionaries spanning from the pre-Islamic to the modern era.

2. *use* NLP tools to structure the collected dataset by extracting attributes needed for building relational networks between persons in the biographical entries.

3. *propose* to use the semantic similarity metrics to create topological networks between an eminent individual and his contemporaries as well as the clustering of a group of individuals around a thematic node.

4. *provide* visualisation methods for graphically interpreting constructed biographical networks from a spatio-temporal perspective.

In order to present our networks, in Section II we review previous related research on network analysis for Digital Humanities (DH) purposes and how information in Arabic biographical dictionaries has been tackled by other scholars. Then, in Section III we describe the methodology followed for constructing the topological networks. In this section, we explain how the biographical data is compiled, the challenges we faced to extract the network attributes and the NLP tools we utilised to extract information from the data for the purpose of network analysis. In Section IV we illustrate our experimental results for the construction and visualisation of representative cross-sections of the constructed networks with the objectives mentioned in 3 and 4 above. Finally, in Section V we discuss our conclusions on the overall experiment and suggestions for future work.

II   RELATED WORKS

In this section we review the existing related works and organise them in the following two sections. In Section 2.1 we review how network analysis tools are used in DH research. In

Section 2.2 we review studies conducted on Arabic biographical dictionaries.

2.1 Network Analysis for Digital Humanities

Network analysis has recently been one of the computational techniques used in DH to interpret large-scale data particularly in the field of historical qualitative analysis Conroy [2021], Claveau and Herfeld [2018], Kienle [2017]. It is used as a complement to manual analysis of documents since networks can identify relational patterns across a vast number of documents. This approach has proven quite adequate in interpreting historical conclusions drawn from a robust network analysis of large data. As Wilken wilkens2015digital points out in his research on the use of computational methods in DH, tools such as network analysis provide the means for the quantification of great masses of resource material that have long been analysed almost exclusively in qualitative terms by humanist researchers. Painter et al., networkanalysis provide examples of how a network approach of the history of science can enrich the understanding of co-authorship relations in a dataset of evolutionary medicine publications. They construct a dataset of 6,456 publications that appeared from 1971 through 2017 in the International Society for Evolution, Medicine, and Public Health global directory. They use the metadata for each publication (e.g. author names, institutional affiliations, and date) to explore the various relations among authors. Although they use automated tools such as network-level statistics to analyse the network relations, they tackle textual challenges in their dataset manually. For instance, they resort to human effort to disambiguate variant names of the same author. To mitigate the human effort, our approach for constructing biographical networks in this research employs NLP tools to extract necessary elements for the relational analysis.

Similar to Painter et al., networkanalysis relational approach to the history of science, So and Long so use networks to understand the influence of global modernism on literary works by analysing a corpus of literary journals from the United States, China and Japan written early in the twentieth century. They provide new insights into interpreting the relationship between the influence of modernism and the poetic form by using the metadata of the publication records of their corpus, mainly, by counting the number of poems each author published, the year of publication, and the name of the journal. From their network analysis, they observe important

differences among the three national contexts: the US, China and Japan. Their network clusters show that there is much more diversity in the literary work in the US as opposed to centralised literary production in Japan and China where the literary work clusters are limited to a few number of poles in their network.

Scholars interested in network analysis of historical data have also considered the structure of networks as ontologies for distilling complicated connections between historical people, objects and places. Langmead et al., ontologies examine the methodological considerations behind designing interoperable ontologies for historical data through relational networks. They advocate for adapting an ontological standard and structure in accordance with the purpose of the humanistic enquiry rather than adopting a uniform network structure. Ladd et al., programming also show that it is possible to draw important conclusions by using programming tools to explore network statistics and metrics. They illustrate how network statistical analysis can help identify the most significant individuals in clusters of the social network of a mid-seventeenth century Protestant Christian society.

Network analysis in humanities has also been leveraged to construct software tools to help researchers interpret historical biographical data. Chen and Chang chinese provide a relationship map tool (CSNRMT) for exploring characters' social network relationships in Chinese ancient books. They provide researchers with an application programming interface (API) for interpreting historical texts through relational networks constructed from archived databases such as China Biographical Database. By conducting surveys on 21 users of this API, they show that their platform has significantly helped humanists in interpreting ancient texts as well as analyse characters' social network relationships in the ancient Chinese heritage.

The visualisation of relational networks in DH research has also markedly contributed in the analysis and interpretation of important links in datasets of significant size. Tamper et al., visualise show that using knowledge graphs for constructed networks can help examine individual relations in biographical and prosopographical research. They create a network of 13,100 biographies from the collections of the Biographical Centre of the Finish Literature Society. They illustrate how visualisation of relational networks can help identify and extract relations be-

tween entities in a large biographical dataset. On a smaller-scale data, During Hermeneutics illustrates how the visualisation of social networks for 1,400 people between the years 1942 and 1945 helped him to discover highly important contact brokers of Jewish refugees who significantly helped survivors of the Holocaust.

As for the Arabic and the Islamic tradition, there has been a recent interest in applying network analysis for different purposes. For example, Boella et al. boella2011salah use unsupervised segmentation and linguistic analysis of Arabic texts of Prophetic tradition حَديث to automatically segment each text unit in a transmitter chain and a text content. They analyse each segment according to a set of regular expressions chunks in transmitter chains in a graph labelled with the relation between transmitters as well as a morphological analyser to annotate lexically and morphologically the text content. A graph with relations among transmitters and a lemmatised text corpus is the final output of their analysis. In the next section, we briefly describe how the Arabic biographical entries have been studied.

2.2 Study of Arabic Biographical Dictionaries

The study of Arabic biographical dictionaries in humanities have focused primarily on two themes: the development of the genre and the classes of persons covered on the one hand Wilkens [2015], Young et al. [1990], Cooperson [2000], and the Arabic biographical dictionaries as a valuable source of historical material on the other Caine [2018], Nazmabadi et al. [2014]. Young et al. [1990] provide an extensive review of the distinguishing characteristics of classical Arabic biographical dictionaries and how they diverge from the Roman and Greek biographical tradition. They illustrate the arrangement system for Arabic biographical notices and the development of their theme since the emergence of Islam and during the different Islamic Caliphates. He points out the fact that the Arabic biographical literature exceeds that of any other culture in the ancient and medieval periods. Similarly, Cooperson classical studies the origin and development of Arabic biographical dictionaries without addressing any textual or relational analysis of their content. He focuses primarily on the terminological ambiguities related to the categorisation of Arabic biographies as well as the various purposes for which each was written. On the other hand, Al-Qadi arabicbios views Arabic biographical dictionar-

ies as the scholars alternative history of the Muslim community across different eras. Following the same suite, Bray arabicbio2 studies the biographies as a historical resource but zooms in on only the Medieval and Modern Arabic periods. There were also scholars who used this genre to study the history of a particular region in the Islamic world. For example, Sharkey arabicbio3 used biographies to study the history of Sudan. There were also researchers who used the data in the Arabic biographical dictionaries to study the history of exemplary lives of a particular group of individuals: for example, Booth arabicbio5 studies the feminist roles across different eras, Yusoff arabicbio6 studies the lives of the Hadith transmitters and Osti arabicbio7 studies the portrait of scholars as given in the Medieval biographies.

