TR 13-017

Towards the Analysis of Narrative Networks

Muhammad Aurangzeb Ahmad

May 23, 2013

Towards the Analysis of Narrative Networks

Muhammad Aurangzeb Ahmad
Department of Computer Science
University of Minnesota
Email: mahmad@cs.umn.edu

***Abstract*—The literature on social networks has explored a large number of classes of human networks e.g., citation networks, online social networks, co-authorship networks, event networks, co-location networks, genealogy networks etc. The coverage of networks in the literature is however not uniform, one such neglected area is narrative networks. These networks constitute a class of networks which are formed by chains of narrations from one person to another person, these networks can even span generations. One of the reasons that the field has been neglected is because of the lack of datasets for analysis. In this paper an outline for the network science of narrative networks is given and a historical narrative network is constructed from a 9th century Middle Eastern source book. This network is used for network analysis and to illustrate the major problems in this field and lastly a set of research questions are proposed that should be addressed by researchers.**

I. Introduction

Socialization and connectivity with others is one of the defining traits of human beings. The relationships that people form with one another may be formal, informal, in the context of an organization, familial ties, friendship ties etc. Because these relationships can be expressed in the form of graph, the field of social network analysis arose to analyze these relationships in a variety of contexts e.g., citation networks, partnership networks, co-authorship networks, clandestine networks, online social networks etc. The explosion of social data in the last decade or so has resulted in the expansion of the application of social network analysis techniques in areas where it was not possible previously because of lack of availability of data. Additionally the availability of new datasets allows the researchers to also ask new questions which could not have been asked previously. However even with the advancement in data collection and new analysis techniques there are certain types of networks which have not been studied to any appreciable extent, one such class of networks is the class of narrative networks. The reason for lack of studies in this area is not because of lack of interest but rather lack of datasets. In this paper we offer a solution to remedy this problem.

Narrative networks refer to a class of social networks which are formed by creating edges between people who have narrated stories, reports, incidences etc from one another. These can be historical or contemporary in nature. The graphs used to represent these networks are directed graphs where the incident node corresponds to the person from whom the narration was originally taken. The networks closest to these networks in terms of semantics are networks used to study rumor-spreading, chains of reports etc. We note that one major difference between these types of networks and narrative networks is that while the former uses already existing ties for the spread of information the narrative network not only use already existing ties but new ties may be be added to the narrative networks for the sake of narrative continuity [5]. To the best of our knowledge datasets related to narrative networks are not available to the research community and thus progress in this area has been limited as a result.

One emerging sub-field in the area of social network analysis is the field of analysis of historical networks [23]. While there are many issues with respect to construction of historical social networks e.g., historical omission, entity resolution etc [24][23] which effect historical social network analysis, historical sources also offer a new source of network data as well as a new way of looking at history. There are also many historical documents that can be mined to construct social networks and historians have only started to look at these documents from a network perspective [26]. To date almost all of historical network analysis has been limited to constructing networks from Western sources and languages. Since a large amount of historical sources are in non-Western sources and the reason and the compilation of these sources are often different from their Western counterparts, they can offer us alternative and in some cases even richer sources of data[18].

In this paper we construct a historical social network of narrators spanning over the course of a hundred years from an Arabic *hadith* book, from the ninth century. A *hadith* refers to the sayings of Prophet Muhammad, the foundational figure of the religion of Islam. Hundreds of thousands of such sayings were recorded in the collection of *hadiths* [5] after his death. Since these were written almost a century after his time, each narration from Muhammad is recorded as a chain of narration from the person who wrote them down to the people that the collector heard the narration from and from the person that this person in turn heard it from and so on up to the person who had heard the narration from Prophet Muhammad himself. These *hadiths* are further classified on the basis of their authenticity. Out of total around fifteen thousand or so are considered to have the highest level of authenticity, these hadiths are called *Sahih*. In this paper we use one of the most famous and authentic collection of hadiths called *Sahih Bukhari* [5] as a test case for the analysis of narrative networks.

The analysis of *hadiths* from a network perspective has the potential to not only expand the range of subjects available to network science but it also has the potential to have a transformative effect on other fields as well e.g., early Islamic Historiography [5], the study of social networks in pre-modern societies as well as non-Western societies and lastly *hadiths* have always have had an important role in Islamic legal thought [15]. The rest of the paper is organized as follows: In section

II related work is discussed, section III describes some of the terminologies associated with Narrative Networks, IV given an overview of the data as well the historical background of the data, section V gives the analysis of the narrative networks, open problem in this area off research are discussed in section VI and lastly the conclusion section not only summarizes the conclusions but also discuss the current and future efforts in this area.

II. Related Work

While the field of social network analysis goes all the way back to 1920s and plethora of various types of social networks have studied [22] in this field, to the best of our knowledge narrative networks have not been systematically studied before. As described in the introduction the unavailability of datasets precluded the analysis of narrative networks to be undertaken previously. The use of rich historical sources can solve this problem. The approach that we take is to mine such networks from historical sources in Arabic. With respect to network analysis of social networks extracted from historical Islamic sources the work by Senturk which deserves a mention [20]. Senturk looked at networks of the scholars of hadiths i.e., people who studied hadith rather than the networks of narrators. He limited his study to a list 1,161 prominent *hadith* scholars from the seventh century to the fifteenth century to determine how the transmission of hadiths and cannonization effected and was effected by the structure of these networks. There is already a large corpus of studies on the text of the *hadiths* as well the people who are mentioned in them [1] but they have not been studied from a network perspective. As a consequence the study of the texts themselves are of less relevant for our purpose.

