Introduction to Social Network Analysis

Davor Salihović

One of the primary goals of the DiplomatiCon project is to examine the development, maintenance, and structures of the various formal and informal networks that existed across the medieval Mediterranean, as well as their interrelations. While it is well known that the "Christian" west and north and the "Islamicate" east and south maintained contact throughout the medieval period, the project seeks to investigate the processes through which these contacts emerged, how these processes were governed by spatial and structural characteristics of the networks and by the personal characeristics of individuals, and to what extent they were influenced by the varying intensity of contact across different networks. An investigation into these dynamics over time allows not only for a mapping of the local forces through which cross-Mediterranean networks emerged, but also for a formal testing of our current understanding of these interactions and the introduction of a pioneering formal description of phenomena that have previously been understood only descriptively.

As such, the project lies at the intersection of history and network science, employing tools from graph theory and network analysis to model and investigate these phenomena. Network analysis is based on modeling connections and interactions between entities, conceptualizing them as points (called "nodes") and the lines that connect them (called "links" or, in mathematical terms, "vertices" and "edges"). While network science and graph theory both develop the metrics and tools needed to model, analyse, and understand graphs theoretically—as mathematical entities and models—findings from these fields have been successfully applied to the analysis of empirical networks across a wide range of disciplines. Network analysis has now been applied to phenomena as diverse as animal behavior,[1] scientific collaboration,[2] cultural culinary preferences,[3] genetics and disease,[4] protein interactions,[5] to name a few, and—perhaps most successfully—social interactions.

Like any other mathematical model of social (or other) phenomena, graphs aim to introduce mathematical rigour and formal language into the analysis of data on (historical) social interactions. Networks—and the various metrics and tools developed for their analysis—enable us to formalise theories, assumptions, and data; quantify the structures and dynamics characteristic of networks of interest; apply statistical inference; test established theories and propose new ones; and describe findings in a precise, interdisciplinary language. As with other models, their purpose is to support the human mind—remarkably limited in its ability to intuitively grasp the intricacies of complex systems, having evolved to function in the "Middle World"[6]—in disentangling the underlying processes that govern correlations between diverse dynamics, be it in the emergence of consciousness or the success of a job hunt. Today's computational power has further enhanced mathematical modeling, enabling rapid calculation and quantification, as well as accelerating the development of new formal methods. All this allows us to move beyond the limits of human intuition in conceptualizing and analysing complex phenomena such as (social) interactions and connections.

The last point seems especially relevant to the study of history, where formalism is often lacking and "common sense" heuristics have long served as the cornerstone of methodology—both in general and in the study of social networks in particular. Although social network analysis has gained popularity among historians in recent years—and some studies do employ formalism and sophisticated methods—network analysis in historical scholarship is still often limited to

visualisation, with little engagement in the deeper analytical or theoretical potential of the approach. While network visualisations can produce striking imagery—perhaps by catering to the strengths of human perception or conveying an illusion of analytical sophistication, or both—they often contribute little, and usually nothing, to our understanding of the structural forces and dynamic processes that give rise to networks, govern the formation and maintenance of links, or shape percolation, flow, and other network phenomena. The types of visualisations themselves should be chosen based on concrete criteria—criteria that are often overlooked. Visualising networks is a rigorous practice in its own right, governed by precise mathematical rules that determine how nodes and links are plotted. Arbitrarily selecting a visualisation method can mislead analysis more often than it offers meaningful insight.