There were some attempts to analyse relations in Arabic biographies as a means for a better understanding of their content. Romanov algorithmic focuses on one bio-bibliographical collection written in 1919 by Ismāīl Bāsha Al-Baghdādī which covers the period before the beginning of Islam in the seventh century AD up to the end of the nineteenth century. He uses manual and semiautomatic tagging of important features in the textual data to conduct an algorithmic analysis of the biographical entries. Mainly, the semi-automatic tools used are regular expressions that tag descriptive names, place names, and dates in each biographical entry. He uses this tagged data to discover spatial and chronological patterns among different persons included in this dictionary. In our research, however, we tackle a collection of biographical dictionaries with a wider range of themes rather than one in particular.

III METHODOLOGY

3.1 Data Compilation and Challenges Involved

There is a huge number of Arabic biographical dictionaries which differ not only in terms of the classes of persons they document but also in their length and structure. Also, the time-span covered by these dictionaries expand from quite an early age in Islamic history as researchers point out that the Prophet Muhammad has inspired the first Arabic biographies, known as Al-Sirah Al-Nabawiyyah (he Prophetic biography); the earliest was by Wahb ibn Munabbih (654-725 AD). Moreover, scholars have noted that Arabic biographical material on poets, singers,

Quran readers and jurisprudents are as old as the ones on Hadith transmitters Cooperson [2000]. Faced with this abundant biographical material, we searched for a source that has a collection of Arabic biographical dictionaries rather than analysing individual ones dedicated to one field of knowledge.

As far as the authors' knowledge, the largest compilation of Arabic biographical dictionaries is provided by *"Mawsūat al-Tarājim wa-l-Alām"*, 'The Encyclopedia of Biographies and Imminent Persons',Tarajim [2022]. This online Encyclopedia is a collection of 30 biographical dictionaries with around one million biographical entries. The collection includes some of the most well-known classical Arabic biographical dictionaries which record the lives of notable persons belonging to different categories such as companions of the Prophet (e.g. طَبَقاتُ الفُقَهاء by Ibnul-atheer 1305 A.D), Islamic exegetes, (e.g. أَسَدُ الغَابَةِ في معرِفَةِ الصَحابة e.g, by Al-Shirazi 1014 A.D) and literary figures (e.g. مُعجَمُ الأُدَباء by Yaqut al-Hamawi 1184 A.D). We refer the reader to the online site of the Encyclopedia[1] for the detailed list of the 30 biographies included in their database.

The biography of each person in the Encyclopedia comprises a collection of all the biographical entries on that person mentioned in the biographical books included in the its data. Thus, for example, the entry for the famous Islamic warrior Khaled ibn al-Waleed consists of entries extracted from 9 different biographies that describe his line of ancestors as well as major events in his life. The designers of the online Encyclopedia state that the time-span of the biographical dictionaries extend from the pre-Islamic to the modern era Tarajim [2022]. Moreover, the entry search in the Encyclopedia is gender-based. Thus, biographical entries are indexed as belonging to names Men, Women, or an epithet by which they are known.

Since our aim is to provide an illustration of how relational networks can unearth significant connections between individuals in different biographical dictionaries, we opted for scraping a sample dataset of 45,500 entries extracted from the biographical dictionaries on the Encyclopedia website. We used the Beautiful Soup Beautiful Soup [2022] Python library to pull out the data from the encyclopedia website in a computer-friendly format in order to be able to utilise

---

[1]https://www.taraajem.com/books

NLP tools for building up relational networks. We did not select the entries from one particular biography but we selected biographies for both men and women gender.

In our analysis, however, we targeted specific categories in our dataset that we thought would bring about informative relations in our network. Thus, we extracted what is known in Arabic biographical genres as مَجاهيل, "the Unknown", which refers to individuals that are not eminent in any particular field but are only notable because of their relation to a well-known person. Those مَجاهيل biographical notices of unknown men or women were traditionally collected by classical biographies to document the authenticity of a Hadith by the Prophet or an interesting anecdote recorded about an outstanding person in which the unknown man or woman was involved. We used this category to reveal any unexpected cross-connections between networks of the other groups, mainly prominent men and women extracted from the other biographical entries included in the Encyclopedia website.

The second category that was of interest is women biographical notices. The Encyclopedia of dictionaries included separate entries for women notable in different fields which were collected from classical and modern Arabic biographical dictionaries such as الطَّبَقاتُ الكُبري by al-Zuhari (785 A.D) and الأعلام by al-Zarakli (1927 A.D). We focused on this category to bring about any cultural, religious or historic commonalities between female prominent figures in classical and modern times.

After collecting the data, however, we were faced with two challenges. First, since the biographical entries were collected from different dictionaries, they lacked structural consistency. Thus, the length of notices varied widely. Some were very brief (e.g. الأعلام dictionary would have the following notice of a woman : "Amina Bint Anaan: a noble woman from Baghdad, she recited hadith in Baghdad and Mosul, lived and died in Mecca"), others were of considerable length (up to several pages) Al-Zarakli [1927]. The length of a person's entry was broadly proportional to his or her importance in the respective field. Despite the length variance, the encyclopedia consistently included shorter introductory notices of longer entries which summed up the essentials of the persons character in a few words. To avoid bias towards lengthier entries by NLP tools used for relational analysis, we opted for extracting the shorter notices of longer

biographical entries to have relatively consistent length for biographical notices in our dataset.

