Introduction to Social Network Analysis: Basics and Historical Specificities

Martin Grandjean
University of Lausanne

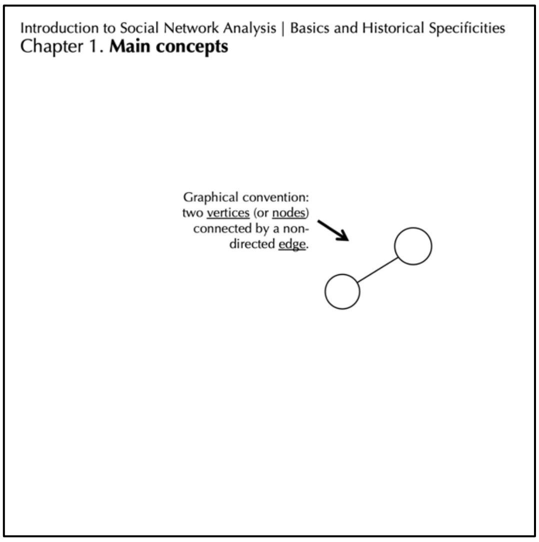
HNR+ResHist Conference 2021
June 30, 2021

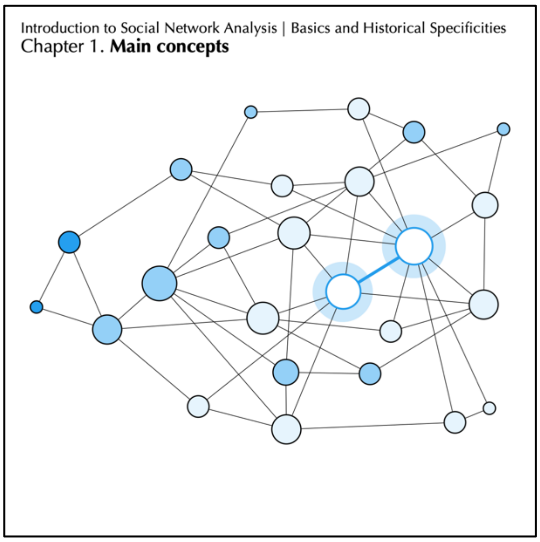
Chapter 1: Main concepts

Welcome to this introduction to network analysis. The purpose of this presentation is to provide a very quick introduction to network analysis with a particular focus on practices in the historical sciences. It is divided into five short chapters, which obviously cannot be exhaustive, but which aim to arouse curiosity to go further. The first chapter presents network analysis and its main concepts.

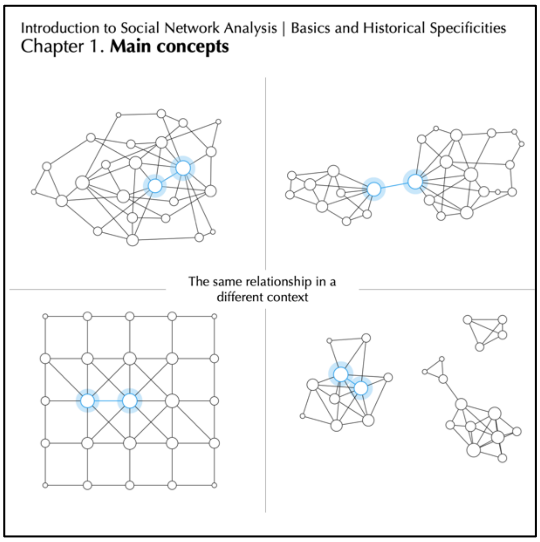
Contextualization

As a preamble, I propose an example that allows you to understand how analyzing a structure is different from any other statistical analysis: it gives context around the element we are studying. This is a very simple example, which allows us to approach the way we graphically represent a graph: here, two vertices connected by an edge represent an undirected relation between two elements. It is a type of representation that we are used to because we often see it around us: a metro map, an airline advertisement, a network of connected objects, etc. This kind of graph can be a network of people, organizations, places, objects, etc. For the purposes of this demonstration, imagine that it is about two people writing letters to each other.

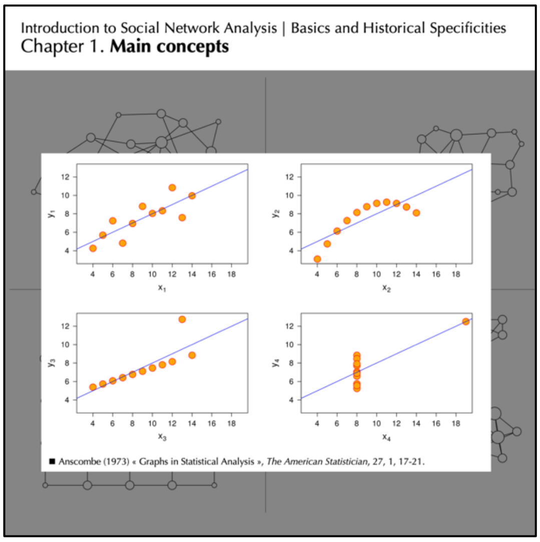




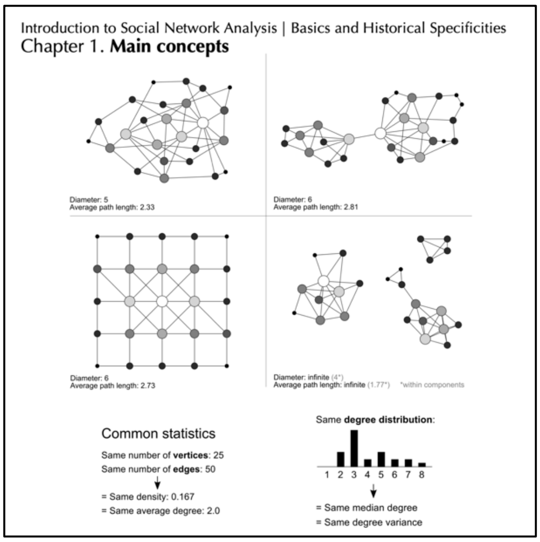
This relation therefore represents a certain number of letters, say ten. We can obviously qualitatively analyze the content of these letters between these two people. But they can also be placed in the context of other letters written and received. As soon as we add the other people who correspond with our two subjects, we realize that some of them can be common relations. It's very interesting to know that there are people (in this example, three), who can make the "link", or be a "bridge", between the two people in the center. Of course, these neighbors can also have relationships with each other, regardless of the relation between the subjects of our analysis. And knowing that these relationships exist is very important to take some distance, to realize that even if it is the two highlighted individuals that interest us, they are not the only factors of relationship, they are not necessarily the center by default. This decentralization is even stronger in a situation where people can send and receive letters without ever having a connection with the two people we are interested in. It's an additional level of context. And so, these neighbors of neighbors can also have neighbors, etc. And we can imagine going much further than that, it all depends on the corpus of archives we are working on.



To make you understand how important the context is to understand this specific relationship, these ten letters, let's imagine that it appears in different situations, structurally speaking. These four graphs all contain exactly the same number of people and the same number of relationships, or letters exchanged, but they are not distributed in the same way. Here, we understand that this relationship, highlighted in blue, takes on a completely different meaning if these ten letters are exchanged at the center of a network built around our two people (top left) or if they connect two groups that have nothing else in common (top right). Or if this relationship takes place in a group that is not connected to other groups (bottom right). The content of these ten letters between these two people never changes, so that if we limit ourselves to analyzing them qualitatively, the interpretation will always be exactly the same. But their context of appearance is so different that we can give then another meaning, or another status.

