Antonio Maria Fiscarelli

Social network analysis for digital humanities

Challenges and use cases

1 Introduction

The field of digital humanities has grown rapidly in recent decades thanks to the greater availability of online digital sources, and new software and tools. Nevertheless, there are still some challenges that must be faced. During the same period, and due to the growing computing power and availability of online databases, network analysis has gained popularity: researchers from different fields have jumped on the network science bandwagon, and words such as "network" and "complexity" have become increasingly commonly used.

Network analysis can be used to model different systems such as biological networks,[1] the World Wide Web,[2] organizations, and societies. A social network can be described as a collection of "social actors" who are connected to each other if they form some sort of relationship. Social network analysis focuses on the relationships among these social actors and is an important addition to standard social and behavioral research, which is primarily concerned with the attributes of social units.[3] Not only is it important to acknowledge that social relationships are relevant, but also to understand how ties such as this work and how they relate to the many underlying social mechanisms governing these networks.

Social network analysis is one of the tools that have become particularly popular among humanities scholars. Even though social networks may seem to be a fairly recent invention, with the term calling to mind Facebook and other

1 Hawoong Jeong et al., "The Large-Scale Organization of Metabolic Networks," *Nature* 407, no. 6804 (2000): 651–54.

2 Réka Albert, Hawoong Jeong, and Albert-László Barabási, "Internet: Diameter of the World-Wide Web," *Nature* 401, no. 6749 (1999): 130–31.

3 Stanley Wasserman and Katherine Faust, *Social Network Analysis: Methods and Applications*, vol. 8 (Cambridge University Press, 1994).

Acknowledgments: I would like to thank Marten Düring (University of Luxembourg) for his insightful feedback on earlier drafts of this paper, and the Luxembourg National Research Fund (FNR) (10929115), who funded my research.

https://doi.org/10.1515/9783110723991-002

online platforms, they are in fact not limited to modern days.[4] For example, analysis of social networks has been used to model networks as diverse as the marriage and business relationships of the Medici family in fifteenth-century Florence,[5] the evolution of women's social movements in the nineteenth century,[6] the personal support network of Jewish refugees during the Second World War,[7] and visibility networks of Neolithic long barrows in the United Kingdom.[8]

The rest of this article is organized as follows: Social network analysis and some of its tools are introduced in Section 2. Section 3 presents an in-depth review of the latest historical network research. Finally, a use case drawn from my collaboration with a historian colleague is presented in Section 4.

1.1 Challenges in digital humanities

The first challenge in digital humanities is of a methodological nature.[9] On the one hand, and particularly in the use of network analysis, there is a risk that humanities research will limit itself to the "drawing of complicated graphs"[10] – yet the use of a certain method or digital tool should not be the main objective of research. On the other hand, some scholars may be hesitant to introduce digital tools into their research, fearing that these will take them out of the realm of history. It is therefore important to understand what digital tools can really offer in support of historical research.

4 Bonnie H Erickson, "Social Networks and History: A Review Essay," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 30, no. 3 (1997): 149–57.

5 John F. Padgett and Christopher K. Ansell, "Robust Action and the Rise of the Medici, 1400–1434," *American Journal of Sociology* 98, no. 6 (1993): 1259–319.

6 Naomi Rosenthal et al., "Social Movements and Network Analysis: A Case Study of Nineteenth-Century Women's Reform in New York State," *American Journal of Sociology* 90, no. 5 (1985): 1022–54.

7 Marten Düring, "The Dynamics of Helping Behavior for Jewish Refugees during the Second World War: The Importance of Brokerage," *Online Encyclopedia of Mass Violence*, 2016.

8 Tom Brughmans and Ulrik Brandes, "Visibility Network Patterns and Methods for Studying Visual Relational Phenomena in Archeology," *Frontiers in Digital Humanities* 4 (2017): 17.

9 James E. Dobson, *Critical Digital Humanities: The Search for a Methodology* (University of Illinois Press, 2019).

10 Claire Lemercier, "Formal Network Methods in History: Why and How?," in *Social Networks, Political Institutions, and Rural Societies, ed.* Georg Fertig (Turnhout: Brepols Publishers, 2015), 281–310.

The second challenge relates to the interdisciplinary nature of digital humanities. Humanities research can manifest in two forms. In the first case, scholars may show interest in a digital tool, start experimenting with it, and include it in their workflow. This approach could lead to the tool being used rather as a "black box" – i.e. given some input, the black box will produce a certain output, while everything in between is unknown. Therefore, it will not be possible to understand how the tool works, how to interpret the output, or how to recognize any potential bias inherent in that tool. In the second case, scholars may seek help, or a collaboration with an expert from another field, for example a computer scientist with a solid background in a specific method or tool. In this case, there is the risk that the humanities scholar will become a simple "data provider" for the model maker.[11] It is also essential to find a common vocabulary and be able to conciliate the two different perspectives in this scenario. Only if this is achieved can the two researchers start negotiating new forms of knowledge and successfully undertaking historical research together. In fact, my role in this project was to assess all these issues and ensure a fruitful collaboration between humanists and computer scientists.