Figure 1 displays an array of networks—some simulated, some empirical; some generated through similar dynamics, some based on different underlaying forces. In most cases—particularly as networks grow large, as seen in networks E or even B—the human eye and mind are incapable of discerning even the most basic structural characteristics, let alone calculating likelihood functions for parameters correlated with the probability of nodes sharing links. Without comprehensive training, one would struggle to identify which networks are computer-generated and which are derived from empirical data—if such a distinction is possible at all. While network A, a random graph generated via the G(N, p) Erdős-Rényi model[7] appears notably different from, for instance, graphs C or F, graphs B and C are also simulated, but although they appear rather different from each other, they were generated using the same dynamics of preferential attachment and growth in the Barabási-Albert model.[8] There, acquiring links over time is related to one's existing degree, governed by the preferential attachment parameter (approximated in the probability that a node with degree k is chosen: $\Pi(k) \sim k^{\alpha}$). This model more closely approximates real-world networks than random graphs, most notably by producing a power-law degree distribution ($P(k) \sim k^{-\gamma}$, the probability of a node interacting with k other nodes decays as a power law, meaning that the probability is relatively high for low degrees and much lower for high degrees), a common feature in many empirical networks. This is why graph C, for instance, resembles the three graphs in the lower panels—all of which are based on empirical data. Graph D shows the network of accusations in a fourtheenth-century inquisitorial trial.[9] Graph E depicts human protein-protein interactions (simplified for visualisation purposes),[10] while graph F shows the spatial network of correspondence across the late-medieval Mediterraenan, drawing on DiplomatiCon data on the flow of correspondence between locations. Network F is also shown in Figure 2, but under a different layout. There, node placement is based on geographic coordinates; in Figure 1F, the positions are determined using the popular Fruchterman–Reingold layout algorithm, which treats nodes as repelling particles (like particles with the electric charge of the same sign) connected by spring-like links, leading to a structure based on network topology rather than geography.

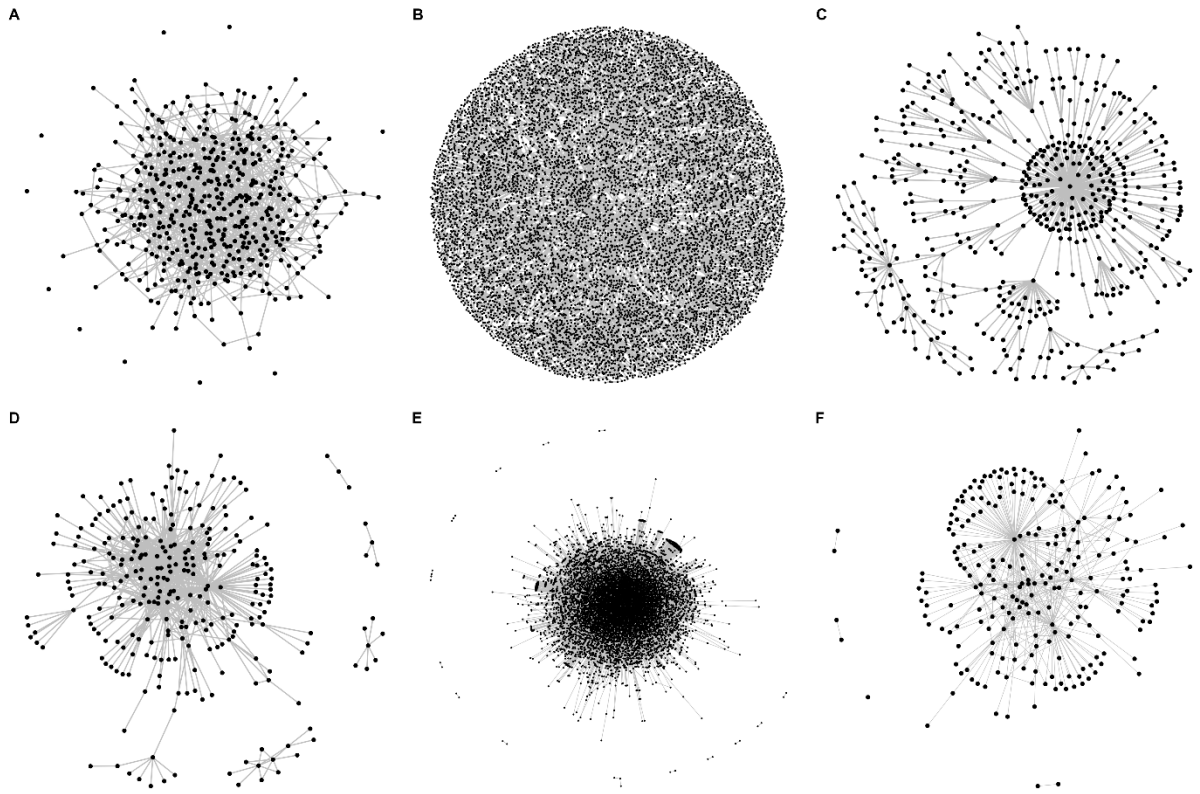

Figure 1.