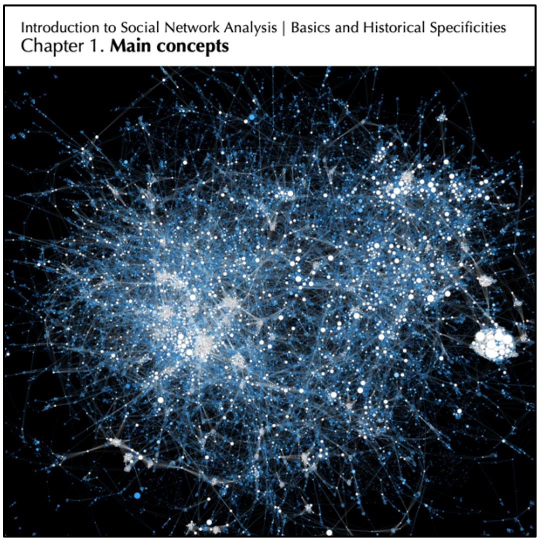


There is something else interesting about these four situations. In fact, these four situations are a version of Anscombe's Quartet adapted to network analysis. In his 1973 article, Francis Anscombe shows that it is important to visualize data even though it is more popular in the field of statistics to rely only on calculations. He shows that four data sets with very different distributions have identical statistical characteristics: same mean, same variance, same correlation and linear regression line, etc. If we rely only on statistics to interpret these data sets, we miss a very important information.



As I said, these four networks are all composed of the same number of vertices and edges. So, they have the same density. But above all, the number of connections of each vertex is always distributed in the same way: there is always a vertex which has 8 connections, two vertices which have 7, etc. So, if we simply produce a simple statistical report about these networks, we are missing out on extremely important structural information. This remark is therefore an encouragement to visualize the networks. Or at least, it allows to nuance the sometimes very strict positions that say that the purity of numbers, mathematics and statistics, does not need graphic representations, imperfect by nature.

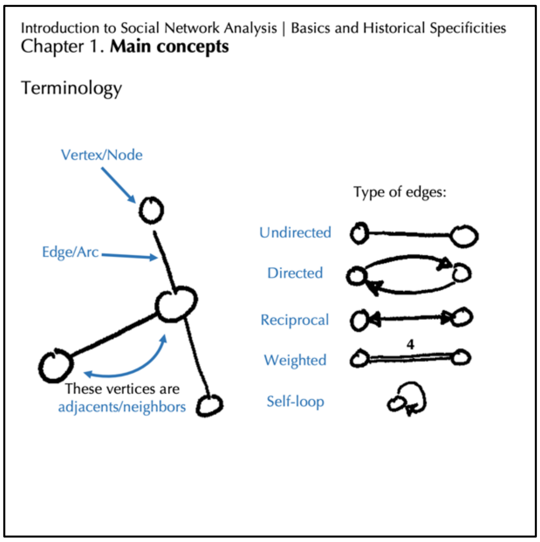
In fact, we know that graphic representation is problematic, and that visualizations are sometimes inadequate to make our subject clearer for our readers, but it is a very useful exploration tool to understand the data. Obviously, the four situations we have here are very simple and caricatural. What happens if we extend our example and imagine that our archive contains thousands of letters, between thousands of correspondents? This is when we will have to find new ways to "read" a network, because a complex network of several hundred vertices is already difficult to read. It is even more difficult when there are thousands, or tens of thousands, as here. This complex and massive network, this impenetrable galaxy, is what we call a "hairball" or a "big spaghetti monster". We see groupings, clusters, less dense regions, but we cannot read it with our eyes: we will need mathematics to decipher it. This does not mean that this graphical representation of these hundreds of thousands of relationships is not useful, it means that it will not be used in the same way: it is not a research result as such, it is an exploration interface.



Terminology

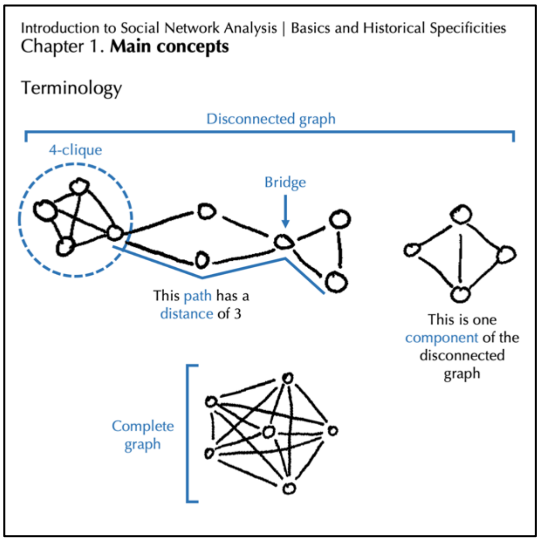
Now, this preamble was intended to show the essence of network analysis, which is to analyze a structural context, but now let's get back to the basics of terminology. Basically, the object we manipulate, the graph, is composed of points and lines, usually called vertices or nodes, connected by edges or arcs.

It is an extremely simple way to represent a relationship between two elements, a very high level of abstraction. As a result, we have to be very careful about what we put behind these abstractions: real life situations cannot be reduced to such abstractions except at a high price: that of an extreme simplification. This is called data modeling, the creation of a model that can be applied systematically. If we are aware of the reduction of complexity that this induces, and that we can therefore be critical of this process, then this systematic approach allows us to use all the tools of graph theory with great efficiency. If the vertices of the network are generally quite easy to define, when modeling (individuals, organizations, objects, etc.), the choice of the type of relationship is more critical.



An edge can be undirected or directed, but it can also be reciprocal, which does not necessarily have the same meaning as undirected. In general, we study graphs in which the relations are encoded in the same way, but our modeling may lead us to situations where some relations have directions (a sent letter, for example) and others do not (a friendship). It is important to note that in most empirical research networks, relationships can be weighted, they will be represented with more or less thickness depending on the intensity of the relationship, or if the same relationship appears several times in a row. More rarely represented graphically, but sometimes very real, self-loops are also possible.

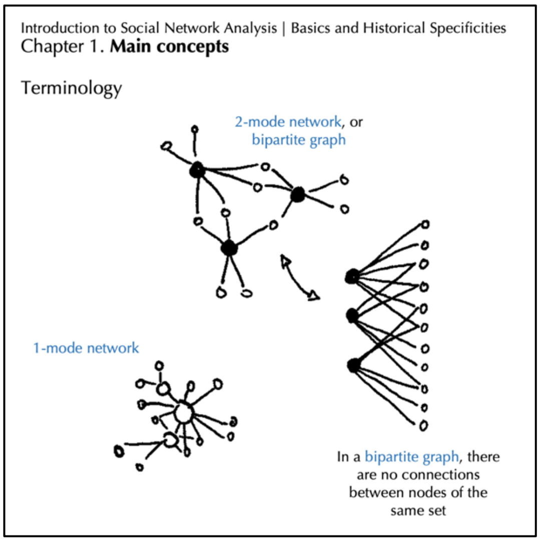
Let us now look at the qualification of the different situations that we find when these vertices and edges are connected. We distinguish for example connected graphs from disconnected graphs: the first ones are made of a single component, we can find a path between all the vertices of the graph. Disconnected graphs are composed of several components. It sounds like a completely obvious detail, but the fact that a network is continuous or divided into several components is rarely discussed in papers that are not purely theoretical: usually only the main component is commented on, forgetting the vertices that are not connected to the largest group.