The closest example of networks that are similar to the narrative networks that have been studied before are pedagogical networks e.g., adviser-student networks [12], genealogical networks [25], networks constructed for clandestine information sharing [3],[10] and the phenomenon of rumor spreading in networks [14]. The last topic has especially been extensively studied in the social network analysis literature [16]. However the dissimilarity of narrative networks to all of these networks warrant them to be studied separately. Another area which is relevant to our work is the somewhat nascent field of analysis of Historical Social Networks [23]. The main issue in this field has been the same i.e, issues related to the construction of social relationships given the scarcity of data [24] and thus there is the potential of the current work to contribute to this area as well.

III. Network Science of Narrative Networks

In this section we give a graph theoretic formulation of the terms associated with narrative networks which will be used in the rest of the paper.

Narrative Network: A narrative network is a directed graph $G = \{V, E\}$ where an edge $e_{ij} \in E$ represents a report from node v_i to node v_j. The graph G consists of the union of k paths such that paths of the form $p = \{v_1, v_2, ..., v_m, v_1\}$ i.e., loops are not allowed.

Narrative Chain: A narrative chain is a directed path $p = \{v_1, v_2, ..., v_n\}$ in the narrative network $G = \{V, E\}$ such that each path semantically represents the transmission of a narration from the source v_1 to the destination v_n. A node v_i in the path p is stated to have heard the narration from the node v_{i-1}.

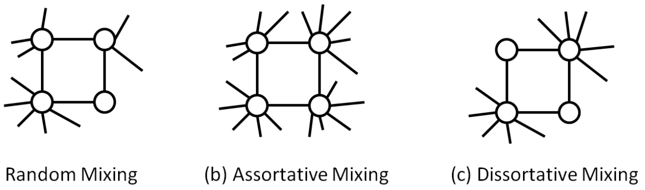


Fig. 1. Mixing Types in Networks

Single Source Narrative Network: In some networks there may be just a single source where all the chains of narrations are traced to that source. This is especially true for networks where are biographical in nature. A single source narrative network is a directed graph $G = \{V, E\}$ where an edge $e_{ij} \in E$ represents a report from node v_i to node v_j and all paths $p_i \in P$ have a primary node $v_a \in V$ as the first element of the path p_i.

Diameter of Narrative Networks: The diameter of the narrative network is the maximum eccentricity off any vertex in the network. It should be noted that this not the same as the maximum path length of a narrative path since two or more chains have nodes in common which may thus combine to create paths that are greater in length than any of the constituent nodes.

Narrative Assortativity: Network Assortativity refers to the correlation between a node's degree with its neighbors' degrees. There are three main types of assortativie mixing. If there is no or negligible mixing then it is referred to as random mixing. Assortative mixing is where thee is positive correlation and corresponds to the cases where well-connected nodes are connected to other well-connected nodes and the same applies to poorly connected nodes. Disassortativie mixing is when there is negative correlation and well-connected nodes are connected to poorly-connected nodes. The three types of assortativities are illustrated in Figure 1. We use the scheme described by Newman et al [17] for measuring assortativity.

Influential Nodes: Given that narrative networks are historical as well as temporal in nature the idea of influence can be defined in multiple ways. **Local Influence:** The local influence $inf_k(n)$ of a node v is the influence of a node measured by only considering the subgraph of the node up to k hop distances from the node v. This metric is agnostic to the formulas that may be used to compute influence e.g., PageRank, Degree Centrality, Betweenness etc. **Topical Influence:** Given a directed graph $G = \{V, E\}$ such that each node $v_i \in V$ has a vector of topics $T_i = \{t_{i1}, t_{i3}, t_{i3}, ..., t_{in}\}$ associated with it. The topical influence of the node is with respect to a given topic $t_k \in T$. Similar to local influence, topical influence is also agnostic to the metric used.

Narrator-Subject Matrices: Given that a narrator may narrate on multiple subject, and alternatively different terms may occur more prominently with one person as compared to another person even if the subjects are not given beforehand, it is possible to construct narrator-subject matrices where the

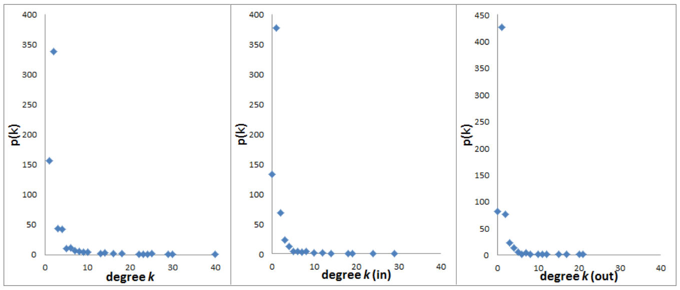


Fig. 2. Degree Distribution of the Accumulative Network

data is available. Given such a matrix M, an entry m_{ij} refers to the narrator v_i and the subject matter s_j. The values in the matrix refer to the normalized frequency of the presence of such terms. The idea is similar to Term- Frequency matrices used in the field of Information Retrieval [19]. This opens up the possibility of using the whole repertoire of techniques from the field of information retrieval to the analysis of *hadiths*.