The second challenge is related to extracting information essential for creating networks by NLP tools. To build relational networks, we needed to hard-code attributes that are common among different persons, for example, common era, geographical proximity, common places of birth, death or living, common historical events or a combination of any of these attributes. Although Arabic biographical dictionaries typically start with genealogical and geo-temporal information about a person, the style of presenting this information differs from one dictionary to another. For example, one dictionary would incorporate birth and death information thus: "Ibn Al-Naqīb: Egyptian Shafii (school of jurisprudence) his birth and death in Egypt..." Al-Zarakli [1927]. While in others, the geo-temporal information is not as clearly presented. For example, the place of birth is indirectly expressed in the biographical notice of Ibn Al-Dabbaāgh: "A linguist *from the people of Baghdad.*." Adh-Dhahabi [2006]. In more intricate examples, the historical period is to be deduced from context as is the case in the following biographical notice:

> **Ibn Mājid** : *The Lion of the Sea, one of the prominent Arab navigators in the Red Sea, the Mediterranean, the Persian Gulf, the Indian Ocean, and the Sea of China. He is the Sea Captain who guided Vasco da Gama, the head of the Portuguese fleet, to the route in his trip from Malinda on the Eastern coast of Africa to Calcutta in India....*Al-Zarakli [1927]

As seen from the quote above, information about the particular historic period can only be inferred from the fact that Ibn Mājid and Vasco da Gama were contemporaries, i.e. the 15th century AD. These inconsistencies in the textual structure of the scraped biographical data also apply to other information such as the birth/death place and profession. It is also not uncommon to have the biographical notice stating that the exact date of birth or death is not available or cannot be determined.

It is considerably difficult to extract network data from such unstructured text since the network nodes and edges are primarily determined by having clearly defined features for each individual, such as place of birth, death, position, important dates or events, and historical era. Extracting commonalities among individuals in terms of these attributes is essential for the purpose of

network analysis and structure. Manual extraction of these attributes from our large dataset would be time-consuming and exhausting since it requires full concentration at all parts of the biographical entry. Here NLP tools come in value. We use NLP methods such as semantic matching, document clustering, and regular expressions to extract the attributes needed for an easier access of the biographical information in the data through relational networks. The NLP tools used for this task are explained in the following section.

3.2 Data Structuring by NLP Tools

Creating a successful network from an unstructured data requires clearly defined ties between nodes in the network, in our case individuals in the biographical entries. The ties that we were seeking are mainly the attributes that are typical of an Arabic biographical dictionary. These attributes are: date of birth, date of death, place of birth, place of birth, era, and the specialised field of knowledge in which the person is prominent. We also added gender as one of the distinguishing attributes. As mentioned in the previous section, information on the attributes are not textually expressed in a consistent style. So we utilised a number of NLP methods to extract these seven attributes.

We extracted places of birth and death in two steps. First, we used regular expressions to extract all trigrams (three-word phrases) that can precede or follow the place of birth or death. After examining the data, we extracted phrases that are normally used in Modern Standard Arabic (e.g. 'his birth place was Mecca', 'he died in Aleppo', 'his birth and death was in Damascus' etc.) to match the Arabic version used in modern biographies in the dataset. To capture phrases used in the old biographies, we extracted Classical Arabic phrases used to denote the place of birth (e.g. 'he is from the people of Baghdad', 'Cordoba is the place where his head has fallen (i.e. where he was born)'). The output of this stage was truncated phrases that have the word indicating the location of birth or death somewhere in the middle. Second, in order to have the place as an attribute, we needed to extract only the name of the city or the country from these phrases. We experimented with two methods. The first is what is known in NLP as Named Entity Recognition (NER) which is the method of automatically classifying named entities such as person names, organisations and locations from large unstructured text. We

used CAMel NER tool[2], a Python toolkit for Arabic NLP, to extract all the locations of birth and death from the phrases produced in step one Obeid et al. [2020]. The NER CAMel model is trained on modern names of cities in the world, so it can extract cities such as Baghdad, Aleppo, Mecca and so on. It is not trained, however, to extract cities that appear in classical biographical dictionaries from the 12th or 13th centuries such as Constantinople (present day Istanbul), Khorāsān or Bukhāra (present day Uzbekistan), or cities in Spain under the Islamic ruling such as Sarqasta (present day Zaragoza) and Mayorqa (present day Majorca). For these older city names, we used a programming look-up dictionary based on a 14th century Arabic biography for city names, The Observations of Names of Cities and Places by áafiyy Al-Dīn Safi Aldin [2008]. Finally, with the help of programming string functions, we managed to extract only the name of the city or country of birth and death whenever available as well as their geographical co-ordinates as separate attributes for each biographical entry.

The second category of attributes was the dates of birth and death. The scraped data from the Encyclopedia website contained dates in a special dual format typical of modern Arabic Biographical dictionaries: the Hijri (i.e. Islamic lunar calendar) denoted by the Arabic letter 'ه' (e.g. 1268 هـ) followed by the Gregorian calendar denoted by the Arabic letter 'م' (e.g. 1547م). We used both regular expressions and programming string functions to separate dates of birth into the two independent categories: AH (Hijri) and AD (Gregorian). As for the era, it was easily separated as the scraped data had a separate category for the era of each biographical entry. Our dataset comprised biographical entries from 10 eras: Pre-Islamic (Before 610 A.D), Islamic (610-632 A.D), (Rashidūn Caliphate (during the reign of the four major caliphates after the death of the Prophet Muhammad(632-662 A.D)), Umayyad Caliphate (661-750 A.D), Abbasid Caliphate (750–1258 A.D), Fatimid Caliphate (909-1169 A.D), Ayyubid dynasty(1174-1260 A.D.), Mamlūk Sultanate (1250-1517 A.D), Ottoman Empire (1299–1922 A.D) and the Modern period (1800s and 1900s AD).

The last and most important of the attributes which we aimed to extract was the position of the individual, i.e. the field in which he or she is distinguished or famous for. Again, extracting this information from this large unstructured data constituted a particular challenge due to the

---

[2]https://camel-tools.readthedocs.io/en/latest/api/ner.html

Clustered Category	English Translation
مُفَسِّر	Islamic Exegete
تَفسِير	Islamic Exegesis
فُقَهاء	Jurisprudents
عَالِم	Scholar
مُحَدِّث	Hadith Narrator

Table 1: Examples of Elements of One Cluster in a Biographical Dictionary

huge variety of fields documented on the one hand, and the different stylistic features of each biographical dictionary on the other. Due to this, we opted for extracting the most frequent positions in the dataset by the NLP methods for document clustering. First, we transformed the corpus of the biographical dictionaries into vector space using term frequency-inverse document frequency (tf-idf) for each biographical summary. Tf-idf is a statistical measure which refers to the frequency of a word in relation not only to the document it occurs but to other documents in the corpus as well. Words with a high tf-idf score are assumed to contain more meaning in relation to the document and hence, in our case, may be representative of the position of the individual. Then, we used the tf-idf matrix of each document to run a k-means clustering algorithm which assigns groups of documents with similar tf-idf values to distinct clusters. Table **??** shows an example of the output of the clustering algorithm for the documents in a biographical dictionary. This cluster is grouped under the word 'القرآن'(al-Qurān) and the words with the highest tf-idf scores included positions such as 'Islamic Exegete (i.e. an expert in the critical explanation of the Qurān), Jurisprudence, áadīth narrator and so on. We manged to compile a list of the most common positions and used it to couple each position keyword with its respective entry.