If all the vertices of the graph have connections with all the others, we speak of a complete graph. This is rarely the case in real network analysis situations, but it allows us to discuss the completeness of the network by comparing it to a perfectly complete situation.

The search for complete graphs is also expressed in the long tradition of detecting cliques, which are those regions of the graph where all vertices are connected to each other and thus form groups of maximum density. The graph on the left here contains one 4-clique and five 3-cliques (triangles joining three vertices). When the removal of a vertex would result in the disconnection of two components of the graph, it is called a bridge. The detection of such bridges is very important for the evaluation of the robustness of a network, we will come back to this.

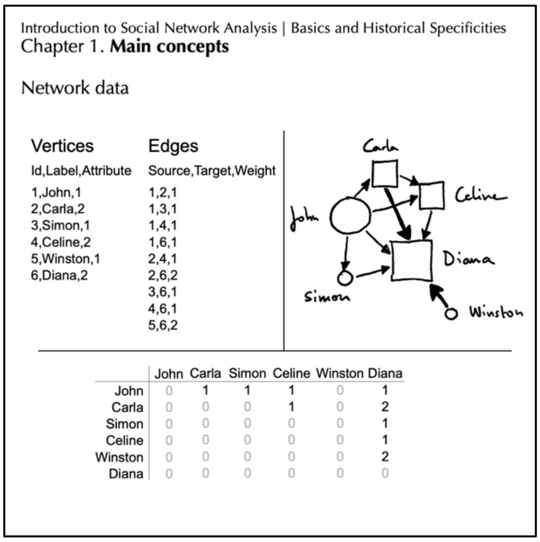
If we now come to the categorization of network types, we distinguish two main families: 1-mode networks, which are composed of a single set of vertices and 2-mode networks which contain two different sets of vertices. It is not the nature of these vertices that is important, but their structural characteristic: in a 2-mode network, which is basically a bipartite graph, there is a connection only between nodes of a different type. To take an example, we can very well represent people, organizations, objects, concepts in the form of a single 1-mode network, with relations between all these elements. But if we want to formally analyze the affiliation between individuals and institutions, for example, we are only interested in this "vertical" relationship, between vertices of a different type, and no longer in the relationships between the individuals themselves. This is an important distinction: most of the time, we use 2-mode modeling to produce 1-mode graphs, which are then easier to analyze.



Network data

Without going into too much detail on this point, which could be the subject of an entire course, here are some elements concerning the formatting of data. Basically, a network is an adjacency matrix, as you can see below, which allows you to read the relations between all vertices. Note that in this matrix, the relationships are directed (it must be read from left to right) and weighted (see values 1 and 2).

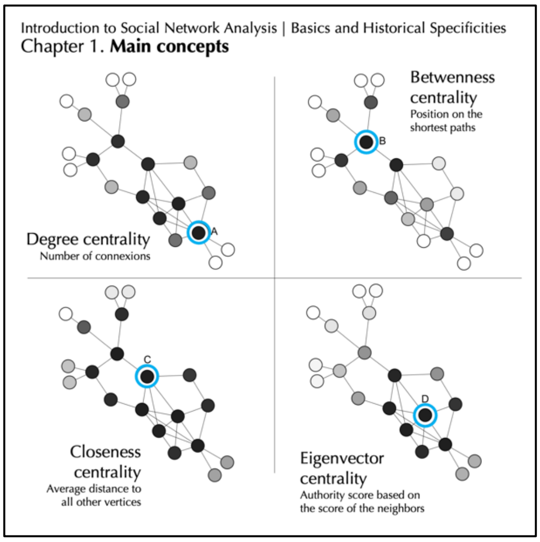
Such a matrix may be sufficient for us to analyze small to medium sized graphs, but the encoding itself is usually done with a simplified list of relations, as seen on the left in the "edges" column. On the first line, we read that vertex 1 is linked to vertex 2 by a relation of size 1.



	John	Carla	Simon	Celine	Winston	Diana
John	0	1	1	1	0	1
Carla	0	0	0	1	0	2
Simon	0	0	0	0	0	1
Celine	0	0	0	0	0	1
Winston	0	0	0	0	0	2
Diana	0	0	0	0	0	0

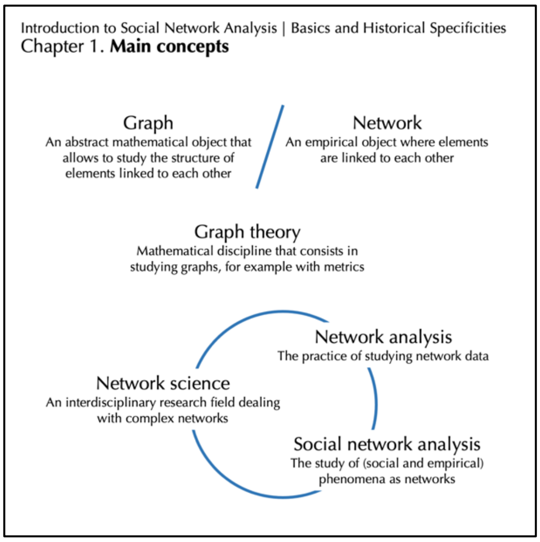
A list of vertices, here on the left, gives us the key to decrypt the identifiers as well as additional variables, such as a label to display, or attributes to sort the vertices. This matrix and this adjacency list both give us the small network on the right.

Metrics



If the interpretation of the metrics provided by graph theory will be discussed in more detail in a following chapter, we must nevertheless mention the principle here. To explain how these centrality measures are calculated, let us take a very simple network. Here, it is duplicated to compare the result of the calculation: the darker a point, the higher the score. The point highlighted in blue is the vertex of the graph that gets the highest value. At the top left, we find the degree centrality, a metric that is easy to understand since it consists in counting the number of connections of each vertex. If we are still talking about a network of letters, then this value is simply the number of people who correspond with the selected individual. Here, the highest score is reached by individual A, who has 7 connections. In this conception of centrality, an individual is central if he has many connections. At the bottom left, we find the closeness centrality. It consists in measuring how far are all the vertices of the graph from each other. The one who has the smallest average distance with all the others is therefore the one who is closest to all the others on average. Here, vertex C has the highest mean proximity. In this conception of centrality, an individual is central if he is in the middle of the graph, in terms of overall topography. At the top right, the betweenness centrality consists in detecting all the shortest paths between the vertices of the graph and then counting how many times a vertex is on the path between two others. We can therefore say that the most central individuals according to this measure are those who connect different parts of the network that would not necessarily be connected to each other. They are the "bridges", the "information carriers". Here, it is vertex B which is most often on the path between all the possible pairs of vertices. In the case of a correspondence network analysis, we would probably not focus on this vertex B, but the betweenness centrality reminds us that it is the only path to 5 vertices of the graph (above it, on this image), which represents 25% here! In the next chapters, we will see that the interpretation we can make of this statistical and structural information depends on how our data has been extracted from historical sources.