Another issue relates to the data themselves. Historians nowadays have access to much larger amounts of data than their predecessors, whether from digitized classical sources (scans of books, digitized old photographs and recordings) or born-digital sources (websites, social networks). They can also access these sources at high speed and relatively low cost. For that reason, historians may be experiencing a paradigm shift, going from a scarcity to an abundance of sources,[12] while traditional methods used by historians may be failing to deal with such a volume of information. One example of such methods is close reading, which may fail in its purpose when the researcher is faced with very large collections of texts without the support of computer-based techniques. The easy accessibility of data comes with new questions too. Which sources have been digitized, which were discarded and what criteria were used to select those retained? It is also important to identify the origin of such sources. What was the provenance of the original sources? For born-digital sources, how were they generated?

Data storage has also changed with the advent of the digital era. The use of new technologies has made storing data far easier – a single hard drive can now store thousands of documents, and is cheap, small, and easy to transport. It can be easy to think that digital data will last forever. Unfortunately, data

11 Lemercier, "Formal Network Methods," 281–310.

12 Roy Rosenzweig, "Scarcity or Abundance? Preserving the Past in a Digital Era," *The American Historical Review* 108, no. 3 (2003): 735–62.

stored in digital form do not have any intrinsic meaning without the specific software or technology that can read them, and these technologies can become obsolete within a decade or even less. One may also think that digitally stored data is safe from aging. Indeed, unlike analog sources, digital data do not deteriorate. However, a single malfunction of the storing volume could render an entire data collection inaccessible and irretrievably lost.[13]

1.2 Project summary

The main objective of my doctoral project is to show how humanities research can benefit from network analysis by providing PhD students from other disciplines – such as history, psychology, linguistics, and archaeology – with the right tools to help them answer their historical questions and by adapting these tools to their research projects. In this way, a fruitful collaboration is sought, where each side can benefit from the other: humanities scholars gain a critical understanding of digital tools and their functionalities, while computer scientists find new use cases and applications, at the same time learning to appreciate the needs of humanists. Understanding each other's needs is crucial to the collaboration. Instead of two distinct groups with separate interests, I envision humanists and computer scientists joining forces to share their knowledge and expertise in order to tackle the new challenges that are emerging in digital humanities. Only with a common goal and a shared vision can this collaboration be effective and still worth the time and effort required.

2 Social network analysis

Historically, the first encounter with network analysis is seen in the "Seven Bridges of Königsberg" problem.[14] The then Prussian city of Königsberg was built on four main areas: the two sides of the Pregel River and two small islands, connected by seven bridges. The problem consisted in finding a route that reached all the areas of the city by crossing each bridge exactly once. Euler

13 Christine Barats, Valerie Schafer, and Andreas Fickers, "Fading Away . . . The Challenge of Sustainability in Digital Studies," *Digital Humanities Quarterly* 14, no. 3 (2020).
14 Norman Biggs, E. Keith Lloyd, and Robin J. Wilson, *Graph Theory, 1736–1936* (Oxford University Press, 1986).

modeled this problem using what we now call graph theory – representing the city areas as nodes and the bridges as edges connecting nodes – and proved it to be unfeasible: it has no solution.

2.1 Complex networks

Complex networks are those that exhibit unusual properties that make them different from other, simple networks. Some of these properties have played an important role in the development of the field of social network analysis and are worth examining.

2.1.1 Some definitions

A graph, or network (the terms are often used interchangeably), can be *directed* or *undirected*, depending on whether the direction of a connection is relevant. It can also be *weighted* or *unweighted*, where the weight represents cost, strength, or the importance of a connection.

The *degree* of a node v_i represents the number of incident edges it possesses – in other words, the number of the node's direct connections. In the case of a directed network, its *in-degree* and *out-degree* are also defined, and these refer to the number of ingoing or outgoing edges of a node.

The *average path length* of a network is defined as the average shortest path between any two nodes in that network. The *diameter* of a network is defined as its maximum shortest path. These two metrics represent how easily information can travel through a network.

The *clustering coefficient* of a network is defined as the average local clustering coefficient of each node in the network. The *local transitivity* of a node is the ratio of the triangles connected to the node and the triples centered on the node.[15] This metric is related to the concept of transitivity: given that v_i is connected to v_j, and v_j is connected to v_k, what are the odds that v_i is also connected to v_k?

15 Christine Barats, Valerie Schafer, and Andreas Fickers, "Fading Away . . . The Challenge of Sustainability in Digital Studies," *Digital Humanities Quarterly* 14, no. 3 (2020).

2.1.2 Small world phenomenon

The *small world* phenomenon was first identified during Milgram's experiments regarding social networks.[16] The experiments' objective was to send a letter from a source person in Nebraska to a target person in Massachusetts. The source person was asked to send the letter to whichever of their acquaintances was most likely to be connected to the target person, with the objective of reaching the target within as few steps as possible. Milgram noticed that source and target were, on average, between five and six people apart. This average path length figure was much lower than the number of people involved in the experiments, and became associated with the term "six degrees of separation."

Later on, Watts and Strogatz discovered that many real-world networks – such as the Western US power grid, the brain network of the nematode species *C. elegans*, and the World Wide Web – even though of different types, all had the same two properties: low average path length and a high clustering coefficient.[17] The network models known at that time – regular lattices and the random network model developed by Erdős and Rényi[18] – failed to capture these properties. In fact, regular lattices have high average path lengths and high clustering coefficients, while random networks have low average path lengths and low clustering coefficients. Watts and Strogatz proposed a model that, starting from a regular lattice, randomly rewires edges according to a certain probability p between zero and one. If this probability is properly chosen, the model can generate small-world networks. In fact, these networks still preserve the high clustering coefficient of regular lattices, but the rewiring of a few edges makes the distance between nodes much smaller.