After extracting the seven attributes from the biographical dictionaries, the final format for each entry was structured as shown in Figure 1. The rows represent the entries and the columns represent the attributes of each individual. The columns in the figure show that each 'name' is connected with its biographical notice (summary), date of birth and death in AD and AH, (DOB_AD, DOB_AH, DOD_AD, and DOD_AH, respectively), its position, and the place of birth and death. We utilised different combinations of these attributes to create topological

age	gender	name	summary	DOB_AD	DOB_AH	DOD_AD	DOD_AH	Position	place_birth	death_place
المملوكي	نساء Women	بنت الشيرجي	بنت الشيرجي مدللة بنت أبى بكر محمد بن إلياس ا...	not available	not available	1272	670	فاضلة Noble	دمشق Damascus	دمشق
الحديث	نساء	مريانا مراش	مريانا مراش مريانا بنت فتح الله بن نصرالله بن...	1848	1264	1919	1337	شاعر Poet	حلب Aleppo	حلب
المملوكي	نساء	مريم بنت أحمد	مريم بنت أحمد مريم بنت أحمد بن أحمد ابن القاضي...	1319	719	1402	805	قاض Judge	القاهرة Cairo	القاهرة
الحديث	نساء	مريم نحاس	مريم نحاس مريم بنت جبرائيل نصر الله نحاس: ادي...	1856	1272	1888	1305	not available	بيروت Beirut	مصر
المملوكي	نساء	مريم الحرة	مريم الحرة مريم بنت شمس الدين بن العفيف: زوجة...	not available	not available	1313	713	زوجة السلطان Wife of a sultan	not available	جبلة

Figure 1: Example of Attributes Extracted for Each Individual

networks between persons in our biographical dataset which can aid scholars in research as well as provide insight into relational ties that may go unnoticed in large unstructured biographical data. An illustration of how the relational networks can help in analysing Arabic biographies is explained in the following section.

IV   EXPERIMENTAL RESULTS

4.1   Relational Network for Eminent Figures

We experimented with different semantic similarity measures in order to create a topology of networks between the persons included in our large dataset. The first was an exploration of the ties between a node that has a prominent figure in one field and his contemporaries in surrounding nodes. We chose Ahmad Ibn áanbal (780–855 CE/164–241 AH), a major collector and critic of Hadith, a traditionalist [3], and founder of the áanbali school of Sunni jurisprudence Melchert [2012]. To measure document similarity, the words in the biographical text need to be transformed into numerical vectors and the distance between any two given document vectors defines their degree of relatedness. There are several algorithms in NLP research for transforming words into numerical values representative of their semantic value. The first method we used was Sentence-BERT (SBERT), which is a deep learning algorithm that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings Reimers and Gurevych [2019]. Sentence embeddings are a type of numerical representation that allows sen-

---

[3]An Islamic Traditionalist rejects taking religion from rationalistic Islamic theology in favour of strict textualism in interpreting the Qurān and áadīth.
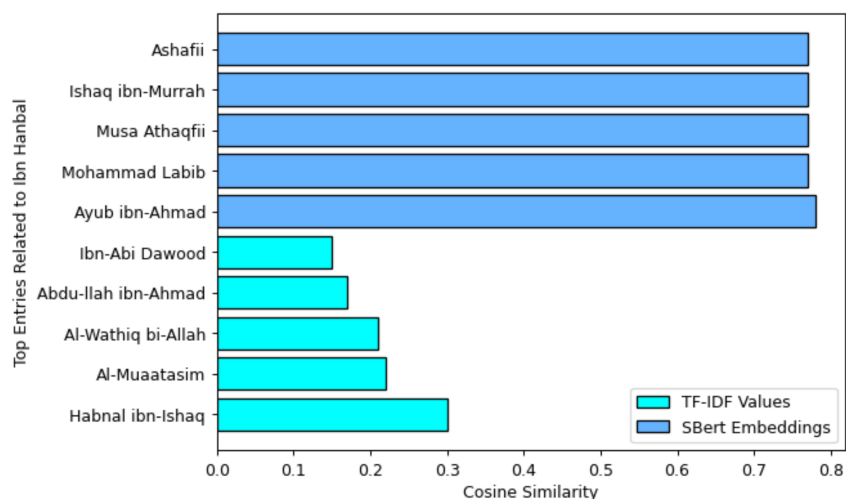
Figure 2: Top Cosine Similarity Related Entries to Ibn áanbal

tences with similar meaning to have a similar representation in vector space. The second type of numerical transformation we used was the tf-idf frequency values explained in Section 3.2. We experimented with the SBERT embedding values as well as the tf-idf values of Ibn áanbal's biographical entry and entries of other individuals in our dataset.

To calculate the semantic distance between Ibn áanbal's document and other entries, we applied the cosine similarity model. Cosine similarity is a commonly used metric, which measures similarity as the angle between two vectors; in our case vector representations of biographical entries Xia et al. [2015]. Figure 2 shows the cosine similarity values of the top five most related individuals in our dataset to Ibn áanbal based on the SBERT embeddings and TF-IDF values, in blue and cayen respectively. Despite the fact that SBERT embeddings are considered state-of-the-art semantic representations in NLP research, we found that the tf-idf values provided more informative similarity measures than the SBERT embeddings which produced indistinguishable similarity values with a large number of entries with Ibn áanbal's. We, therefore, opted for using the cosine distance between the vector of the tf-idf values of Ibn áanbal's biographical summary and the tf-idf vectors of other entries in our dataset to create the relational network. For the visualisation of this network, we used NetworkX[4] to represent these scores in a network where Ibn áanbal is a node and the similar documents are connected to it by edges that varies according to the similarity weight. Figure 3 is a graphical representation of a section from Ibn

---

[4]NetworkX: a Python package for complex network creation.