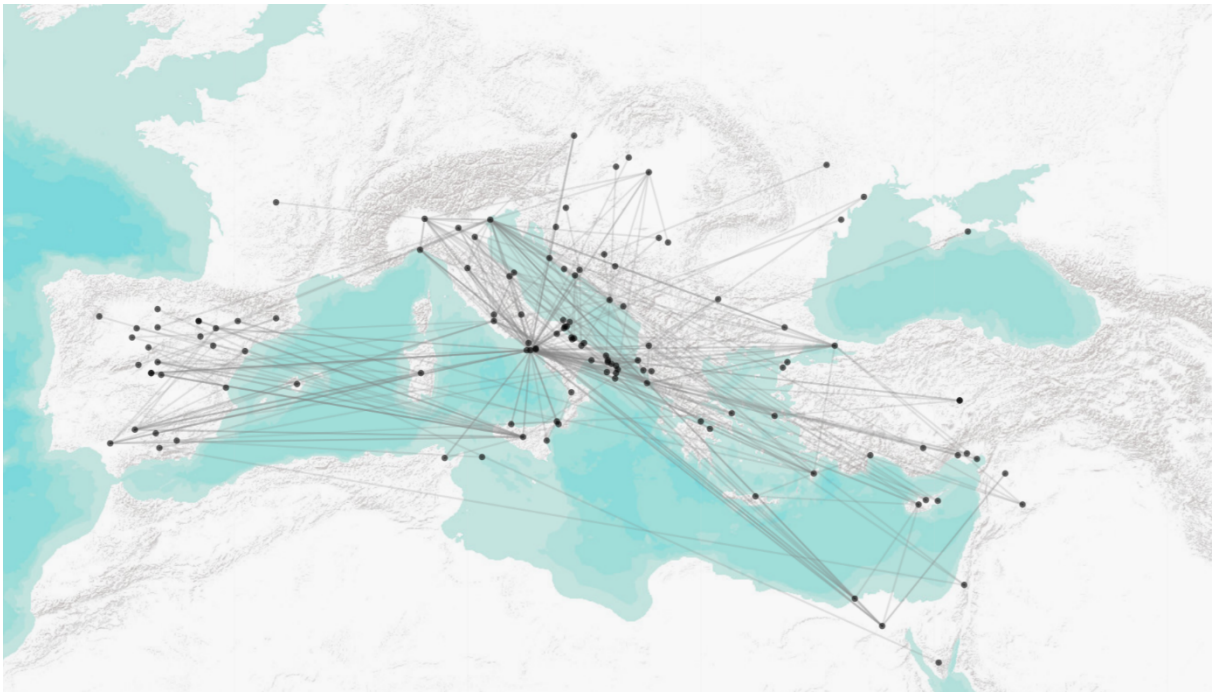

Figure 2.

Graphs are mathematical structures that happen to suit the modeling of social relations. As such, they are most succesffuly represented by a matrix—typically the adjacency matrix—which uses

ones and zeros (or other values, in the case of weighted networks) to indicate which nodes are connected. By convention, the matrix is read from rows to columns; that is, in the case of a non-symmetric matrix representing directed links, a one at position (i, j) indicates that row node i sends a link to column node j:

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \quad (1)$$

It is for this reason that matrix operations and linear algebra are central to the analysis of social (or other types of) networks, as the metrics and statistics that describe various structural features of networks are derived from them. Note, for instance, that a matrix is simply composed of vectors—whether read row-wise or column-wise. These vectors are no different from the two-element or three-element vectors we are used to working with, such as those representing a point in two-dimensional space—for example, vector $\vec{v} = [3, 5]$—or in three-dimensional space, such as $\vec{a} = [3, 5, 1]$. The only difference is that a vector from an adjacency matrix typically has more elements. It follows, then, that an n-element vector describes a point in an n-dimensional space, which we—with our "middle-world" brains—cannot intuitively visualise if $n > 3$. It also follows that a person's profile, in terms of to whom they send their links—or from whom they receive links—is just a vector (or a point in a multidimensional space). It is little wonder, then, that one can formulate a measure of similarity between two such profiles—or between two people—in terms of their outgoing or incoming ties, based on how their vectors relate to each other in this space. A well-known operation involving vectors is calculating the cosine of the angle (and therefore the angle itself) between two vectors. In the two-dimensional case, for two vectors $\vec{a}$ and $\vec{b}$, the equation looks as follows:

$$\cos\alpha = \frac{a_1 b_1 + a_2 b_2}{|\vec{a}| \cdot |\vec{b}|} = \frac{a_1 b_1 + a_2 b_2}{\sqrt{a_1^2 + a_2^2} \cdot \sqrt{b_1^2 + b_2^2}}$$