Conclusion



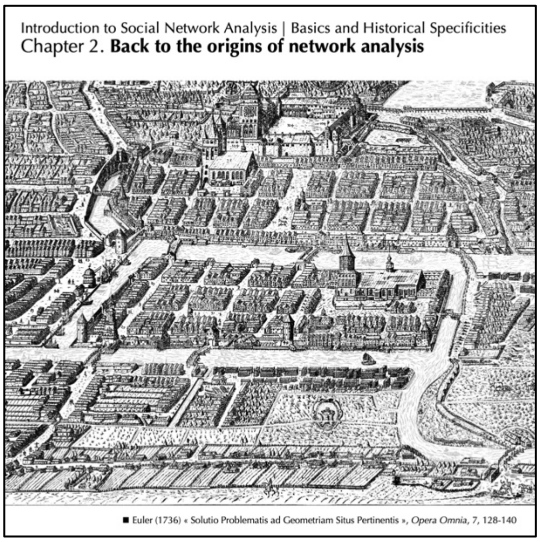
As a conclusion to this introductory chapter, I would like to continue our terminology exercise by also applying it to these sometimes unclear concepts. They are often mixed, and this is normal since science is an iterative and above all social and cultural process: concepts are produced without there always being a consensus on the definition or on what a field really covers.

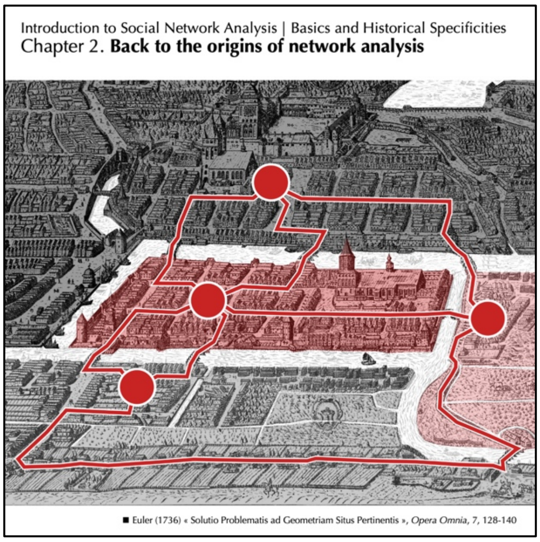
However, the difference between a graph and a network is quite simple. The graph is an abstract mathematical object, while the network is its concrete counterpart. This means that when you want to analyze a network, you model it as a graph. And once you have analyzed it with the tools of graph theory, you translate the results into the language of your object, of your network. The difference between network science, network analysis and social network analysis is more subtle since these concepts seem to be interchangeable. But in fact, even if they describe things that overlap to a large extent, these concepts are positioned at different levels: network analysis is the analysis of network data, it is a technical practice, whereas social network analysis is much closer to the phenomena that create these networks, it is the analysis of the phenomenon, not only of its data. For its part, network science is a field of research, almost a discipline, which has evolved towards the study of complex networks. The same could be said of Historical Network Research, it is a community of practice around a set of methods and tools.

Chapter 2: Back to the origins of network analysis

In this second chapter, we will review three seminal examples of network analysis. First of all, I think it is important to understand where the concepts we use today come from and to be aware of the temporality that separates us from these milestones. Because we always assume that what we do with new technologies is a revolution. We will see that our tools allow us to go much further than what was done several decades ago, but the principle is relatively similar. Second, these three examples are part of the common culture of our community, it is important to have these references to understand current research and publications. Finally, these are very inspiring examples: even though they are very simple, they raise questions that are still valid today.

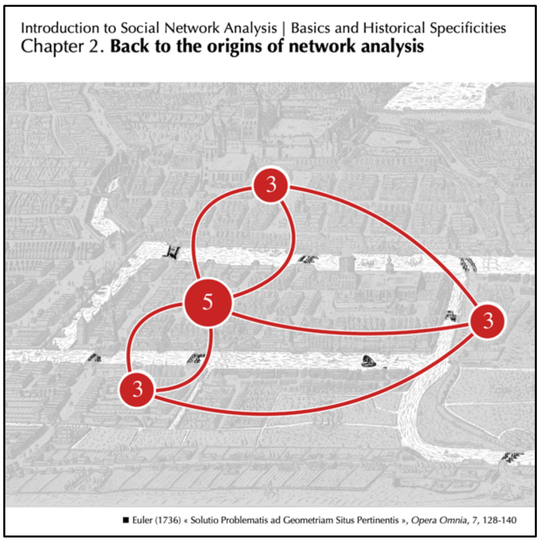
Leonhard Euler





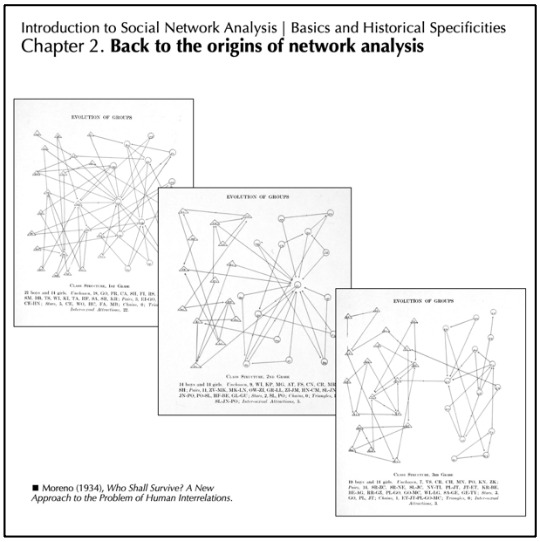
Our first milestone is the so-called "Königsberg Bridges" problem, or the "seven bridges of Königbserg", solved by Leonhard Euler in 1736. Located in East Prussia, the city of Königsberg, now called Kaliningrad in Russia, is built on the river Pregel and includes two islands in the city center. There are four main districts, separated by seven bridges. The question is whether it is possible to take a walk across all the bridges without passing through the same one twice. What Euler shows is that topography or street layout is not important. All that matters are the bridges themselves. He will therefore reduce the problem to a mathematical abstraction. Even if he does not represent it graphically in the form of a network, we can synthesize the situation as follows: the districts are the vertices of the graph, the bridges are the edges between them. This also means that we can represent the problem in a simplified way, for example like this. Leaving the map here allows to focus on the problem in an abstract form. And Euler's reasoning is that if you enter a district by a bridge, you must be able to leave by another bridge. This means that the number of bridges connecting a district must be an even number. There are two exceptions: the departure and arrival districts can have an odd number of bridges, but all the others must have an even number of connections to be crossed. In fact, this is exactly the same as calculating the degree centrality of each of the districts. The proof is then easy to do: if all vertices have even

degree, then the path exists. If two of them have an odd degree, it is also possible as long as we do not try to return exactly to the starting point. But if there are more vertices with odd degree, then the path does not exist. And unfortunately, this is the case in Königsberg: all districts are connected to each other by an odd number of bridges, which means that the perfect walk is not possible.
The solution to this problem is considered the first theorem of graph theory. What is interesting for us is that it prepares the foundations of a computational approach to structure: it is by calculating the number of relations of the vertices of the graph that we obtain a key information.