2.1.3 Scale-free networks

Barabási and Albert noticed that, for many complex networks, the degree distribution does not follow a Poisson distribution with a peak around the mean value, but rather a power-law distribution.[19] This means that a very small number of

16 Stanley Milgram, "The Small World Problem," *Psychology Today* 2, no. 1 (1967): 60–7.
17 Duncan J. Watts and Steven H. Strogatz, "Collective Dynamics of 'Small-World' Networks," *Nature* 393, no. 6684 (1998): 440.
18 Paul Erdős and Alfréd Rényi, "On the Evolution of Random Graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5, no. 1 (1960): 17–60.
19 Albert-László Barabási and Réka Albert, "Emergence of Scaling in Random Networks," *Science* 286, no. 5439 (1999): 509–12.

nodes (or hubs) in the network have a very high degree – something that the Watts-Strogatz model was missing. Barabási and Albert realized that many real-world networks show a preferential attachment: nodes do not connect randomly but, rather, favor more "popular" nodes. For example, novice researchers in a collaboration network are more likely to aim to collaborate with researchers who are further on in their careers and already have many connections. Furthermore, complex networks are not static but instead grow in size. In fact, every year, new researchers start their careers and are added to the network. Barabási and Albert proposed a model that, based on these two mechanisms, can generate networks with a power-law degree distribution. The network starts with a fixed number of nodes. New nodes are then added and are connected to other nodes with a probability based on their degree. The networks generated with this model are called scale-free networks.

2.1.4 Emergence of communities in complex networks

Another important property of complex networks is their organization into communities. A community consists of a group of nodes that are highly connected to each other but loosely connected to the rest of the network.[20] For example, researchers in a collaboration network tend to connect to other researchers in the same field, resulting in the emergence of communities that represent similar research topics. Communities can be disjoint if nodes can only belong to a single community, or overlapping if they can belong to many.

2.1.5 The importance of weak ties

So far, we have seen that complex networks show high transitivity. Because of transitivity, nodes become highly connected to each other – and as a consequence, the network self-organizes into communities. We have also seen that, in a complex network, the average path length must be low. Therefore, it is necessary that some nodes act as "bridges" between communities. These connections are called *weak ties*. Sociology identifies two different kinds of ties in social networks: *strong ties* represent established interpersonal relationships, and are found in intracommunity connections; weak ties represent acquaintances, and are found in intercommunity connections. Granovetter, in his study, showed that

20 John Scott, "Social network analysis," *Sociology* 22, no. 1 (1988): 109–27.

people are more likely to find a new job through their acquaintances rather than through close friends.[21] This proved that weak ties are very important when it comes to the transmission of information within the network. While individuals in the same community can only share information that most of them probably already know, acquaintances can provide access to novel information.

2.2 Centrality metrics

Centrality metrics represent an important tool for the analysis of social networks. These metrics are defined on the nodes, and they rank nodes according to their position in a network.[22] *Degree centrality* measures the number of direct connections of a node and can be used to identify actors who are highly connected. *Betweenness centrality* is computed as the number of shortest paths between any two nodes in the network that go through a certain node. It measures to what extent an actor has control over the information flowing between other actors and can be used to identify those actors who occupy strategic positions in the network in terms of information exchange. *Closeness centrality* is computed as the average shortest path between a node and any other node in the network, and measures how long it will take for information to flow from one node to the rest of the network. The first person to experiment with centrality metrics was Bavelas, who showed that centrality measures were linked with group performance and that centrality metrics can help identify people with different roles in the network.[23]

2.3 Orbit analysis

Graphlets are small connected graphs with a size of between two and five nodes. Graphlet analysis is a useful tool for analyzing the global topological structure of networks and, locally, of a node's ego network. Figure 1 shows all the graphlets with up to four nodes. Some well-known examples are the "star" graphlet and the "triangle" graphlet. Some graphlets are characteristic of certain

21 Mark S. Granovetter, "The Strength of Weak Ties," *American Journal of Sociology* 78, no. 6 (1973): 1360–80.

22 Ulrik Brandes, "A Faster Algorithm for Betweenness Centrality," *Journal of Mathematical Sociology* 25, no. 2 (2001): 163–77.

23 Alex Bavelas, "A Mathematical Model for Group Structures," *Applied anthropology* 7, no. 3 (1948): 16–30; and Alex Bavelas, "Communication Patterns in Task-Oriented Groups," *The Journal of the Acoustical Society of America* 22, no. 6 (1950): 725–30.

types of network. For instance, the triangle is more likely to be found in social networks, due to high transitivity, while the star is more likely to be found in visibility networks. Graphlet counts, defined as the number of times that each graphlet appears in a network, can be used to characterize networks.

Nodes within a specific graphlet can have different roles. For example, in the star graphlet, one node can be identified as the center and the other three nodes as the leaves. Similarly, an orbit count can be defined as the number of times a node appears in each orbit, and can be used to identify groups of nodes that play different roles in the network. The orbit count for the central position of the "brokerage" graphlet can, for instance, be used to identify "mediator" nodes in collaboration networks.

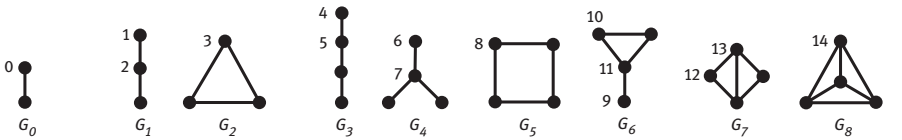

Fig. 1: Graphlets with up to four nodes, with their different orbits. 2020. © Antonio Fiscarelli.