áanbal's network[5] based on cosine similarity scores. It should be noted here that the graph shows only a cross section of Ibn áanbal's network, the network can be expanded as far as the research purpose entails. We aim here to show how network analysis and visualisation of Arabic biographical dictionaries can highlight relations between individuals in an insightful and elegant way that are not feasible by manual research.
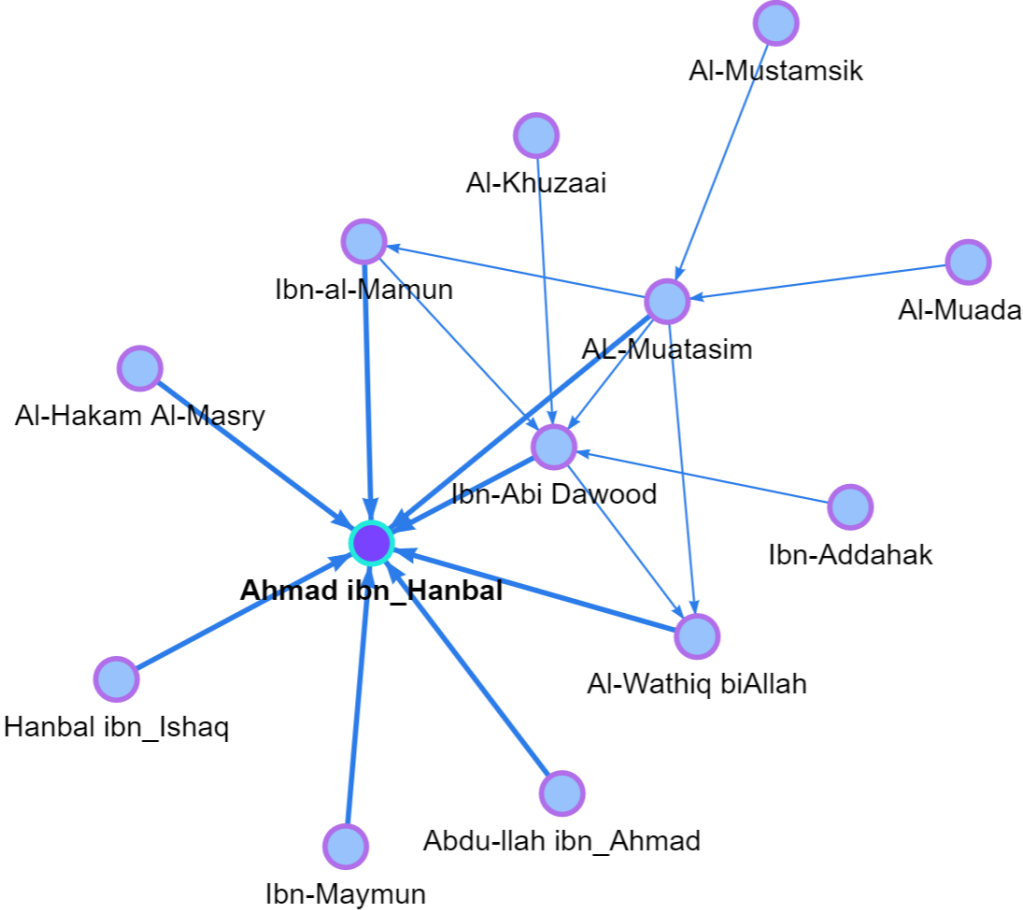

Figure 3: A Section of a Graphical Relational Network for Ibn Hanbal

In Figure 3, the thickness and length of the edges are directly proportional to the similarity scores between Ibn áanbal and the individuals in surrounding nodes. The arrows indicate the direction of relation and their thickness reflects the strength of relation. This part of the relational network between Ibn áanbal and other personalities in our dataset unwraps different types of relations. First, the network shows that Ibn áanbal is not only connected to those in his field of knowledge such as his cousin and protege áanbal Ibn-Isááq and his son and follower Abdullāh, but also to political figures such as al-Muatasim (833–842 AD), the eighth Abbasid

---

[5]The names in the network graphs are translated into English for an easier visualisation.

caliph. Network diagrams become meaningful when they are part of a dialogue with data and other sources of information Düring [2015]. According to context knowledge, Ibn áanbal is closely connected to al-Mutaáim due to his refusal to accept the Mutazila view of the Quran which asserts the religious authority of the caliph. As a consequence, al-Mutaáim imprisoned and tortured Ibn Hanbal; this constitutes a pivotal incident in Ibn áanbal's life captured by textual similarity scores. Moreover, as shown in the figure, both Ibn áanbal and al-Mutaáim are connected to Al-Wāthiq Billāh (842-847 AD), who was the son of al-Muatasim and his successor. He is known for his tolerance with Ibn áanbal as he pardoned him and the different biographies of Ibn áanbal indicate that it is during Al-Wāthiq's reign where Ibn áanbal lived in peace. Another interesting bond in the network is between Ibn áanbal and one of the unknown Islamic Exegete of his age, i.e. Al-áakam Al-Miáriyy. By checking the biographical notice of Al-Miáriyy, we found that the established connection with Ibn áanbal is due to the fact that they both rejected the Mutazila view, but due to little biographical information about him, he did not receive as much fame as Ibn áanbal concerning this incident. Thus, as shown in Figure 3, only some of the top cosine similarity scores between Ibn áanbal and his contemporaries reveal typical and non-typical connections between famous and non-famous individuals. A researcher can go as deep or as shallow with creating such a topological network according to the objective of research.

4.2   Relation Network to a Thematic Vector

The second type of networks we created was between women biographical entries and a vector constituting of tf-idf values of battle-related synonyms (e.g. 'fight', 'Jihād', 'Ghazwa' (a battle in which the Prophet was involved), 'war', etc.) which we called 'Women Fighters'. The objective of this network is to locate women across different ages who participated in the battle field or in combat-related activities. We also created a parallel network based on the cosine similarity of women in the 'Women Fighters' network and all other women biographical entries in our dataset. Figure 4 shows a section of this multi-graphical visualisation of these two networks.