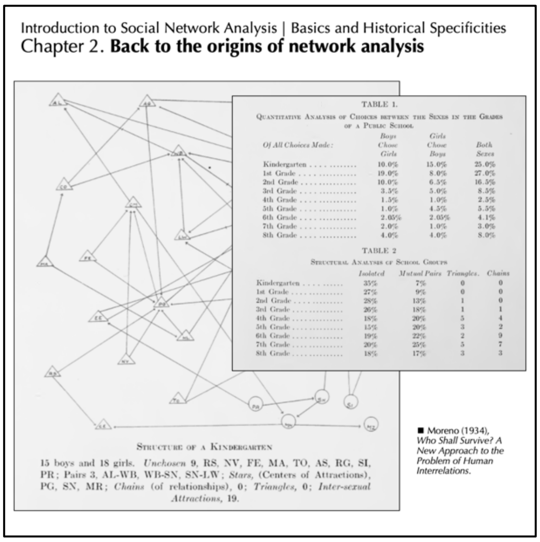


Jacob Moreno

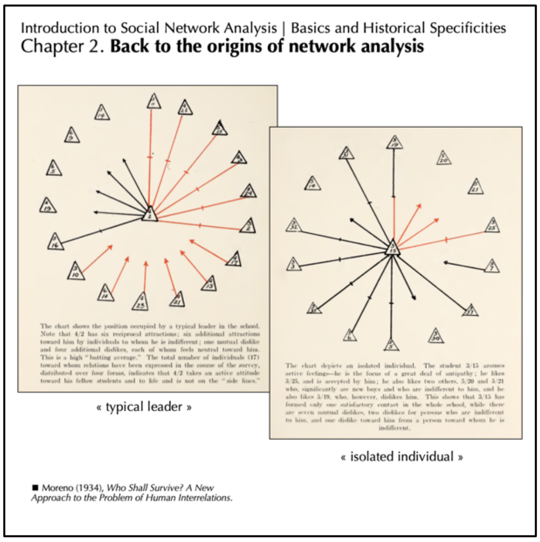
For our second milestone, we jump two centuries in time to the twentieth century, in the 1930s, to discover the work of Jacob Moreno and Helen Hall Jennings. Forerunners of sociometry, Moreno and Jennings are interested in the micro-societies that form in the New York Training School for Girls in Hudson, a reform school for adolescent girls. They map the friendships and enmities of these young women and show why and how an epidemic of runaways develops in the institution. But the example we are interested in today is based on another dataset: an analysis of mixed school classes of all grades. For each of these classes, they observe the process that leads, at the beginning of the year, to the choice of who the students choose to sit next to. This means giving each student a piece of paper and asking them to name the two people they would like to sit next to. The researchers then compile these pieces of paper and summarize the situation in what they call a sociogram.



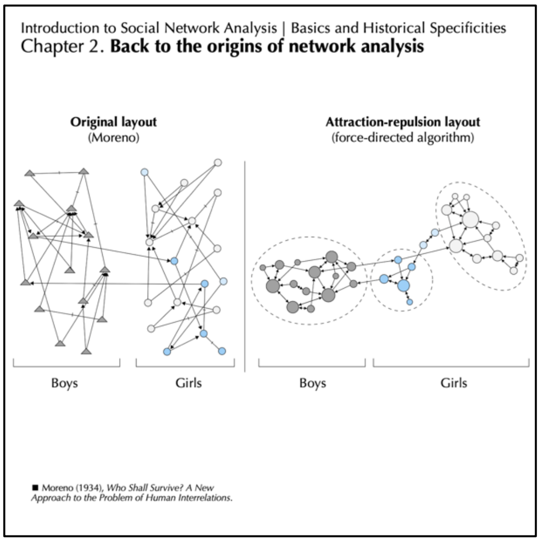
Published in 1934, these sociograms are the first instance of a formal analysis of social networks. Here you see three examples, the boys are on the left, represented by triangles, and the girls are on the right, represented by circles. It is important to note that Moreno and Jennings do not use the term "network", they speak of "sociometric diagrams". The term "social network" was introduced into the field of interpersonal studies by John Barnes twenty years later. Moreno and Jennings do the same analysis with classes of all grades and show in particular that the attraction between boys and girls is fluctuating: in the early years, it is common to have boys and girls side by side, while this changes in adolescence, to come back from time to time in the last years of school. In the table on the right, we see that it is not only a visual analysis, we see that they also count the percentage of isolated individuals, the frequency of mutual attraction, the number of triangles and chains.



These statistics lead to the creation of a whole terminology. For example, individuals who are frequently chosen are called "stars", not by analogy with famous people but because the arrows form a star around



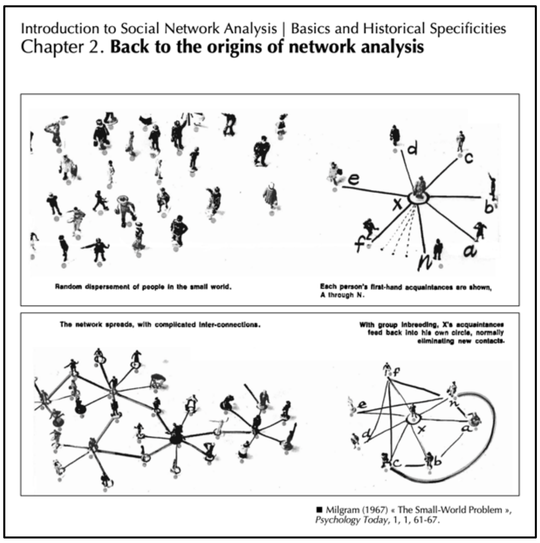
their node. In the representations presented here, we go even further since we represent positive (in red) and negative (in black) relationships. These are directed, can be mutual or not symmetrical. As in the Euler example, we are really counting relations. But we don't go further, we don't analyze the global structure but only the position of a specific element in the network in relation to the others. This is of course due to the technical means: the algorithms that allow us to calculate the shortest paths and test all the possibilities do not exist yet.



What is very interesting in Moreno and Jennings sociograms is the fact that these technical constraints and their hypotheses on affinities between the sexes lead them to spatialize their network in an original way: the boys are placed on one side and the girls on the other in order to clearly see the edges that lie between these two groups. Of course, you have to imagine them with these dozens of pieces of paper, trying to transfer all this onto a sheet of paper, or onto a large flat surface, they don't have algorithms that can optimize the position of each element according to its connections. Today, we use mostly force-directed algorithms, which make the connected vertices attract each other and, on the contrary, make the vertices that do not have a common edge repel each other.

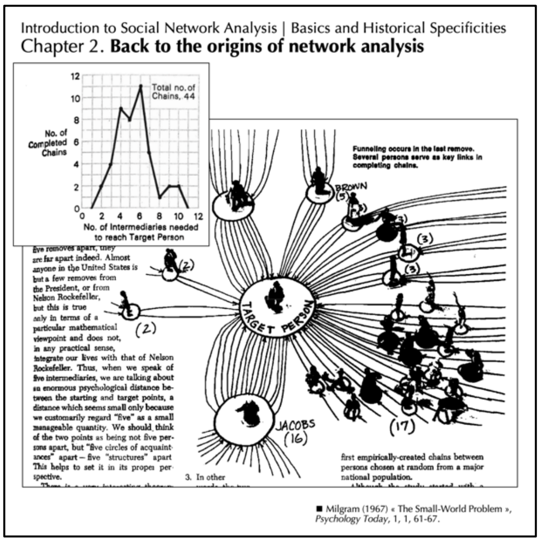
However, if we reproduce this exercise almost a century later, with recent tools, we observe different things than what Moreno and Jennings wanted to show. What we see in this example is that there are not two groups facing each other, boys and girls, but that the girls are in fact divided into two groups: the handful of girls shown in blue here are no more connected with the rest of the girls than they are with the boys' group. This is structural information that is very difficult to see in the original visualization. And it changes the whole interpretation! My goal here is not to develop further, but to show that our visualization decisions have an impact on our reading of the network, even if we always have exactly the same data set. This is why network visualization is mostly an exploratory process: we play with the data, we turn them around, not to make them say what we want but to try to understand as best as possible how they are organized.