IV. Dataset

A. Historical Background

Sahih Bukhari is considered to be the most authentic book of collections of sayings of Prophet Muhammad [8][9][5][1]. The book was compiled by Imam Bukhari in the ninth century after traveling far and wide in the Islamic world and collecting the narrations from individually from people [5]. The book contains more than seven thousand narrations and each narration is accompanied by a chain of narrations of the form ”person A heard from person B who heard it from person C .... who heard prophet Muhammad saying this narration.” Taken together the book contains thousands of narrators and thousands of chains, the existence of the chains with reliable narrators was considered necessary to ascertain the reliability of the *hadiths*. The field of *Ilm'ar'rijal* was developed around the same idea to to determine the authenticity of narrations by not only recording information regarding the character of the narrators but also their biographies [9][5]. The field of *Ilm'ar'rijal* contains a large number of books which are essentially biographical registers of such people. The number of people whose brief biographies were recorded eventually ran into tens of thousands. Even Imam Bukhari himself compiled such a book, *al-Tarikh al-Kabir*, in the field of biographical evaluation[5]. The availability of information regarding the narrators can be used as an additional data for analysis in the future. The *hadith* related terminology that we use in this paper is taken from Ibn al-Salah's terminology on the field of *hadith* [8] which is considered to be the standard in the area of *hadith* studies [5].

The process that we used to reconstruct the social network of the narrators can be described as follows: Multiple people first go through the text of each narration and extract the names of narrators. In many cases the chain is not always given as an ordered list or even the names of the people are not given in a straightforward manner. These issues are discussed in detail in section IV-B. The chains are first recorded in Arabic since it ensures that the names are standardized as there are multiple ways to transcribe the same name from Arabic to English e.g., Umar and Omar are two different ways to transcribe the same Arabic name in English. The chains are then re-transcribed in English, once the transcription has been done the names of each person in the network is checked for duplicates. Entity resolution in this case is not just about the variants of the names of the person but also that teknonym of the same person may used in different places. This issue is also discussed in more detail in the next section IV-B. Transcribing the data in English also has the advantage of making the data easily available to the larger scholarly community as we plan to release the data to the wider research community in the future. We plan to partially automate the process of extracting these networks with some human supervision in our future work. Since all of the narrations were collected by a single person (Imam Bukhari) and these can be traced back to a single source (Prophet Muhammad), we do not include these two nodes in the network. Given the limitations of time and resources and the current manual nature of network reconstruction, we were only able to extract the networks from around five percent of the text from *Sahih Bukhari* but even that corresponds to a network of 640 people.

B. Data Preparation Issues

The narrative networks were extracted manually from the Arabic text by going through each individual narration and noting down the person who is part of the chain of narration. From a data transcription and data quality perspective there are a number of issues with respect to how the narrations are recorded. In some cases the partial names of the narrators are given and in other cases full names are given e.g., the name Hisham ibn Urwa vs the name Urwa, both of which refer to the same person. We thus consulted domain experts who are able to manually do entity resolution for this purpose. In some cases the relationship of one of the narrators to another narrator is given without any information regarding the name of the person. Thus narration 37 in the chapter *Kitab Muwa'qeet As'salah* (Book of Prayer Times) terminates with ”Salim ibn Abdullah heard it from his father.” In this case the identity of his father has to be disambiguated based on additional knowledge since the same person may be recorded under his name in another narration. An analysis of other texts [9] reveals that his father is Abdullah ibn Umar who is known to have narrated a large number of *hadiths*.

There are other cases where general information is given regarding the narrators instead of individual names. Thus in narration 61 of the same chapter the first link in the chain from Prophet Muhammad is Umar but instead just his name the narrator says that ”Ibn Abbas heard it from a number of people including Umar.” This implies the presence of multiple people but it is impossible to determine who these people are. In the Arab culture the teknonym of a person, known as *kuniya*, is often used instead of his or her given name i.e., the name of the person is recorded as the father or mother of so and so. As an example, Abu Musa Alashari is the same person as what the person in a chain who is recorded as just the father of Abu Bakar bin Abu Musa.

There are additional cases where more detailed information about family relationships is required. Consider the following illustrative example: In narration 24 of the same chapter it states that ”Hisham bin Urwa heard it from his father who was the nephew of Ayesha.” The identity of these people is not