The numbers on the edges in Figure 4 show the cosine similarity scores between different women in relation to the 'Women Fighters' vector as well as to other women in the network.
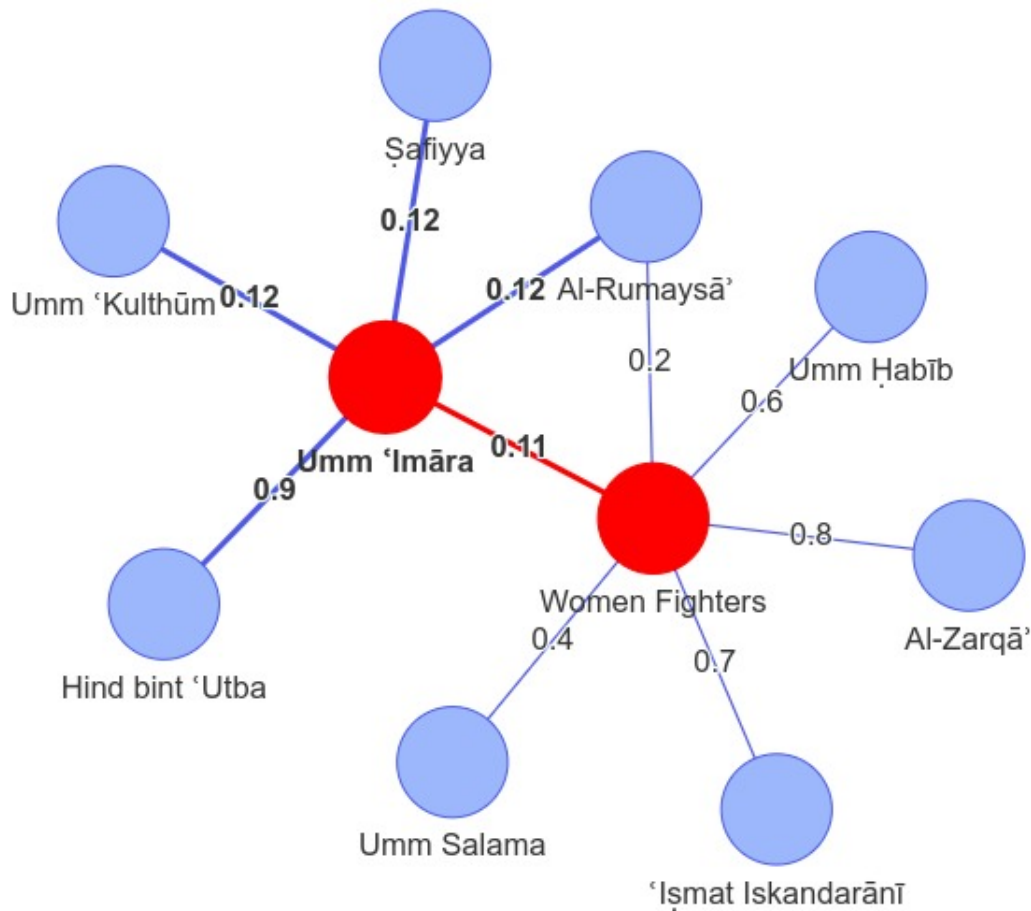
Figure 4: A Section of a Multigraph representation of Women Entries

The most interesting part of this network is that it was able to correlate women fighters across different ages. For example, it shows that several prominent women figures from the Islamic era such as Umm imāra and áafiyya (the Prophet's aunt) are connected to women in later ages such as 'issmat Al-Iskandarāni', an author whose father died in the Crimean War (1853-1856 AD) between Turkey and Russia during the Ottoman Empire era. The latter documented a number of battles in her writings. We also attempted to calculate the centrality measures of these two connected networks. Through centrality measures, we learn how to find the most important nodes (individuals) in the network. The first centrality measure we calculated was the 'degree of centrality' which is based on the assumption that important nodes have the largest number of connections in a network Opsahl et al. [2010]. We found that the highest degree of centrality (0.5) for women figures in this part of the multi-graph was to 'Umm imāra'(634 A.D). She was a woman warrior who fought several battles during the Islamic and Rāshidūn Caliphate era Al-

Zarakli [1927]. As can be seen from Figure 4, 'Umm imāra' is a central node connecting women figures to the 'Woment Fighters' network. For example, 'Umm imāra' has a cosine similarity score of 0.12 to áafiyya. Although the latter is not a woman warrior, her biographical entry mentions an incident where she had to kill a spy who attacked the women's camp during one of the Muslims battles against Quraysh. Thus, the centrality of this node highlights the thematic connections between women biographical notices. Moreover, network centrality measures can highlight not only how many individuals one is connected too, but the importance of a node based on the type of people it is connected to it. Eigenvector centrality measure provides exactly this Ruhnau [2000]. It measures the importance of a node based on how many other important nodes are connected to it. The eigenvector centrality measure showed that three women figures are particularly important in this section of the network: 'Umm imāra', Hind bint utbah, and Al-Rumaysā with eigenvector centrality scores of 0.46, 0.32, 0.32, respectively. Not all the three women are equally well-known in the Islamic history Al-Zarakli [1927]. The first two are involved in famous incidents during the Prophet's battles but the third is not equally famous. However, by surveying their biographical entries, it was clear that the biographies of these three figures highlighted their blood-relationship with other women fighters as well as their combat-activities and bravery in the battlefield (e.g. the biographical entry of Al-Rumaysā starts with describing her as '*the fighter with the dagger in wars and battlefields*'). Thus, it can be seen that the network as well as its visualisation can aid in connecting individuals based on a common theme and bringing out close and remote connections to figures across different ages. As previously mentioned, this is only a section of the women biographical networks, this network can be further expanded thematically and temporally.

4.3 Spatio-Temporal Networks

Another type of topological networks with which we experimented was geographical networks. We aimed to link individuals eminent in particular field with respect to the geographical proximity of their place of birth and death. For this type of networks, we supplemented the information we extracted from the Biographical Encyclopedia with context knowledge. As mentioned in Section 3.2, the classical biographies mention cities and countries whose names and borders have changed in the geography of the modern world. We searched the modern names and