Stanley Milgram



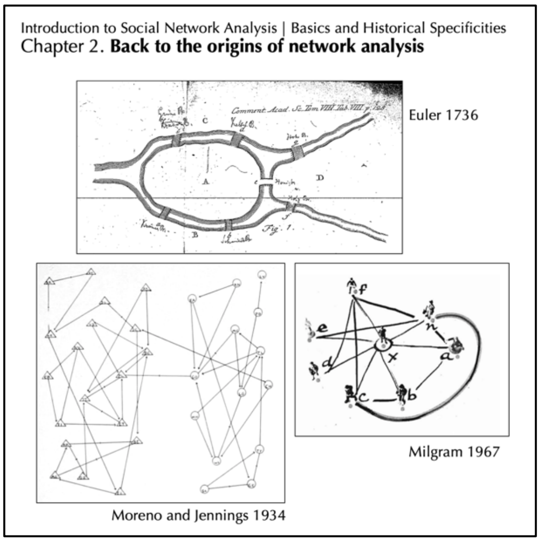
The Small-world problem of Stanley Milgram is our last milestone. Taking the idea of a game created at the beginning of the twentieth century and consisting of finding the shortest path between two individuals in the world which postulated that there was an average of 6 degrees of separation between two people taken at random, Milgram published in 1967 an article in a popular magazine in which he relates his experiment.

He sent over hundreds of letters to random people in Nebraska asking them to forward them to the person they thought was closest to a designated person in Massachusetts. Obviously, this also involves documenting the process, asking everyone in that chain to also send him the address of the next one, etc. The practical conditions of this experiment are obviously a bit problematic, and the results have been widely criticized afterwards. But the experiment itself has generated a lot of enthusiasm. He has indeed shown that there was a degree of separation of six people (but this is questioned), but what interests us here is not the result but, on the one hand, the very pedagogical illustrations of his article and, on the other hand, the concept of small world that he puts to the test. Here you see all the chains converging on the target person. This brings us to the concept of "small world", a type of network where it is very easy to go from one point to another in few steps, despite the very large number of vertices, their very heterogeneous distribution, in lots of small dense groups linked to a few very connected elements. A good example would be the current airline network: there are thousands of airports, most of them rather small and connected only at the national level, but it only takes a few international hubs to make any journey in the world in two or three steps.



Conclusion

These three examples are very complementary when discussing the relationship between measurement and visualization. In the case of Euler, we do not visualize since we only work on an abstraction. This abstraction leads to a rigorous counting, which is sufficient in itself for the demonstration. In the case of Moreno and Jennings, we are in a counting logic which can hardly take place without visualization. It is even this visualization that leads to conceptual developments, when they speak of stars, for example, or to show the relationship between boys and girls. But in the end, it is a somewhat disappointing visualization because it conditions the interpretation a lot. Milgram's study is the opposite of Moreno's. In his case, he does not start from an existing network that is going to be modeled, but from the abstract concept of network to give it concrete life. His visualizations are there to make us understand the process, but they are not data visualizations as such, since he does not have enough letter returns to really map the path of all the mail.



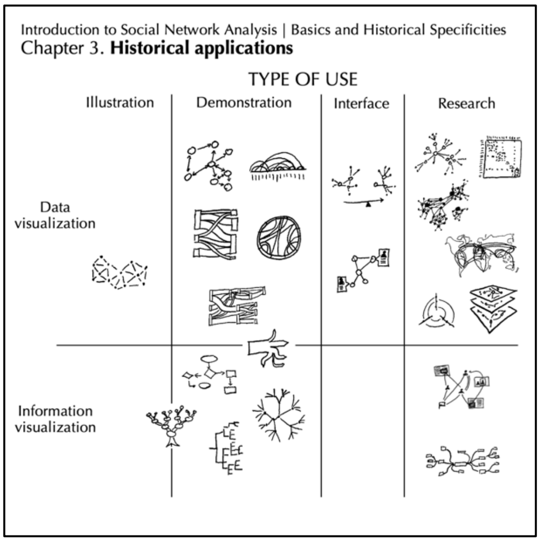
It is interesting to see that these examples raise very contemporary questions about the use we make of network analysis. This is the subject of the next chapter.

Chapter 3: Historical applications

In this chapter, we will focus on how we integrate network analysis and visualization into the research process. The focus here is on historical research objects, but these considerations are valid in most fields. So, what use do we make of these networks in the historical sciences? To help us with this question, I would like to approach it from two angles: what is the status of the visualizations we produce? And how do we use historical data to produce networks?

A typology of network visualization

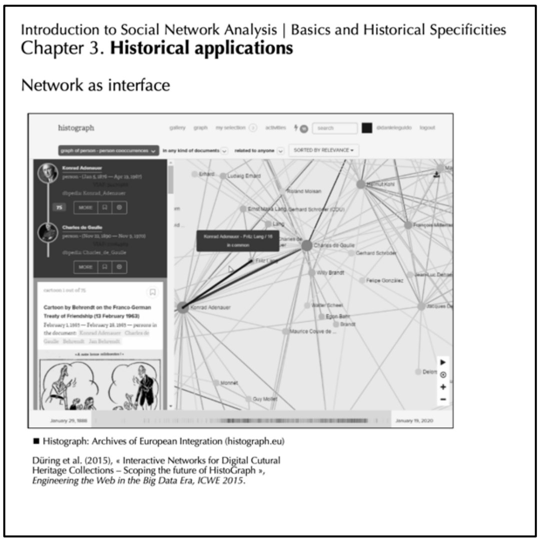
Let's start with network visualization. The typology I propose here is a fairly classical classification and can be applied to all data analysis and visualization in general. Classification is an exercise that is necessarily imperfect, but I find it more useful to position ourselves at such a conceptual level to discuss our uses of network analysis than to make a huge catalog of everything that has been done so far. I propose two axes: the type of use, on the horizontal axis, and the category of data analyzed on the vertical axis.



In the wake of the "exploratory data analysis" thematized by John Tukey in 1977, a distinction was made between the illustrative approach and a more experimental approach made possible by the processing of statistical data on a larger scale. This distinction between what we will call "demonstration" visualization and "research" visualization is at the heart of the question of the uses and integration of data analysis and visualization in the research process. This may be a bit theoretical, but I feel that what is expected from such an introduction is not to leave this kind of questions in the dark. We often realize in our own day-to-day practice that we are not clear enough about the purpose of network visualization: we produce something that we intend for an audience, but which is in fact unreadable or, on the contrary, is intended to be a part of our historical enquiry and end up as a mere cover illustration.