2.4 Exponential random graph models

Exponential random graph models (ERGMs) are a family of statistical models that help us discover and understand the processes underlying network formation.[24] They have been used extensively in social network analysis and are popular in various fields such as sociology,[25] archaeology,[26] and history.[27] ERGMs

24 Carolyn J. Anderson, Stanley Wasserman, and Bradley Crouch, "A p* Primer: Logit Models for Social Networks," *Social Networks* 21, no. 1 (1999): 37–66; Garry Robins et al., "An Introduction to Exponential Random Graph (p*) Models for Social Networks," *Social Networks* 29, no. 2 (2007): 173–91; and Garry Robins et al., "Recent Developments in Exponential Random Graph (p*) Models for Social Networks," *Social Networks* 29, no. 2 (2007): 192–21.

25 Steven M. Goodreau, James A. Kitts, and Martina Morris, "Birds of a Feather, or Friend of a Friend? Using Exponential Random Graph Models to Investigate Adolescent Social Networks," *Demography* 46, no. 1 (2009): 103–25; and Thomas U. Grund and James A. Densley, "Ethnic Homophily and Triad Closure: Mapping Internal Gang Structure Using Exponential Random Graph Models," *Journal of Contemporary Criminal Justice* 31, no. 3 (2015): 354–70.

26 Tom Brughmans, Simon Keay, and Graeme Earl, "Introducing Exponential Random Graph Models for Visibility Networks," *Journal of Archaeological Science* 49 (2014): 442–54.

27 Abraham Breure and Raphael Heiko Heiberger, "Reconstructing Science Networks from the Past," *Journal of Historical Network Research* 3, no. 1 (2019): 92–117.

provide a model for networks that includes covariates – variables that relate to two or more nodes – which cannot be addressed using traditional methods. They can represent effects such as:

- *homophily*: the tendency of similar nodes – i.e. nodes having the same attributes – to form relationships.
- *mutuality*: the tendency of node B to form a relationship with node A, if node A is connected to node B.
- *triadic closure*: the tendency of node C to form a relationship with node A, if node A is connected to node B, and node B is connected to node C.

ERGMs also provide maximum-likelihood estimates for the parameters governing these effects. For example, they can estimate the increased likelihood of a tie existing between two nodes when these nodes have the same attributes. ERGMs also provide a "goodness-of-fit" test for the model, in order to verify whether the effects included in the model are sufficient to explain the structure of the observed network. Furthermore, they can simulate networks that match the probability distributions estimated by the model. In other words, they can be used to generate artificial networks that reflect the characteristics of the observed network.

3 Current trends in historical network analysis

There are already several examples of historians incorporating network analysis into their research. In this section I review some of their work, including how they translated historical questions into a social network analysis perspective, and identify what I consider to be the missed opportunities in these studies.

Breure and Heiberger, in their study, argue that eponyms serve as a proxy for contact and are a promising way to explore historical relationships between natural scientists.[28] Eponyms are used in taxonomy when an author describes a new species for which they use the name of a person – usually a field collector or colleague.

Breure and Heiberger tested this hypothesis on the community of malacologists (i.e. zoologists studying mollusks) in the nineteenth century, analyzing the recorded activity of malacological authors between 1850 and 1870. The dataset used contained authors' information such as age and home country, as well as performance measures like their numbers of publications, pages, coauthored publications, and coauthors. Each connection between authors was

28 Breure and Heiberger, "Reconstructing Science Networks," 92–117.

classified as an eponym, an exchange of material, or a coauthorship. Therefore, these authors had, effectively, built a collaboration network, in particular a multiplex network, where nodes interact within different layers (depending on the type of interaction) but there is no interaction between the different layers themselves. This network, consisting of 476 nodes and 1,822 edges, can be considered of medium size. The authors in the network were ranked according to their number of publications, and elite authors were identified as those who contributed to 80 percent of the total publications.

Breure and Heiberger noticed that few authors published a large number of papers, something that has been widely recognized in bibliometrics. They also identified two heavily linked communities that represented authors dealing with recent shells and those dealing with fossil (paleontological) shells. They manually assigned authors to one of the two communities, depending on their research interests. It would have been interesting to use a community detection algorithm to compare the communities found with the ones identified by the authors, using metrics such as normalized mutual information[29] or adjusted randomized index[30] to quantify the agreement of the result, and thus assess any bias in the manual assignment.

The authors used ERGMs to find out what effects had shaped the network of collaboration and found that authors from the same country were more likely to connect with each other, and that higher publication numbers increased the odds of a tie between authors. They also discuss how eponyms could result in a collaboration between authors, but this hypothesis was not tested, even though ERGMs offer the possibility of testing whether a tie in one layer increases the odds of a tie in a different layer.

Fernandez Riva, in his work, introduced a new method for analyzing shared manuscript transmission of medieval German texts, based on network analysis.[31] Medieval manuscripts contain several texts that were brought together according to certain criteria – both cultural (common genre) and practical (availability, size, etc.) – rather than being randomly grouped. Fernandez Riva modeled the transmission of shared manuscripts as a network, where nodes represent texts

29 Leon Danon et al., "Comparing Community Structure Identification," *Journal of Statistical Mechanics: Theory and Experiment* 2005, P09008; and Zhao Yang, René Algesheimer, and Claudio J. Tessone, "A Comparative Analysis of Community Detection Algorithms on Artificial Networks," *Scientific Reports* 6 (2016): 30750.