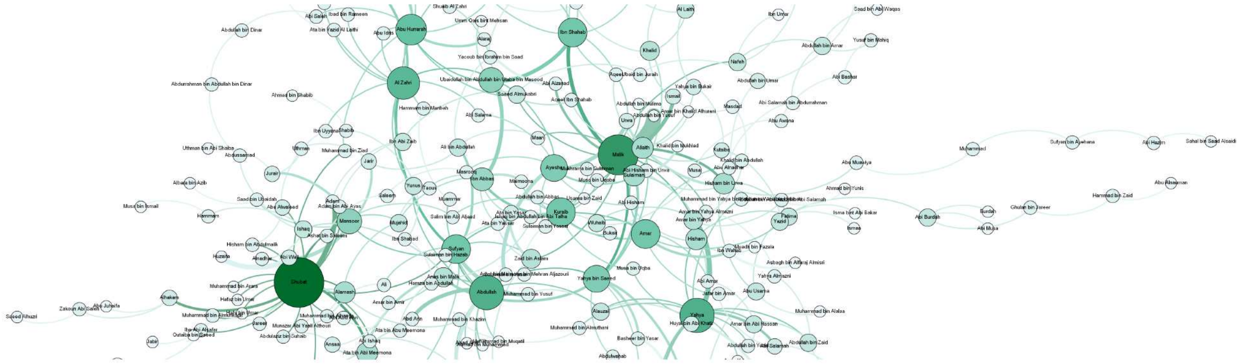


Fig. 3. A snapshot of the Narrative Network from *Kitab Jummah*

evident until one consults their genealogy from other source and it is established that the father of Hisham bin Urwa is Urwa bin Zubair bin Alawam whose father Zubair bin-Alawam was married to Isma bint Abu Bakar who was the sister of Ayesha, the wife of Prophet Muhammad and the first person in the chain of narration. Lastly there are cases where the narration starts or ends with "and he said" and thus determining the name of the person in this case is a challenging task. We note that even after the rigorous process described in this section there are likely to be issues of data quality that were missed. We plan to constantly refine the quality of the data after each release of the dataset.

V. Analysis of the Sahih Bukhari Network

A. Network Characteristics

Sahih Bukhari is organized in many chapters where the heading of the chapter refers to the subject matter being discussed in the chapter e.g., Book of Prayer times, Book of Prayer of Repentance etc. Each chapter is referred to as a book. Thus it is possible to construct separate social networks for each chapter and do a comparative analysis for each chapter. Given the size of the networks it is not possible to provide good visualizations along with the labels, thus a partial snapshot of one such network is given in Figure 3. The size of the nodes represent the relative importance of the person in the network i.e., how many *hadiths* did he or she narrate or how many were narrated from that person. Similarly the size of the edges represent strength of relationships. The prominent nodes are the ones that have been noted to be prominent in the *hadiths* literature. These are discussed in detail in the next sub-section.

There are 54 chapters in total and we have currently reconstructed the narrative network of 5 different chapters as given in Table IV-B. Since each chapter also corresponds to a different subject matter one can also compare the network graphs from each chapter to determine which narrators talked about which subjects. From the table it is clear that even though some of the networks have relatively large diameters, the average path length is relatively small as evident in Table IV-B. This implies that in most cases the distances between the narrator and the source i.e., Prophet Muhammad is relatively small. This makes sense given that Imam Bukhari compiled *Sahih Bukhari* in the second century of Islam.

If we consider the accumulative network then it is observed that the degree distribution follows a long tail distribution observed in most other networks except the difference being that more nodes have 2 connections as compared to just one connection as given in Figure 2. A relatively large number of nodes, around ten percent, have an out-degree equal to zero which implies that they do not narrate from any other source. These are the people who directly heard the saying from Prophet Muhammad. We also compared the overlap in narrators across the various chapters as shown in Table V-A. The overlap is quite low which is to be expected since over the course of time the subject matters that people narrate becomes more and more specialized[20]. This implies that common subject matter across narrators would become less likely over time. Also as expected the longer networks have greater average degrees but lower density.

TABLE II. Jaccard Index for Common Nodes across *Sahih Bukhari*

	Jummah	**al-Eidaan**	**al-Khauf**	**Muwaqeet**	**al-Wudu**
Jummah	1	0.034	0.035	0.112	0.113
al-Eidaan		1	0.081	0.043	0.034
al-Khauf			1	0.036	0.028
Muwaqeet				1	0.105
al-Wudu					1

Figure 5 gives the visualization of the narrative networks from all 5 chapters as well as the network that is obtained by the union of all the other networks. From the visualizations it is clear that some chapters have more chains and thus more narrators as compared to others. Another interesting thing to note is that in this accumulative network approximately 98 percent of all the nodes belong to the largest connected component. This also implies that these men and women were part of a community of scholars and transmitters and most of them knew one another. The same is true for the larger individual chapter and one can conclude that the isolated components that one sees for smaller chapters is because of the small sample size.

B. Analysis of Individual Narrators

We now consider some of the prominent nodes in the network and corroborate this information from what is already known from historical sources. Table III gives the in-degree, out-degree, PageRank (used to determine the relative importance of a node) [4], EgoNet(3) refers to the percentage of nodes which are connected to the node within three hop distances from that node and the Betweenness Centrality of node.