To detail these categories, we can distinguish between network visualizations intended to illustrate a point, those that are used to demonstrate or present a situation through data, those that will serve as an interface, as an access point to interact with the data, and finally those that are intended to try to bring new knowledge about the object studied in the framework of a research process. On the vertical axis, the nuance between the notions of "information visualization", or infographics, and "data visualization" is difficult to perceive since one is often used for the other. However, the difference is important in order to understand how the approach differs, even if the result is sometimes very similar in appearance. Producing a representation based on a compilation of information is an act that involves a graphic and often manual "layout". On the other hand, the visualization of a dataset is an operation that can be performed automatically by software.

As we will see, a significant portion of network analyses and visualizations meet the definition of demonstration visualization: the dataset could be analyzed simply with a spreadsheet, but its graphical representation adds some readability and better communicates the result to the audience. In this category, we find sociograms, flow diagrams, arcs, or even simple 1-mode and 2-mode networks. In contrast to these objects intended to be shown, as soon as the network becomes very massive, or is particularly complex and intricate, its value as a visual object decreases. It is not always good enough to be published in an article, so much so that one has to master the codes of network analysis to be able to draw something from it. This is typical of a research visualization, which is a tool for the researcher in his exploration process. If the complex and scale-free network is typical of the "research" network visualization, all multilevel representations can also belong to this category, as well as network matrices and maps that call for a more topographic reading. Between these two categories, we find the uses of networks as an interface to facilitate the access to the data. Of course, all research visualizations are interfaces between the researcher and his data, but these visualizations are designed for this purpose: interactive sliders to play with temporal evolution, graph databases that allow access to additional information, etc. A good example might be the Histograph interface, which allows you to click on the edges and have the documents that create the relationship displayed. Search parameters also allow to select a time range, to play with tags or to explore the corpus directly from the reproductions of the annotated documents.



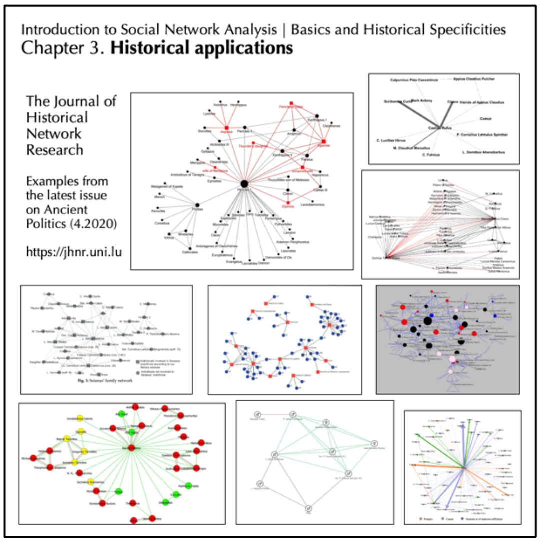
And we should not forget that sometimes network visualization is used as pure illustration. This is a bit problematic since no data visualization is supposed to be pure decoration, but the notion of network being a very powerful concept, it is not uncommon to see networks on a book cover but to find none inside...

I can't resist sharing this meme about our temptation to make use of the graphic representations we produce even if they are not useful. In this case, it's a very "meta" meme about the outputs of the Gephi software, known for producing colorful and attractive visualizations, since it's one of the creators of Gephi himself who publishes it, aware of the risks that such a tool can bring.

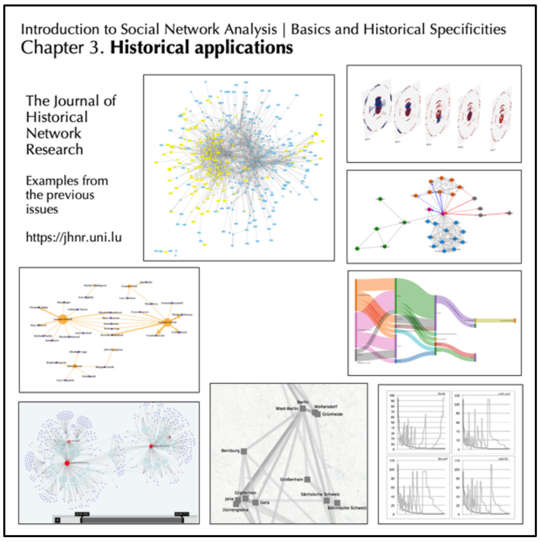
I also find it important not to relegate information visualization (here below) to the background: often the network we need is not the product of data analysis but rather of information formatting. This is for example the case for family trees, or flowcharts, which can be real synthesis and demonstration tools. And just because network visualization is the result of information compilation and not quantitative data processing does not mean it cannot also be heuristic. Mapping a subject by gathering in a space all the relational information we have, like a profiler in the middle of an investigation, can help us to bring out patterns and show us where to keep digging. The same is true for mind maps, but more from a conceptual point of view: at a high level of detail and intricacy, this kind of object is quite a research tool.

Examples

To briefly complete this reflection on uses, there is nothing like taking a look at what has been published recently, for example in the Journal of Historical Network Research. Here are some networks taken from the last issue, in ancient studies. In most cases, we are obviously working with medium-sized datasets, which remain easily readable. In several cases, a handful of vertices concentrate most of the edges, which means that this result was already detectable in the data file: the visualization is therefore not strictly speaking heuristic. It is nonetheless useful to "demonstrate", or simply "show"



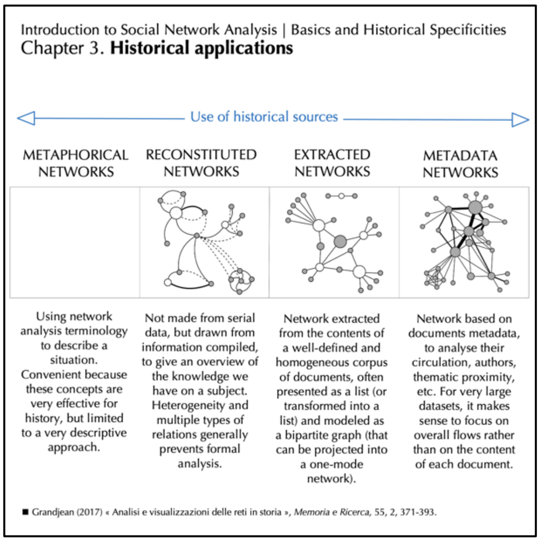
the data to the readers. The previous issues of the journal contain a little more diversity, which is normal since this time they are not thematic issues. There are some more massive networks, as well as multilayer representations, flow diagrams or interactive interfaces allowing to play with temporality. I think we learn best by example, so I can only encourage you to delve into the literature to understand the journey of other researchers and see how they use network analysis in their research and historical narrative process.



For this, the Historical Network Research Collaborative Bibliography is a very good resource. As I said, the purpose of this introduction is not to make an inventory of all historical network analyses, but it is good to know that there are places where such inventories exist, even partially. But now that we are clear on the status of network analysis in our research process, it is useful to ask: how do we extract a network from our historical data?

Historical sources and networks

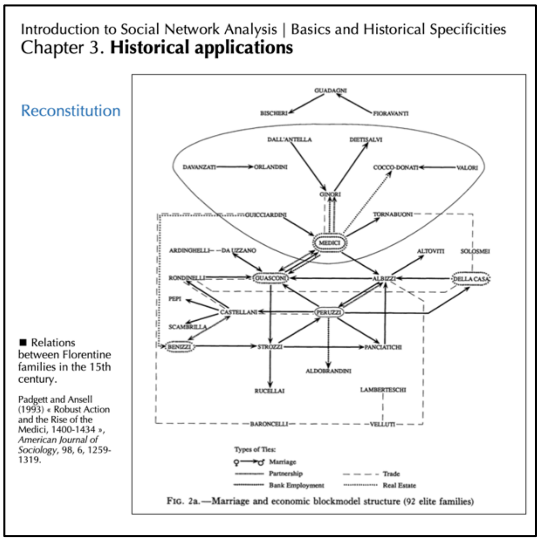
First of all, it must be said that the practices are as multiple as our historical sources are diverse. But as a gateway to this problem, I think we can divide the ways of using historical sources for network analysis into four broad categories.