30 Lawrence Hubert and Phipps Arabie, "Comparing Partitions," *Journal of Classification* 2, no. 1 (1985): 193–218.

31 Gustavo Fernandez Riva, "Network Analysis of Medieval Manuscript Transmission," *Journal of Historical Network Research* 3 (2019): 30–49.

that are deemed connected if they appear in the same manuscript, and a weight is assigned if texts appear together in more than one manuscript. He does not mention the size of the network, however he specifies that the largest connected component of the network included 76 percent of the nodes, while several smaller components (of two to eight nodes) included 6 percent of the nodes, and the remaining 18 percent consisted of isolated nodes. Fernandez Riva decided to name these three different parts of the network "Continent," "Archipelagos," and "Islands." He proceeded by applying a community detection algorithm on the largest component to identify communities, although the algorithm used is not mentioned. Since the nodes had no attribute data – such as genre, time, or location – available, the author manually inspected the outcome of the algorithm to verify whether any of these characteristics correlated with the communities found, and came to the conclusion that there was a high overlap between communities, even for different genres. He used eigenvector centrality to identify texts that tended to appear in large collections, and betweenness centrality to identify texts that connected different communities in the network and fitted into more than one genre. These metrics helped him identify texts that occupied strategic positions in the network, something that would have been impossible by human inspection. Although the author does not really provide statistical methods for his analysis of the network of interest – instead limiting his work to the visualization of the network and the computation of centrality metrics – it must be recognized that the data available to him were rather limited.

Valleriani et al. analyzed the emergence of epistemic communities during the early modern period.[32] They worked on a corpus of printed cosmology textbooks used at European universities, dividing each book into several text parts, representing "atoms" of knowledge. The authors built a directed, weighted, multilayer network where nodes represented books that were connected to each other, on different layers, if they contained text parts that reoccurred in time (i.e. if they contained the same text, adaptations or translations of the same text, commentaries on the same text, or commentaries on the same adaptation), for a total of five layers. The network was a directed one, with the directionality being chronological, from older to more recent occurrences. The weight of connections, on the other hand, was given by the number of text parts that reoccurred in two different books. The corpus contained 563 text parts, but the authors decided to consider only those parts reoccurring at least once, and with at least

32 Matteo Valleriani et al., "The Emergence of Epistemic Communities in the 'Sphaera' Corpus: Mechanisms of Knowledge Evolution," *Journal of Historical Network Research* 3 (2019): 50–91.

one year between reoccurrences. Therefore the network, which can be considered of small-to-medium size, consisted of 239 text parts and 1,625 reoccurrences. The authors also analyzed the aggregated graph, which included the same set of nodes – two nodes were deemed connected if they were connected in any of the five layers. The authors performed a longitudinal analysis by first looking at the age distribution of connections for each layer of the network – computed as the difference between years of publication of the two text parts at the ends of each connection – and found substantial differences between layers. They then looked at the various connected components of the network in order to identify the different epistemic communities. Using a series of plots, they analyzed the distribution of nodes' out-degrees, normalized by the publication date of the text. For each plot, the visualization was further enhanced with different colors representing the nodes' attributes such as in-degree, publication place, book format, and network layer. The analysis is followed by an in-depth interpretation of the results, and discussion on the emergence and evolution of the different families of editions. Again, the methodology provided is based more on data visualization than statistical analysis or advanced modeling techniques. Cline, in her work, has used social network analysis to study political life in Athens between the 460s and 450s BC.[33] She builds three increasingly broad social networks using selected biographies from Plutarch's *Lives*, from which she retrieves all actors and their interrelationships. The first network uses Plutarch's "Life of Pericles" and consists of 54 actors and 79 ties, which essentially equates to Plutarch's ego network. She then enlarges this by adding actors from "Life of Alcibiades." This second version of Athens' social network contains 106 nodes and 145 connections. Lastly, she includes "Life of Cimon" and "Life of Nicias," for a total of 133 nodes and 191 ties across this largest network, formed from all four biographies' actors. These networks are all of a small size, undirected, and unweighted. The author says she is working with a multiplex network, since ties between actors are of different natures (family, work, friendship), even though there is no distinction between these ties in the analysis. Her objective is to demonstrate that the social network of Athens' political life was a small world. Her argument is that democratic institutions in Athens enabled people belonging to different circles and social classes to meet, hence favoring innovation and the diffusion of new ideas. From a network perspective, this would reflect in Athens' social network having a low average path length, high level of transitivity and a core-periphery structure where degree distribution follows a power law, with few

33 Diane Harris Cline, "Athens as a Small World," *Journal of Historical Network Research* 4 (2020): 36–56.

highly connected nodes and most nodes having a low degree. Indeed, she computes transitivity, average path length, and diameter for all the networks, and compares them with the same quantities computed on a random network of the same size. All these measures confirm that Athens at the time was indeed a small world. For the core-periphery structure, Cline computes the degree distribution but does not perform any statistical tests to verify whether a power law is the best fit. She also computes betweenness for each actor to confirm that women tend to occupy central positions in the network, connecting different families via marriage. For this work, information such as gender, family, and social status was available. Therefore, it would have been interesting to test whether any of these attributes had an influence on the network of connections.