This, as one might guess, generalises to n-dimensional vectors **A** and **B** as:

$$\cos\alpha = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

The larger the cosine—and the smaller the angle between the vectors—the more similar the vectors are. In terms of social relations, the profiles of two people, based on whom they connect to (or who connects to them), are more similar when the cosine between their corresponding vectors is larger.

In our directed network, displayed in matrix **A** (1)—which we can imagine as a network of correspondence—the cosine of the angles between the five column-vectors (representing incoming ties, or incoming correspondence) are as follows:

$$\begin{bmatrix} 1 & 0.41 & 0.41 & 0.50 & 0.41 \\ 0.41 & 1 & 1 & 0.82 & 0.67 \\ 0.41 & 1 & 1 & 0.82 & 0.67 \\ 0.50 & 0.82 & 0.82 & 1 & 0.41 \\ 0.41 & 0.67 & 0.67 & 0.41 & 1 \end{bmatrix}$$

The higher the cosine in position (i, j)—ranging between 0 and 1—the greater the similarity in incoming ties between the two nodes; that is, the more similar the sets of people who send correspondence to them. This is why all diagonal entries equal 1—any node *i* has an identical profile to itself.

Similar metrics—from the simplest to more sophisticated ones—are likewise based on linear algebra and matrix operations. One of the fundamental node-level metrics in network analysis is a node's degree: the number of ties it is incident to, that is, the number of ties it sends (out-degree) or receives (in-degree). This corresponds to the sum of the values in the node's row (for out-degree) or column (for in-degree) vectors of the adjacency matrix, $\sum_{j=1}^{n} A_{ij}$ and $\sum_{i=1}^{n} A_{ij}$. A measure of reciprocity—the proportion of ties in a directed network that are reciprocated by alters—is, for instance, operationalised as the sum of the element-wise product of the adjacency matrix and its transpose, divided by the total number of ties in the network. A measure of similarity between two nodes based on their shared neighbours—nodes they send ties to or receive ties from—that accounts for the popularity of those neighbors by penalising high-degree nodes—is the inverse log-weighted similarity. This measure assigns greater weight to less popular (lower-degree) common neighbors. It is defined as the sum of the inverse logarithms of the degrees of all common neighbors of nodes *i* and *j*: $S_{ij} = \sum_{n} \frac{1}{\log(k_n)}$, where *n* are common neighbours and *k* their degree.

Such measures—and much more—have been developed not only within linear algebra but also in the dedicated fields of network science and graph theory, particularly in their application to social networks. Some of the most prominent fora for the dissemination of the latest research in network theory, methodology, and empirical applications include *Journal of Complex Networks, Social Networks*, and *Network Science*. An aspiring student of networks is encouraged to consult the following titles for an extensive introduction to the field, in the order listed below:

1. Borgatti, S. P., Everett, M. G., Johnson, J. C. (2018). *Analyzing Social Networks*. SAGE, and Borgatti, Everett, Johnson, Agneessens F. (2022). *Analyzing Social Networks Using R*. SAGE.
2. Wasserman, S., Faust, K. (1994). *Social Network Analysis. Methods and Applications*. Cambridge University Press
3. Barabási, A. (2017). *Network Science*. Cambridge University Press.
4. Newman M. E. J. (2010). *Networks. An Introduction.* Oxford University Press.
5. Kolaczyk, E. D., Csárdi, G. (2020). *Statistical Analysis of Network Data with R*. Springer.
6. Cranmer S. J., Desmarais B. A., Morgan J. W. (2021). *Inferental Network Analysis*. Cambridge University Press.