TABLE I. NETWORK STATISTICS FROM SAHIH BUKHARI

Arabic Name	Translation	Nodes	Edges	Avg. Degree	Density	Components	Diameter	Avg Path Length
Kitab Jummah	Friday	208	231	1.111	0.005	6	9	3.811
Kitab al-Eidaan (Salah)	Eid Prayers	38	33	0.868	0.023	5	5	2.120
Kitab al-Khauf (Salah)	Prayer of Repentance	29	25	0.89	0.031	4	5	2.338
Kitab Muwaqueet as'salah	Prayer Times	228	271	1.189	0.005	6	9	3.047
Kitab al-Wudu	Ablution	296	386	1.304	0.004	3	14	4.873
	Accumulative	640	888	1.388	0.002	6	19	6.136

TABLE III. LIST OF PROMINENT NARRATORS

Narrator	In-deg.	Out-deg.	PageRank	Betweenness	EgoNet(3)	Historical Notes
Abu Hurrarah	29	0	0.031	0	35.94	One of the most prolific narrators and considered to be an authority on narrations
Anas Bin Malik	24	1	0.028	268	38.59	Another prolific narrator, the last person from Muhammad's generation to die
Ayesha	18	0	0.021	0	28.59	Wife of Prophet Muhammad, considered to be a major narrator
Ibn Abbas	14	2	0.016	318.48	27.03	Paternal cousin of Muhammad, a famous scholars of early Islam
Imam Malik	10	20	0.005	8908.60	50.47	One of the founders of the 4 Sunni scholars of jurisprudence
Ibn Shihab	8	10	0.005	1907.8	33.44	One of the famous collectors of narrations of *Sirah* from the 2nd century of Islam
Abdullah Ibn Umar	4	1	0.005	1410.5	26.88	Son of the second ruler of the Islam Caliphate (Empire)
Imam Al-Ouzai	2	2	0.001	9273.88	28.91	Famous Muslim jurist from the 2nd generation after Muhammad

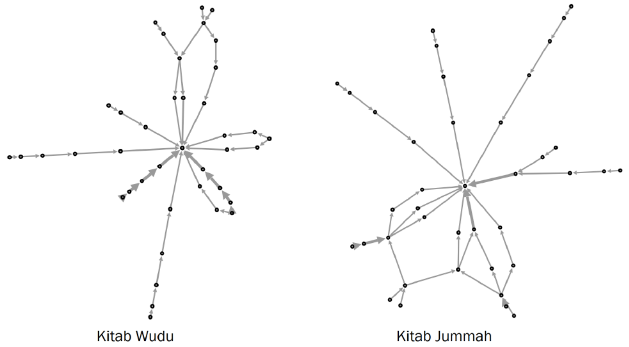


Fig. 4. The ego-Narrative Network of Abu Hurrarah from two different chapters of *Sahih Bukhari*

The top two people with the highest values for PageRank, Abu Hurrarah and Anas Bin Malik are also known to be one of the most prominent transmitters of narrations from Prophet Muhammad [5][9], hence it should not be surprising that their name appeared at the top of the rank list of PageRanks. What is surprising is that the analysis of even a small subset of the hadith bears this out. Consider Abu Hurrarah, if one reconstructs his ego-network then one can even see the relationship between narrators across generations. Thus consider Figure 4 which shows the ego-network of Abu Hurrarah for *Kitab Wudu* and *Kitab Jummah*. The later has a greater branching factor than the former and the chains in *Kitab Jummah* are in average longer than the chains in *Kitab Wudu*.

The person with the third highest value for PageRank is the youngest wife of Prophet Muhammad, Ayesha. Since she lived for fifty years after the death of Prophet Muhammad [13] she had the opportunity to transmit a large number of sayings of Muhammad to a number of people. The next person in the list, Ibn Abbas, was the cousin of Prophet Muhammad who also outlived him for many years and transmitted many narrations from him. The next two people in the list have much lower PageRank values but higher values for Betweenness. Imam Malik is the founder of one of the four main schools of jurisprudence in Sunni Islam [15]. The high Betweenness value also makes sense for Imam Malik given that his connection to a large number of people in the formative periods of Islam as well as having a large number of students gave additional credence to his authority. His position in the network also bears thus out.

The next person in the list, Ibn Shihab is a well known early collectors of the life of Prophet Muhammad as well as hadiths. He is extensively quoted by later scholars of hadith and is considered to be an authority on the subject [5]. Network analysis of the Bukhari network also yields some surprises. We considered the ego-network of another prominent narrators Abdullah Ibn Umar. In one of the chapters, *Kitab Jummah*, he only narrates from father and no one else narrates any hadiths from his father. In other chapters however this is not the case as both of them have other sources and narrators associated with them. This also implies that certain chapters are likely to have exclusive cliques of narrators associated with them.

The last person on the list Imam Al-Ouzai has a very low in-degree and out-degree but a large value for Betweenness Centrality [22]. This indicates that a large number of shortest paths pass through this node in the narrative network. We note that Betweenness Centrality is defined as the number of shortest paths from all nodes to all other nodes that pass through that node. Given a node v, let σ_{st} be the total number of shortest paths from from node s to node t and let $\sigma_{st}(v)$ be the set of paths from node v then Betweenness centrality can be given as follows:

$$g(v) = \Sigma_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

In the current context Betweenness connotates whether the person plays an important role in the transmission of the hadith after the initial set of transmitters from Prophet Muhammad. The role of the person with a high value for Betweenness Centrality is similar to a person with high social capital in other social networks [6] with the difference that in the narrative network the position reflects historical importance in transmission of narrations. While Imam Al-Ouzai is not as prominent as the other narrators in this field but is a well-known jurist from the early centuries of Islam [7]. Having a high Betweenness Centrality score puts him in an advantageous position with respect to connecting within the community of scholars in his era and influencing the development of Islamic scholarship.