Schauf and Escobar Varela[34] used network analysis techniques to identify characters who play structural roles in the Javanese *wayang kulit* incarnation of the *Mahabharata* epic, which involves representations of the series of stories – here called *lakon* – from the epic. They build a weighted, undirected co-occurrence network, where nodes represent the characters of the epic and these characters are deemed connected if they are mentioned in the same scene of any story. Weights indicate how many times two characters appear in the same scene. Each node is enriched with several attributes such as characters' tribe affiliation, origin, species, and gender. The authors also build two different null models that preserve, on average, the degree distribution of nodes. They compute betweenness centrality and closeness centrality for each character in the empirical network, as well as in the two null models. In this way, it is possible to identify outliers whose centrality values are significantly higher or lower than expected, i.e. compared to the same quantity computed in the null models. For example, the authors find that female characters, despite being few in number and appearing relatively infrequently, seem to dominate the top ranks for betweenness. They also propose a variation of these centrality metrics that is based on the attributes of nodes. For example, the inter-faction betweenness centrality is used to identify those characters who act as "bridges" within their tribe, while the faction-world betweenness centrality identifies characters who act as bridges between their tribe and the rest of the network.

One of the challenges that emerges from historical network research working with historical data is dealing with missing and incomplete data.[35] Networked data have to be extracted from sources such as books, bibliographies, and diaries that were originally analog and only digitized later, if needed. These

34 Andrew Schauf and Miguel Escobar Varela, "Searching for Hidden Bridges in Co-Occurrence Networks from Javanese Wayang Kulit," *Journal of Historical Network Research* 2 (2018): 26–52.
35 Erickson, "Social Networks and History."

sources are often incomplete or do not provide enough information to build the network of interest. Additionally, missing data in network research are more critical than in social and behavioral research. Even a small portion of missing data can be problematic if those data are related to crucial nodes (see *hubs* in Section 2.1.3) or ties (see *weak ties* in Section 2.1.5) This is also in contrast to historical research working with born-digital data, such as online databases or data scraped from social networks, where data are rather abundant.

4 Use case: Gender and ethnic collaboration patterns in a temporal co-authorship network

Sytze Van Herck is one of the PhD students at the University of Luxembourg's doctoral training unit in digital history and hermeneutics. Her main research interests are intersectionality and gender within the history of computing – and her work examines occupational segregation, working conditions, and gender stereotypes in advertising from the 1930s until the end of the 1980s. Sytze and I applied social network analysis techniques to analyze the gender and ethnicity gap in computer science research.[36] During the last few decades many bibliographic databases containing the publication records of scientists from different fields have been published online. Starting from these records, a collaboration network can be built where nodes represent authors, and authors are deemed connected if they have coauthored one or more papers together. This network of scientists can provide many insights into collaboration patterns in the academic community.

The dataset that Sytze and I used for the use case discussed here was one derived from a snapshot of the DBLP bibliographic database taken on 17 September 2015 and publicly available.[37] It contains 112,456 papers, written by 126,094 authors and published at 81 different computer science conferences between 1960 and 2015. The dataset includes author gender, which was generated by the Genderize API based on the first forename of an author.[38] For ethnicity data we

36 Sytze Van Herck and Antonio Maria Fiscarelli, "Mind the Gap Gender and Computer Science Conferences," in *This changes everything – ICT and Climate Change: What can we do? IFIP International Conference on Human Choice and Computers,* ed. David Kreps et al. (Cham: Springer Nature Switzerland, 2018), 232–49.

37 Agarwal Swati et al., "*DBLP Records and Entries for Key Computer Science Conferences,*" Mendeley Data, V1, 2016.

38 Genderize API, accessed April 21, 2021, https://genderize.io.

decided to use the R package called *wru* that uses the algorithm implemented by Kosuke and Kabir to predict ethnicity based on last name and gender.[39]

Our research was driven by the following questions:

- Do minorities in computer science demonstrate different collaboration patterns?
- As we saw in Section 2.1.1, metrics such as clustering coefficient, average path length, and diameter can characterize entire networks. A large clustering coefficient can be used to identify densely connected networks with high transitivity, while low average path length and diameter can identify networks in which information flows faster. For this reason, we decided to extract male and female subnetworks from the dataset, as well as networks of white researchers and researchers of color, by considering only the nodes with the selected attribute and the connections between those nodes. We then computed clustering coefficient, average path length, and diameter on these networks and compared the results. We found that the female researchers had a more close-knit network than the male researcher network – and that white researchers, even though they were not a minority, showed a similar behavior.
- Do minorities in computer science struggle to be successful?
- The metrics commonly used to quantify the success or popularity of a researcher are based on the numbers of their publications and citations. We decided, instead, to use network metrics (presented in Section 2.2) that were based on the position that researchers occupied in the coauthorship network and metrics based on a researcher's ego network structure. We computed some local network metrics such as betweenness centrality, closeness centrality, local clustering coefficient, and degree centrality, and then ranked male and female researchers, as well as white researchers and researchers of color. We found that female researchers generally scored lower than their male counterparts in terms of network connections, and had more closely knit networks. However, those ranked at the top obtained better results. Researchers of color, who were mostly Asian researchers, occupied more strategic and central positions in collaborations, outperforming white researchers.
- Do minorities play different roles in the network?
- To answer this question we used orbit analysis (discussed in Section 2.3) to compute the average orbit count for female and male researchers, as well as for white researchers and researchers of color, and compared the results. We

39 Kosuke Imai and Kabir Khanna, "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records," *Political Analysis* 24, no. 2 (2016): 263–72.

found that male researchers dominated central roles, corresponding for example to the central orbit in the star graphlet, while female researchers tended to occupy the peripheral positions. In particular, in the brokerage graphlet, male researchers more often occupied a brokerage position, corresponding to the central orbit of this graphlet, while a pair of female researchers and an individual female researcher were more likely to be found in the peripheral orbits of the same graphlet – implying the male researcher played a mediating role between these female researchers.