You may also wish to consult our introductory course on social network analysis and network statistics, available as an R script on GitHub: https://github.com/davorsalihovic/sna-and-statistics-course. This hands-on, practical introduction covers fundamental concepts, metrics, and inferential methods in social network analysis. It is designed to support further learning and to complement the six titles listed above.

[1] Josefine Bohr Brask, Samuel Ellis, and Darren P Croft, 'Animal Social Networks: An Introduction for Complex Systems Scientists', *Journal of Complex Networks* 9, no. 2 (1 April 2021): cnab001, https://doi.org/10.1093/comnet/cnab001; Sara Vanovac et al., 'Network Analysis of Intra- and Interspecific Freshwater Fish Interactions Using Year-around Tracking', *Journal of The Royal Society Interface* 18, no. 183 (20 October 2021): 20210445, https://doi.org/10.1098/rsif.2021.0445.

[2] M. E. J. Newman, 'The Structure of Scientific Collaboration Networks', *Proceedings of the National Academy of Sciences* 98, no. 2 (16 January 2001): 404–9, https://doi.org/10.1073/pnas.98.2.404; M. E. J. Newman, 'Coauthorship Networks and Patterns of Scientific Collaboration', *Proceedings of the National Academy of Sciences* 101, no. suppl_1 (6 April 2004): 5200–5205, https://doi.org/10.1073/pnas.0307545100.

[3] Yong-Yeol Ahn et al., 'Flavor Network and the Principles of Food Pairing', *Scientific Reports* 1, no. 1 (15 December 2011): 196, https://doi.org/10.1038/srep00196.

[4] Igor Feldman, Andrey Rzhetsky, and Dennis Vitkup, 'Network Properties of Genes Harboring Inherited Disease Mutations', *Proceedings of the National Academy of Sciences* 105, no. 11 (18 March 2008): 4323–28, https://doi.org/10.1073/pnas.0701722105.

[5] Nikolay V. Dokholyan, Boris Shakhnovich, and Eugene I. Shakhnovich, 'Expanding Protein Universe and Its Origin from the Biological Big Bang', *Proceedings of the National Academy of Sciences* 99, no. 22 (29 October 2002): 14132–36, https://doi.org/10.1073/pnas.202497999; Eric J. Deeds, Orr Ashenberg, and Eugene I. Shakhnovich, 'A Simple Physical Model for Scaling in Protein-Protein Interaction Networks', *Proceedings of the National Academy of Sciences* 103, no. 2 (10 January 2006): 311–16, https://doi.org/10.1073/pnas.0509715102.

[6] The term was coined by Richard Dawkins, who has used it to summarise an idea inspired by the thoughts of the great J. B. S. Haldane in his essay on "Possible Worlds". There Haldane wrote: "Now, my own suspicion is that the universe is not only queerer than we suppose, but queerer than we *can* suppose". (Haldane, naturally, used "queer" in the sense of "strange"): J. B. S. Haldane, *Possible Worlds and Other Essays* (London: Chatto & Windus, 1927), 286. Dawkins introduced the elegant notion that human brains are capable of intuitively understanding the world they have evolved in—the middle world which lies between the extremes of quantum mechanics and cosmic scale. While we have little or no problem understanding the trajectory of a thrown rock or Newtonian mechanics, comprehending wave-particle duality or Einstein is at the limit of our capabilities, and this is obviously something we *can* suppose.

[7] Pál Erdős and Alfréd Rényi, 'On Random Graphs I', *Publicationes Mathematicae* 6 (1959): 290–97.

[8] Albert-László Barabási and Réka Albert, 'Emergence of Scaling in Random Networks', *Science* 286, no. 5439 (15 October 1999): 509–12, https://doi.org/10.1126/science.286.5439.509.

[9] José Luis Estévez, Davor Salihović, and Stoyan V. Sgourev, 'Endogenous Dynamics of Denunciation: Evidence from an Inquisitorial Trial', *PNAS Nexus* 3, no. 9 (2 September 2024): pgae340, https://doi.org/10.1093/pnasnexus/pgae340.

[10] https://snap.stanford.edu/biodata/datasets/10000/10000-PP-Pathways.html; Monica Agrawal, Marinka Zitnik, and Jure Leskovec, 'Large-Scale Analysis of Disease Pathways in the Human Interactome', *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 23 (2018): 111–22.