We also discovered a number of instances where a single narration breaks into multiple chains because Imam Bukhari heard the narration from multiple people in different locations

TABLE IV. RANKED LIST BY CHAPTER

Rank	Jummah	Muwaqeet	Wudu
1	Abu Hurrarah	Abu Hurrarah	Abu Hurrarah
2	Abdullah ibn Umar	Anas bin Malik	Ayesha
3	Umar bin Al Khattab	Zaid bin Thabit	Anas
4	Ayesha	Ayesha	Maimoona
5	Anas bin Malik	Ibn Abbas	Ibn Abbas

or cities. Many of these chains have multiple people in common.While the text is the same we consider the chains as three distinct chains from the perspective of network analysis. The following example from Narration 54 from *Kitab Muwa'queet As'Salah* can be used to illustrate this. It refers to the same saying of Prophet Muhammad:

1) Huzaba bin Khalid → Hammam → Abu Jimrah → Abubakar bin Abdullah bin Qais → Abu Musa Alashari
2) Ibn Rajah → Hammam → Abu Jimrah → Abubakar bin Abdullah bin Qais → Abu Musa Alashari
3) Ishaq → Habban → Hammam → Abu Jimrah → Abubakar bin Abdullah bin Qais → Abu Musa Alashari

Consider Table IV which has the ranked list of the top five narrators from the three larger books in *Sahih Bukhari* based on their PageRank and thus we can compare their topical influence. As described in earlier sections and in table III some of the names are the names of prominent narrators like Abu Hurrarah, Anas bin Malik, Ibn Abbas and Ayesha are to be expected in this list. Umar bin Al Khattab is ranked prominently in Kitab Jummah. While he is not known to have narrated as many *hadiths* as others he does have a prominent role in the early Islamic history, being the second ruler of the Islamic Caliphate (Empire) after the passing away of Prophet Muhammad. While the text itself does not mention it but Maimoona who is ranked 4 in *Kitab Wudu* is most likely the other wife of Prophet Muhammad and thus figures prominently in the network.

C. Assortativity in Sahih Bukhari

Assortativity in narrative networks is a measure of how prolific a person's neighbor's are. If a person's narrations are further narrated by prolific narrators then he or she is likely to have greater influence in the long run since it increase the chance of being heard by more people. The assortativity of the network as a whole is -0.0317 which implies that this is a disassortative network i.e., nodes with high connectivity are connected with nodes with low connectivity. Newman observed that most social networks are assortiative in nature while technological and biological networks are dissassortative in nature [17]. The reason that the current network does not follow this trend is because the hadith network seems to have a small number of sources who do not in general narrate from one another and thus are connected with people with low connectivity. Additionally some people further down along the chain who are known to be prolific teachers do not generally have prolific teachers.

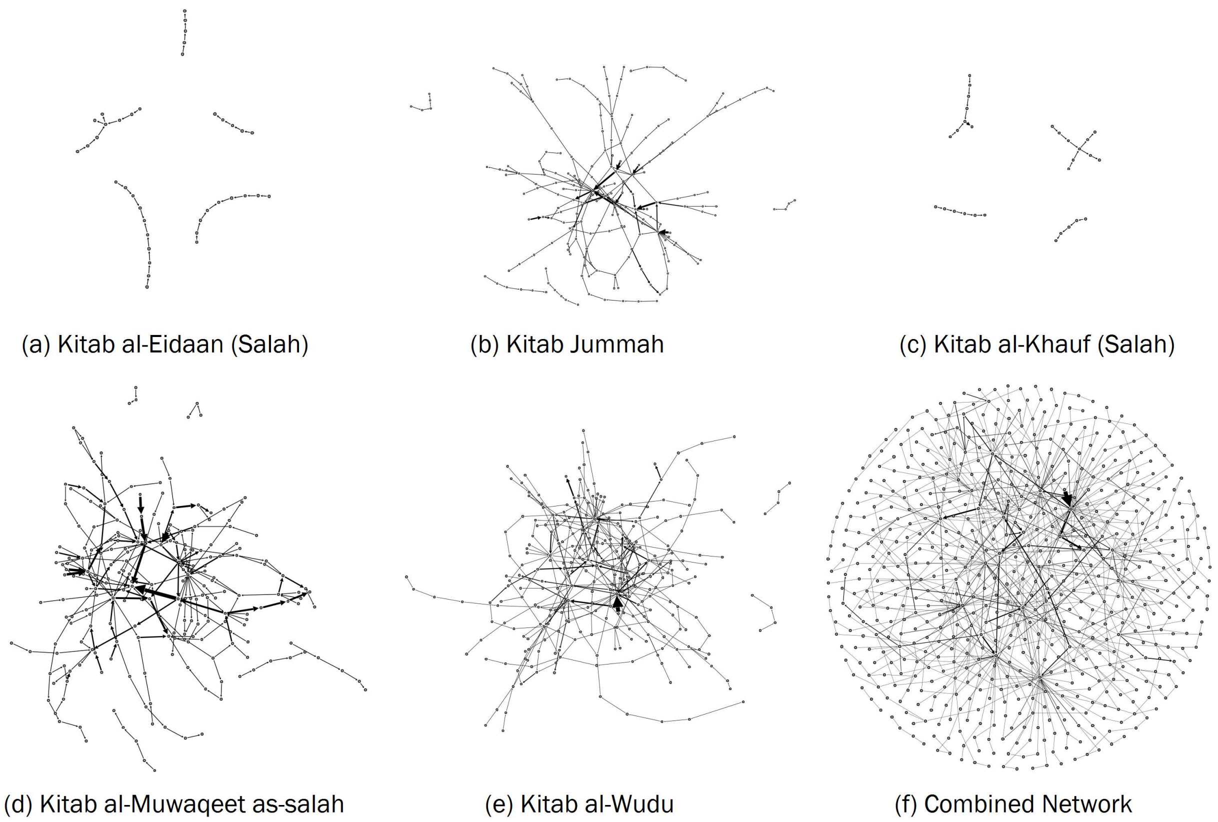
If however we compute the individual networks from each of the larger chapter then a more interesting picture emerges: Both *Mawaqeet as'salah* (0.0275) and *Jummah* (0.0201) chapters have positive values for assortativity (-0.0179) but the chapter *wudu* has a negative value for assortativity. While the difference between the these values may seem small we note that they are in line with the differences that have been observed for technological networks (World Wide Web) vs. Social Networks (Collaboration Networks) [21][17]. A closer analysis of the first two chapters reveal that there are indeed many more cases where poorly connected nodes are connected to other poorly connected nodes as a proportion of the total number of nodes but that is not the case when we take the union of all of these chapters. Based on these observations we conjecture that the network consisting of the complete network of narrators from *Sahih Bukhari* would be a disassortative network.