- Does the minority bias become mitigated over time?
- We built a temporal version of the coauthorship network and answered the same questions to see if there were any changes over time. Firstly, we found that the size of minority groups had expanded over time, with their intragroup homophily increasing even faster. Female researchers performed better at higher ranks only during specific periods, such as in the middle of the 1980s and toward the end of the 1990s. The trend for ethnicity, on the other hand, inverted over time: researchers of color, mostly Asian, occupied more central positions until the mid-1990s, while they have become more closely knit in recent years. In the orbit analysis we found that gender differences had narrowed over time, while we observed a complete inversion of the trend for ethnicity.

4.1 Reflections and challenges

The aim of this collaboration was to build a bridge between the very different disciplines of humanities and computer science. We faced several challenges during this work. The first was related to the algorithmic bias associated with the gender and ethnicity prediction algorithms. The gender prediction was based on the given name (or forename) of an author. This was a generalization that was necessary given the large number of authors and the limited personal information available. First of all, we assumed that gender is binary, rather than more complex. Secondly, the same given name may be more commonly associated with being a male or female name depending on the country of origin. For example, the name "Andrea" is commonly feminine, while it is widely used as masculine in Italy. Additionally, the gender identity of a person may not match their biological sex.

The ethnicity prediction algorithm, on the other hand, is based on the family name (surname) and gender of an author. This is also a generalization, since a person's cultural identity may be different from their ancestry (or indeed from their spouse's ancestry where family names are changed on marriage). For example, many second- and third-generation American citizens have Italian surnames due to their Italian ancestry, while embracing an American identity. We also noticed

that the gender prediction algorithm was less accurate for ethnic minorities. We therefore decided to build two separate networks for our analysis: one containing all authors whose gender prediction had at least 99 percent accuracy (i.e. a 99 percent likelihood of being correctly assigned as male or female), and another containing all authors whose ethnicity prediction score had at least 50 percent accuracy (i.e. 50 percent likelihood of belonging to a certain ethnicity versus all other ethnicities).

The fact that the algorithms do not have 100 percent accuracy shows that the use of digital tools does not remove bias. Algorithms contain an intrinsic bias because they are designed by humans, and researchers also introduce bias when choosing a certain algorithm.

5 Conclusion

The main objective of this project was to show how humanities research can benefit from network analysis, by providing PhD students from different fields with the right tools to help answer their historical questions, and adapting these tools to their research projects. In this way, a fruitful collaboration – where both sides can benefit from each other – may be sought: humanities scholars gain a critical understanding of digital tools and their functionalities, while computer scientists find new use cases and applications, at the same time learning to understand the needs of humanists. Understanding each other's needs is crucial for such collaborations. Instead of two distinct groups with separate interests, I envision humanists and computer scientists joining forces and sharing their knowledge and expertise in order to tackle the new challenges that are emerging in digital humanities. Only with a common goal and a shared vision can this collaboration be effective and still worth the time and effort required.

This article describes how I reviewed the latest historical network research in order to assess the current practices of historians using network-based methods, and discusses some of the challenges faced in digital humanities. As part of this work I translated historical problems for computer science peers and explained the basics of social network analysis to historians. I have also presented a use case here, drawn from my collaboration with a historian colleague, showing how social network analysis can be used to answer historical research questions. In particular, I presented our joint research questions and the tools we used to answer them. Finally, I reflected on the challenges we encountered during our joint work, such as the generalizations that we made in order to model our scenario and the algorithm criticism regarding the gender and ethnicity predictions.

References

Albert, Réka, Hawoong Jeong, and Albert-László Barabási. "Internet: Diameter of the World-Wide Web." *Nature* 401, 6749 (1999): 130–31.

Anderson, Carolyn J., Stanley Wasserman, and Bradley Crouch. "A p* Primer: Logit Models for Social Networks." *Social Networks* 21, 1 (1999): 37–66.

Barabási, Albert-László, and Réka Albert. "Emergence of Scaling in Random Networks." *Science* 286, 5439 (1999): 509–12.

Barats, Christine, Valerie Schafer, and Andreas Fickers. "Fading Away . . . The Challenge of Sustainability in Digital Studies." *Digital Humanities Quarterly* 14, 3 (2020). Accessed November 18, 2021. http://www.digitalhumanities.org/dhq/vol/14/3/000484/000484.html.

Barrat, Alain, Marc Barthelemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge: Cambridge University Press, 2008.

Bavelas, Alex. "A Mathematical Model for Group Structures." *Applied Anthropology* 7, 3 (1948): 16–30.

Bavelas, Alex. "Communication Patterns in Task-Oriented Groups." *The Journal of the Acoustical Society of America* 22, 6 (1950): 725–30.

Biggs, Norman, E. Keith Lloyd, and Robin J. Wilson. *Graph Theory, 1736–1936*. Oxford: Oxford University Press, 1986.

Brandes, Ulrik. "A Faster Algorithm for Betweenness Centrality." *Journal of Mathematical Sociology* 25, 2 (2001): 163–77.