geographical coordinates of places that are recorded in classical biographies. After extracting the geographical data, we used Palladio 1.1[6] to investigate and visualise the results. Palladio is an online toolset designed by Stanford university for the analysis and visualisation of complex, multi-dimensional data. In order to have a meaningful visualisation of the geographical proximity of individuals in our dataset, we opted for narrowing down the research by both the position and age attributes. As an empirical investigation of how far a geographical network is capable of effectively communicating facts about Arabic biographical entries, we narrowed down our dataset to only poets who lived during the Umayyad Caliphate, Abbasid Caliphate (i.e. from 661 to 1258 A.D) and the Modern Period (1800s and 1900s A.D). Figure 5 illustrates a section of the geographical network for the places of birth and death of poets during these three periods.
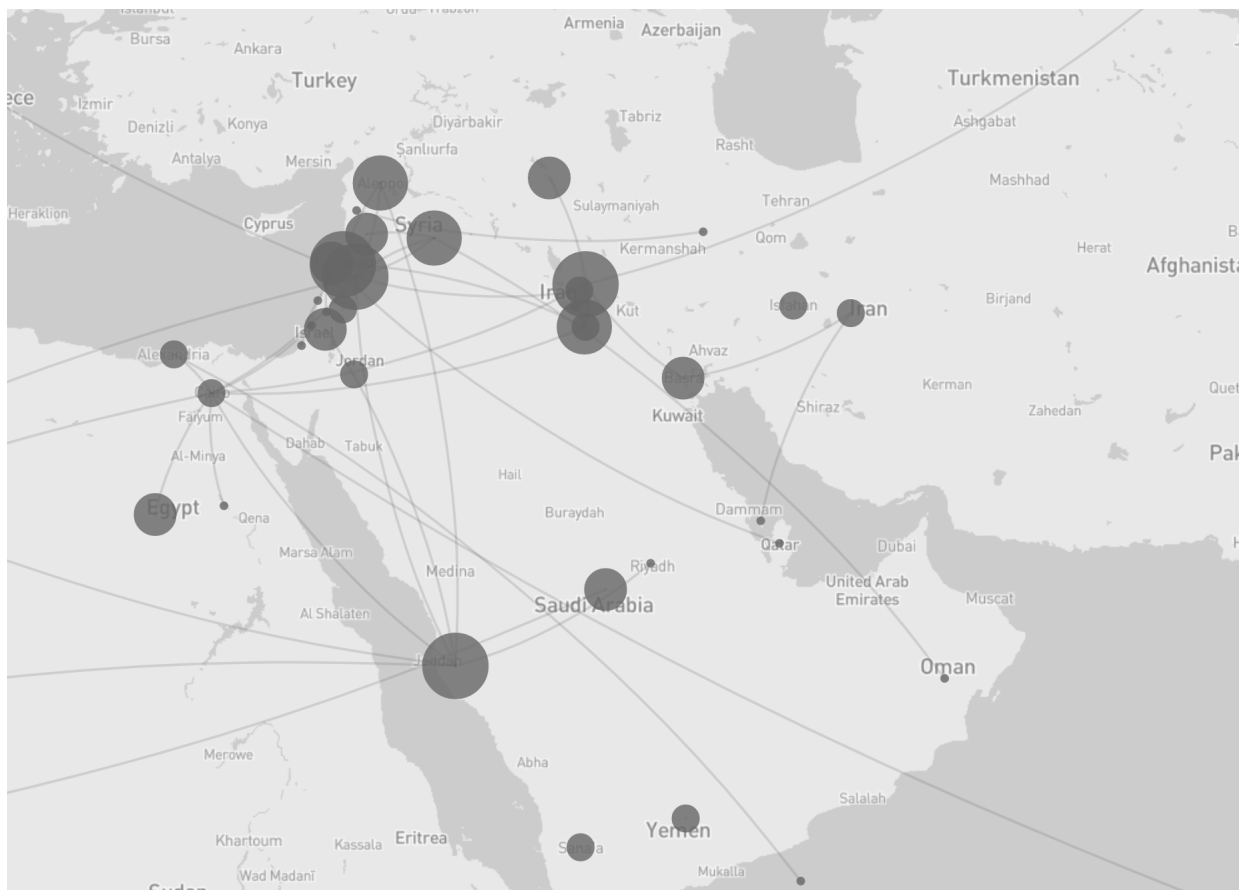

Figure 5: A Geographical Network for Poets Places of Birth and Death

The size of black circles in the figure is indicative of the frequency count of poets related to the respective location on the map and the edges between circles indicate a movement from

[6]Palladio 1.1

one place to another, i.e. birth place to death place or vice versa. It can be seen from the geographical network that the central density of the circles is located in what is historically known as the Levant area. It is a large area in the Eastern Mediterranean region of Western Asia. In modern times, it is equivalent to Syria, Lebanon, Jordan, Palestine and most of Turkey southwest of the middle Euphrates. Also, the edges connected to this focal area is indicative of movements to and from other parts of the world[7]. This shows that the Levantine cities have been a literary capital for poets since the older eras up to the modern times. We also wanted to explore any difference in the geographical pattern of poets' locations in older and modern times. For this purpose, we used the Palladio facet filter to visualise the Umayyad and Abbasid times as one unified period independent from the modern times. Figures 5 and 6 illustrate the older and modern time periods, respectively.

An interesting fact that is revealed by the two figures is that poets in the Umayyad and Abbasid were geographically centred in Iraq and Syria, specifically Damascus and Baghdad, which were the capital cities for these Caliphates respectively. In modern times, however, we have a shift to the West where Egypt, Palestine and Jordan become focal locations for poets either for birth or death. It should be noted that this is a drastically simplified geographical network of poets' locations across different eras. It does point to some facts about the complexities of past events relevant to men and women eminent in one field of knowledge, i.e. poetry. It does not, however, suffice to generate the full insight into the geo-temporal aspects of the biographical data that was collected for this experiment. This example of a poets' geographical network only highlights the potential conclusion that can be arrived at from visualising similar geo-temporal networks for persons distinguished in other fields of knowledge and across diverse periods in the Islamic history.

V CONCLUSION

Many network analysis projects in the social sciences rely on pre-existing data where the attributes are created for network analysis. In this research, we experimented with a large unstructured dataset for a collection of Arabic biographical dictionaries that covered ten eras in