VI. OPEN PROBLEMS

In this paper we have described the problem of the analysis of narrative networks and used a historical network of narrators constructed from a 9th century early Islamic text to show the viability of this enterprise. This work is meant to be a proof of concept as the field is in its infancy and a great deal of work still needs to be done. We now describe some of the problems that should be addressed by the research community. Many of these are motivated by the problems that scholars in the hadith community come across and others are motivated by questions that can benefit the community if they are framed in network terms. Since narrative networks correspond to a different class of networks than what have been studied before, it is not known how these networks evolve and thus they offer a fruitful avenue for network modeling as well as determining the statistical properties of these networks [10]. An important research problem for the *hadith* research community is determining the authenticity of a *hadith*. A large chunk of literature has been devoted to this subject [5][1] but open problems still remain. We note that one can take a network perspective on this subject. Especially techniques from link prediction [2] can be modified for this purpose. In this paper we have used graph as the standard representation of narrative networks, this assumption can also be generalized if one uses hypergraphs to extend the representations to the topics and locations of the narrators as some network problems are more amenable to be solved by hypergraph representations [3].

There are also a number of challenges with respect to entity resolution in this domain: As described previously the names of some narrators are partially recorded in some cases and in some cases the full names are given, in some cases the teknonym of a person is used e.g., father of so and so etc. For some of the well known people in history it is relatively easy to disambiguate their identities. The same cannot be said about more obscure people mentioned in the *hadith* books and thus a better systems needs to be developed to link individual names with biographical registers so that this problem can eventually be automated. While temporal information is not directly available in the text itself it can be partially inferred given that the chain of narrators spans over the course of more than a hundred years and four to five generations for Sahih Bukhari [5]. In other books the time may span for more than two centuries and many more generations [9]. Automatic inference of these cohorts would be greatly beneficial to the scholarly community.

As described previously the *hadith* corpus consists of a



(a) Kitab al-Eidaan (Salah) (b) Kitab Jummah (c) Kitab al-Khauf (Salah)

(d) Kitab al-Muwaqeet as-salah (e) Kitab al-Wudu (f) Combined Network

Fig. 5. Networks for each Chapter of *Sahih Bukhari*

large number of books and we have only considered one book for analysis as a start. Linking narrators across books to topics can also help us determine how do various schools of thought, especially in the area of jurisprudence emerged in the early history of Islam as the development of the various schools of thought in that era was influenced by the availability or the unavailability of certain narratives in different areas [5]. Secondly such an analysis could be extended and connected to the work that Senturk has done on studying the networks of scholars who studied *hadith*. For historians this could become an indispensable tool to study the early history of what of then the emerging religion of Islam. Lastly and most importantly the underlying assumption in many of the aforementioned problems i that a good quality dataset is available for analysis. The problem of entity described previously is just one example of data quality related issues. Outside of the canonical collections of *hadiths* there are still texts where the information regarding the chains as well as the narrators themselves is scattered across multiple texts. From a text mining perspective this offers an interesting challenge to the scholarly community.

VII. Conclusion and Future Work

In this paper we described the problem of the analysis of narrative networks. We noted that the main reason that progress has been slow in this field is because of the lack of availability of datasets. We proposed that this gap can be bridged by constructing social networks from historical sources, in our case a 9th century Arabic text from the early Islamic history. Not only can network analysis of historical social networks confirm what we already know but it has also the potential to yield new insights about historical figures and their interactions. Thus this has the great potential to be a useful tool for analysis for historians. One of the main takeaways from this analysis is that while we have only considered a small subset of the hadith for our analysis one can already see the many salient features of well known facts about the hadith literature emerge. Additionally some non-obvious facts are also beginning to emerge as well.