Breure, Abraham S.H., and Raphael Heiko Heiberger. "Reconstructing Science Networks from the Past." *Journal of Historical Network Research* 3, 1 (2019): 92–117.

Brughmans, Tom, and Ulrik Brandes. "Visibility Network Patterns and Methods for Studying Visual Relational Phenomena in Archeology." *Frontiers in Digital Humanities* 4 (2017): 17.

Brughmans, Tom, Simon Keay, and Graeme Earl. "Introducing Exponential Random Graph Models for Visibility Networks." *Journal of Archaeological Science* 49 (2014): 442–54.

Cline, Diane Harris. "Athens as a Small World." *Journal of Historical Network Research* 4 (2020): 36–56.

Danon, Leon et al. "Comparing Community Structure Identification." *Journal of Statistical Mechanics: Theory and Experiment* 2005, 09 (2005): P09008.

Dobson, James E. *Critical Digital Humanities: The Search for a Methodology*. Chicago: University of Illinois Press, 2019.

Düring, Marten. "The Dynamics of Helping Behavior for Jewish Refugees during the Second World War: The Importance of Brokerage." *Online Encyclopedia of Mass Violence*, 2016. Accessed July 21, 2021. https://www.sciencespo.fr/mass-violence-war-massacre-resistance/en/document/dynamics-helping-behaviour-jewish-fugitives-during-second-world-war-importance-brokerage-se.

Erdős, Paul, and Alfréd Rényi. "On the Evolution of Random Graphs." *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5, 1 (1960): 17–60.

Erickson, Bonnie H. "Social Networks and History: A Review Essay." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 30, 3 (1997): 149–57.

Fernandez Riva, Gustavo. "Network Analysis of Medieval Manuscript Transmission." *Journal of Historical Network Research* 3 (2019): 30–49.

Genderize API. "API to predict the gender of a person given their name." Accessed April 21, 2021. https://genderize.io.

Goodreau, Steven M., James A. Kitts, and Martina Morris. "Birds of a Feather, or Friend of a Friend? Using Exponential Random Graph Models to Investigate Adolescent Social Networks." *Demography* 46, 1 (2009): 103–25.

Granovetter, Mark S. "The Strength of Weak Ties." *American Journal of Sociology* 78 (1973): 1360–380.

Grund, Thomas U., and James A. Densley. "Ethnic Homophily and Triad Closure: Mapping Internal Gang Structure Using Exponential Random Graph Models." *Journal of Contemporary Criminal Justice* 31, 3 (2015): 354–70.

Hubert, Lawrence, and Phipps Arabie. "Comparing Partitions." *Journal of Classification* 2, 1 (1985): 193–218.

Imai, Kosuke, and Kabir Khanna. "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records." *Political Analysis* 24, 2 (2016): 263–72.

Jeong, Hawoong et al. "The Large-Scale Organization of Metabolic Networks." *Nature* 407, 6804 (2000): 651–54.

Lemercier, Claire. "Formal Network Methods in History: Why and How?" In *Social Networks, Political Institutions, and Rural Societies*, edited by Georg Fertig, 281–310. Turnhout: Brepols, 2015.

Milgram, Stanley. "The Small World Problem." *Psychology Today* 2, 1 (1967): 60–7.

Padgett, John F., and Christopher K. Ansell. "Robust Action and the Rise of the Medici, 1400–1434." *American Journal of Sociology* 98, 6 (1993): 1259–319.

Robins, Garry et al. "An Introduction to Exponential Random Graph (p*) Models for Social Networks." *Social Networks* 29, 2 (2007): 173–91.

Robins, Garry et al. "Recent Developments in Exponential Random Graph (p*) Models for Social Networks." *Social Networks* 29, 2 (2007): 192–215.

Rosenthal, Naomi et al. "Social Movements and Network Analysis: A Case Study of Nineteenth-Century Women's Reform in New York State." *American Journal of Sociology* 90, 5 (1985): 1022–54.

Rosenzweig, Roy. "Scarcity or Abundance? Preserving the Past in a Digital Era." *The American Historical Review* 108, 3 (2003): 735–62.

Schauf, Andrew, and Miguel Escobar Varela. "Searching for Hidden Bridges in Co-Occurrence Networks from Javanese Wayang Kulit." *Journal of Historical Network Research* 2 (2018): 26–52.

Scott, John. "Social Network Analysis." *Sociology* 22, 1 (1988): 109–27.

Swati, Agarwal, et al. *DBLP Records and Entries for Key Computer Science Conferences*. Accessed April 21, 2021. https://data.mendeley.com/datasets/3p9w84t5mr/.

Valleriani, Matteo et al. "The Emergence of Epistemic Communities in the 'Sphaera' Corpus: Mechanisms of Knowledge Evolution." *Journal of Historical Network Research* 3 (2019): 50–91.

Van Herck, Sytze, and Antonio Maria Fiscarelli. "Mind the Gap Gender and Computer Science Conferences." In *This changes everything – ICT and Climate Change: What can we do? IFIP International Conference on Human Choice and Computers*, edited by David Kreps et al., 232–49. Cham: Springer Nature Switzerland, 2018.

Wasserman, Stanley, and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press, 1994.

Watts, Duncan J., and Steven H. Strogatz. "Collective Dynamics of 'Small-World' Networks." *Nature* 393, 6684 (1998): 440.

Yang, Zhao, René Algesheimer, and Claudio J. Tessone. "A Comparative Analysis of Community Detection Algorithms on Artificial Networks." *Scientific Reports* 6, 1 (2016): 30750.