---

[7]Due to space limitations, the full map does not show locations between the Levantine cities and other cities in other continents such as London, Brazil and Indonesia.
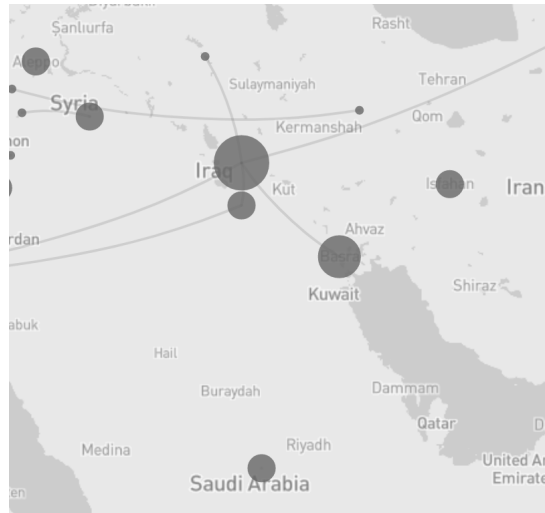
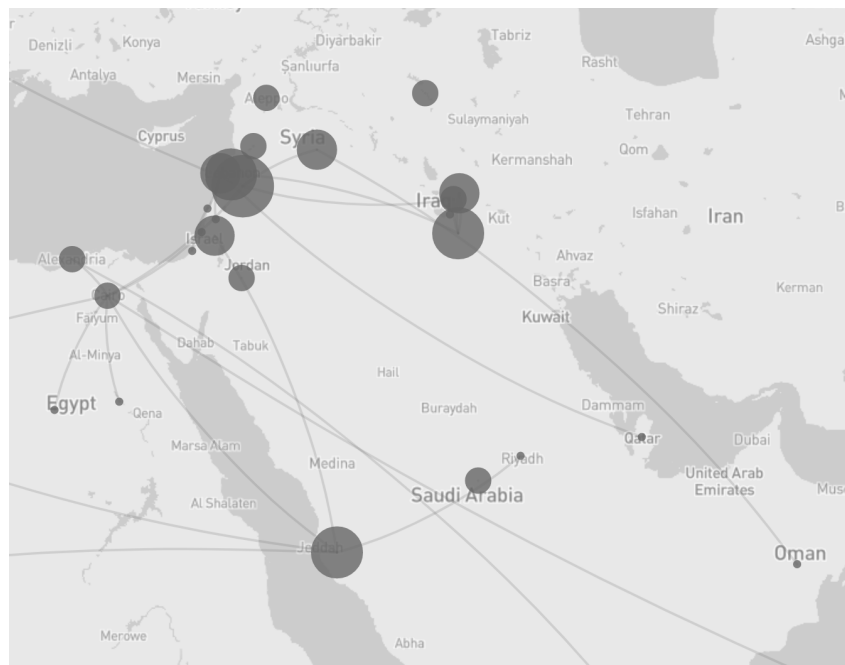Figure 6: Poets Birth and Death Place in the Umayyad and Abbasid Periods

Figure 7: Poets Birth and Death Place in Modern Times

the Islamic History. We managed to extract the network attributes for each biographical entry through different NLP methods. We also used different document similarity metrics to build topological networks which revealed cross-connections between known and unknown figures in different fields of knowledge. The examples of the topological networks created in our experiment have shown that combining NLP tools for text analysis with network theory can bring out complex constellation of relations of different nature, and not exclusively relevant to the individual's field of knowledge. We have also illustrated how the visualisation of networks can help in revealing commonalities between figures, known and unknown, across different eras. More-

over, the geographical networks have proven to be a helpful tool in highlighting a geo-temporal difference for the literary capital of poets across three eras included in our dataset. It remains to be said that our experiments have provided only examples for Arabic biographical networks which revealed hitherto unexpected cross-connections between nodes in the networks. Which aspects of relations between individuals in a biographical network and which attributes matter solely rely on the researcher's viewpoint and research purposes. Our experiment showed that the creation and visualisation of topological networks for biographical data would significantly help researchers in a systematised interpretation of text and unwrap any complex relations that may not be easy to extract from crude unstructured biographical material.

References

Ali Abdul-Fattah. The Prominent Arab Innovators: Arab and Muslim Scientists. *Ibn Katheer Library Publishers*, 2010.

Abu Abdullah Adh-Dhahabi. Biogrphies of the Imminent Nobels. Encyclopedia of Biographies and Imminent Persons. https://www.taraajem.com/, 2006. Accessed: 20-10-2022.

Ahmad Al-Alawnaa. The Modern Arab Scientists Biography. *Al-Bashaar Al-islamya Publishers*, 2011.

Khair al-Deen Al-Zarakli. The imminent: Biography for famous men and women of arabs, non-arabs, and orientalists. Encyclopedia of Biographies and Imminent Persons, 1927.

Korkis Awad. The Arab Scientists Biography. *Nahdit Misr Publishers*, 1986.

Beautiful Soup. Beautiful Soup 4.9.0 , 2022. https://www.crummy.com/software/BeautifulSoup/.

Richard W Bulliet. A quantitative approach to medieval muslim biographical dictionaries. *Journal of the Economic and Social History of the Orient/Journal de l'histoire economique et sociale de l'Orient*, pages 195–211, 1970.

Barbara Caine. *Biography and history*. Macmillan international higher education, 2018.

François Claveau and Catherine Sophia Herfeld. Social network analysis: A complementary method of discovery for the history of economics. *Forthcoming, A Contemporary Historiography of Economics, E. Roy Weintraub and Till Düppe (eds.), Routledge*, 2018.

Melanie Conroy. Networks, maps, and time: Visualizing historical networks using palladio. *DHQ: Digital Humanities Quarterly*, 15(1), 2021.

Michael Cooperson. *Classical Arabic Biography: the Heirs of the Prophets in the Age of al-Ma'mun*. Cambridge University Press, 2000.

Marten Düring. From Hermeneutics to Data to Networks: Data Extraction and Network Visualization of Historical Sources , 2015. https://doi.org/10.46430/phen0044.

Albert Hourani. *A history of the Arab peoples: Updated edition*. Faber & Faber, 2013.

Miriam Kienle. Visualizing networks: Approaches to network analysis in art history. *Artl@ s Bulletin*, 6:4–22, 2017.

Christopher Melchert. *Ahmad ibn Hanbal*. Simon and Schuster, 2012.

Mohammad Nazmabadi, Mohammad Cheloongar, and Mostafa Pirmoradian. A study of historical biographical dictionaries with an emphasis on motives, methods and references for writing them. *History*, pages 23037–23042, 2014.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the 12th language resources and evaluation conference*, pages 7022–7032, 2020.

Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, 32(3):245–251, 2010.

Moshe Perlmann. Gibb, hamilton ar," studies in the civilization of islam"(book review). *Jewish Social Studies*, 26 (4):250, 1964.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Britta Ruhnau. Eigenvector-centrality—a node-centrality? *Social networks*, 22(4):357–365, 2000.

Abdul mumin Bin Abdul haq Albaghdadi Safi Aldin. *The Observationsof Names of Cities and Places, VOL1*. Daru al-Jiil, Beirut, 2008.

Tarajim. Encyclopedia of Biographies and Imminent Figures . https://www.taraajem.com/, 2022. Accessed: 20-10-2022.

Matthew Wilkens. Digital humanities and its application in the study of literature and culture. *Comparative Literature*, 67(1):11–20, 2015.

Peipei Xia, Li Zhang, and Fanzhang Li. Learning similarity with cosine similarity ensemble. *Information Sciences*, 307:39–52, 2015.

M. J. L. Young, J. D. Latham, and R. B.Editors Serjeant. *Religion, Learning and Science in the 'Abbasid Period*, page 168–187. Cambridge University Press, 1990.