In the future we plan to extend the analysis to cover not only the entire collection of *Sahih Bukhari* but also extend it to other books as well. In the future we plan to add a vector of annotations to each narrator from the books of biographical registers described previously e.g., gender, location, ethnicity, topics narrated etc as well as labels for some edges e.g., pedagogical relationships, employment relationship and familial relationships. A more complex network can thus be created and analyzed. Given that the text is available in addition to all of the relationships that we have described here, there is also room for the application of and even extension of information retrieval techniques [19] as well applying and extending current graphical models for topic extraction [11] on this type of datasets. Lastly we plan to release the *hadith* dataset and background on this project to the research community as more and more data is transcribed in the near future. The research community will thus be able to pursue some of the problems described here as well as pursue new directions in research.

VIII. Acknowledgment

The author would like to thank Omar Mehana and Mohamed Elbadry who helped in transcribing the chain of narrations from Arabic to English. Special thanks to Tarig Benega for not only help in the transcription but also in navigating through the issues in transcription in the original Arabic text.

References

[1] Muhammad ibn Jafar Al-Kattani *Al-Risalah al-Mustatrafah* (Beirut: Dar al-Basha'ir al-Islamiyyah) Seventh edition, 2007

[2] Muhammad Aurangzeb Ahmad, Zoheb Borbora, Jaideep Srivastava, Noshir Contractor *Link Prediction Across Multiple Social Networks* Domain Driven Data Mining Workshop (DDDM2010), ICDM 2010 Sydney, Australia.

[3] Muhammad Aurangzeb Ahmad, Brian Keegan, Dmitri Williams, Jaideep Srivastava, Noshir Contractor *Trust Amongst Rogues? A Hypergraph Approach for Comparing Clandestine Trust Networks in MMOGs* 5th International AAAI Conference on Weblogs and Social Media (ICWSM-11) 2011

[4] Sergey Brin, Lawrence Page *The Anatomy of a Large-Scale Hypertextual Web Search Engine* Computer Networks 30(1-7): 107-117 (1998)

[5] Jonathan Brown *The Canonization of Al-Bukhari and Muslim: The Formation and Function of the Sunni Hadith Canon*. Brill, 2007

[6] Ronald Burt *Structural holes: the social structure of competition*. Harvard University Press (1995)

[7] John Esposito *The Oxford Dictionary of Islam*, Oxford University Press, 2003

[8] Ibn al-Salah *An Introduction to the Science of the adith*Translator Dr. Eerik Dickinson Reading: Garnet Publishing Limited, 2006

[9] Ibn Hajar al-Asqalani *Tahdhib al-Tahdhib* Hyderabad, Deccan: Da'irat al-Ma'arif al- Nizamiyya, 1910.

[10] Brian Keegan, Muhammad Aurangzeb Ahmad, Dmitri Williams, Jaideep Srivastava, Noshir Contractor, *Dark Gold: Statistical Properties of Clandestine Networks in Massively-Muliplayer Online Games* IEEE Social Computing Conference (SocialCom-10) Minneapolis, MN, USA, August 20-22, 2010.

[11] D. Koller, N. Friedman *Probabilistic Graphical Models*. Massachusetts: MIT Press 2009

[12] Allyn Jackson *A labor of love: the Mathematics Genealogy Project*, Notices of the American Mathematical Society 54 (8): 10021003. 2007.

[13] Martin Lings *Muhammad: His Life Based on Earliest Sources*. Inner Traditions International (1987).

[14] D. P. Maki *Mathematical Models and Applications, With Emphasis on Social, Life, and Management Sciences*, Prentice Hall. 1973

[15] Christopher Melchert *The Formation of the Sunni Schools of Law: 9th-10th Centuries* C.E., pg. 178. Leiden: Brill Publishers, 1997.

[16] Yamir Moreno, Maziar Nekovee, Amalio F. Pacheco *Dynamics of rumor spreading in complex networks* Physical Review E, Vol. 69, No. 6. (2004)

[17] M.E.J. Newman *Assortative mixing in networks*. Physics Review Letters 89, 208701 (2002)

[18] The Princeton Library of Islamic Manuscripts http://www.princeton.edu/~rbsc/department/manuscripts/islamic.html Retrieved April 11, 2013

[19] G Salton, MJ McGill *Introduction to modern information retrieval*. McGraw-Hill. 1986

[20] Recep Senturk *Narrative social structure: Anatomy of the hadith transmission network* 610-1505. Stanford University Press, Stanford, CA. 2005

[21] Lovro Subelj, Marko Bajec *Clustering assortativity, communities and functional modules in real-world networks*. CoRR abs/1202.3188 (2012)

[22] Stanley Wasserman, Katherine Faust *Social Network Analysis: Methods and Applications*. New York and Cambridge, Cambridge University Press. 1994

[23] Barry Wellman, Charles Wetherell *Social network analysis of historical communities: Some questions from the present for the past* The History of the Family (1996) Volume 1, Issue 1, 1996

[24] Charles Wetherell *Historical Social Network Analysis. International Review of Social History*, 43, pp 125-144. (1998)

[25] Douglas R. White, Paul Jorion *Kinship networks and discrete structure theory: Applications and implications* Social Networks, Vol. 18, No. 3. (1996), pp. 267-314

[26] Colin Wilder *Republic of Literature* https://sites.google.com/site/colinwilder/ Retrieved April 11, 2013