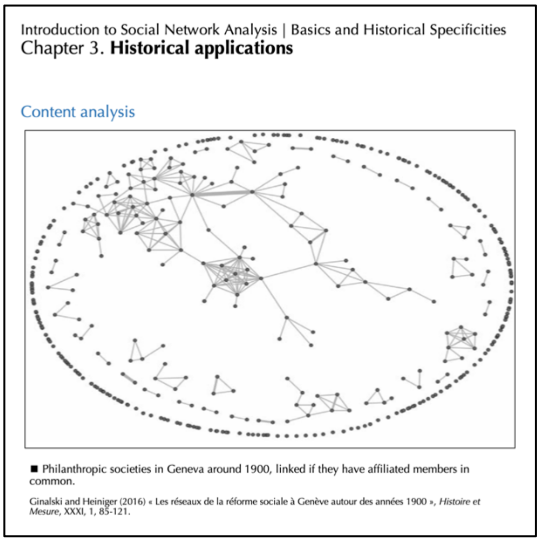
The first is the metaphorical use, which consists of using network terminology without actually doing any analysis or visualization. It works well in some cases, because these concepts are very popular in history, but it's obviously very limited. I don't think we have a monopoly on the network concept. And we saw in the previous chapter that the field is constantly evolving and that it is as successful because the notion of network has a lot of meaning in history. Indeed, we are always looking for connections between people, events, places, etc. And the very term "network" existed long before it was associated with formal approaches related to the application of graph theory. As you will have understood, it is a bit of an exaggeration to mention it here as one approach among others, but it is also to testify to the fact that the term is probably more used in history outside the context of formal quantitative analysis than to describe networks as we mean it here.

The second is the creation of "reconstituted" networks. A bit like when you take a large blank page to write down everything you have found on a certain subject, draw links between this information, etc. It is not always data visualization but more some kind of drawing that gives an overview of the knowledge you have on the subject. Although there may be data sets behind this kind of analysis, the data heterogeneity and multiple types of relationships usually prevent formal analysis and the use of graph theory metrics.

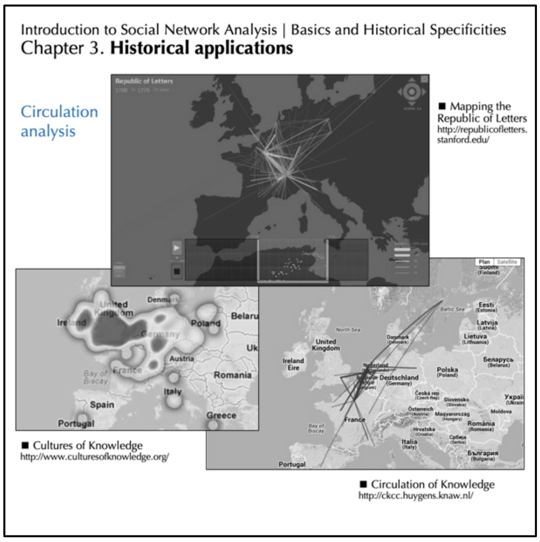
But, depending on the object of research, we sometimes have no other choice, either because the archives are incomplete or very small, or because what interests us is the compilation of everything we know about the subject. A very good example, which shows very well the power of synthesis of this kind of networks, is this analysis of the relations between families of the Florentine elite of the fifteenth century. This visualization here shows only some of the nine types of relationships that the families or individuals composing them could have: marriage, economic, political, financial relationships, friendships and alliances, etc. In this case, the aggregation at the family level, very varied sources and a very uneven level of information depending on whether the family has well preserved archives or not, makes a quantified structural analysis complicated. But this does not prevent an effective discussion of the social, economic and political mechanisms at work in Medici's Florence.



The third category concerns networks produced from the content of archival documents, in general a well-defined and homogeneous corpus, often presented as a list (or transformed into a list). This type of extraction often produces bipartite graphs: indeed, when a list is extracted from a source, the elements are easily affiliated to the entity in which they were found: this can be a list of people mentioned in a text, which will create a text-person link. Or it could be quantities of goods affiliated with a trip between two cities. Or simply people affiliated with institutions or groups, etc. In this example, each vertex is a philanthropic society, they are linked together if they have members in common. It is a 1-mode network resulting from the projection of a 2-mode network of people and organizations. In other words, we "evacuate" the individuals to represent them by a relation between two organizations. If you've never done what's called a projection, it might seem a bit cryptic, but two-thirds of historical network publications use this modeling, consciously or not.



The fourth and final type of network no longer relates to the content of documents but to their metadata: the circulation of a letter, its authors and recipients, an analysis of the proximity of language, etc. This is especially appropriate when working with large datasets, when it makes sense to focus on information flows more than on the content of each document. But size is not necessarily important, what matters is the type of information we get from the historical source: to take an example, it is like studying the route of Milgram's letters across the United States rather than the content of the papers written by Moreno's school children. The cartographies related to projects studying the "Republic of Letters" of the 17th and 18th centuries are excellent examples of these metadata networks. It can probably even be said that they are at the origin of a revival of historical network analysis about ten years ago. They influence many current projects by their aesthetics and their very global approach. A quick word on these cartographic representations: it is interesting to note that they suffer from the same problem as Moreno's sociograms: because the vertices are assigned a position in space which is not the one they would occupy if they were left free, they strongly influence the interpretation. This is an assumed goal here, we want the reader to understand the geographical dimension,

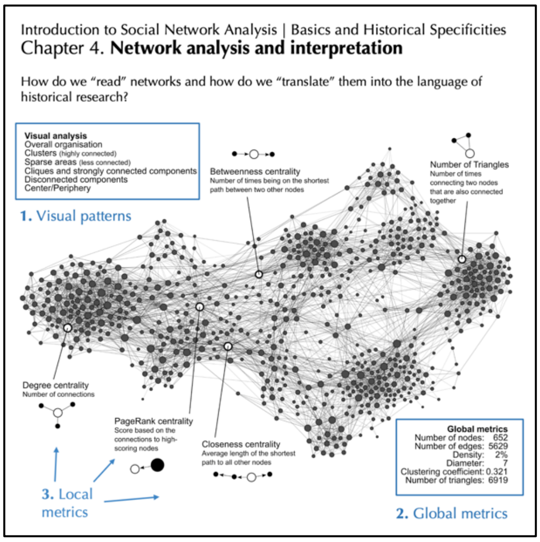


but it implies that a letter written between scholars in Paris and London will be much less visible than if it travels between Rome and Moscow.

Conclusion

To conclude on these historical aspects, we must of course keep in mind that part of our approach is data-driven and that we therefore do not totally choose which process we will use. But being aware of these different functions of network visualization, as well as of these different ways of extracting networks from historical sources, is important in order not to limit ourselves to choosing the simplest and most obvious path.

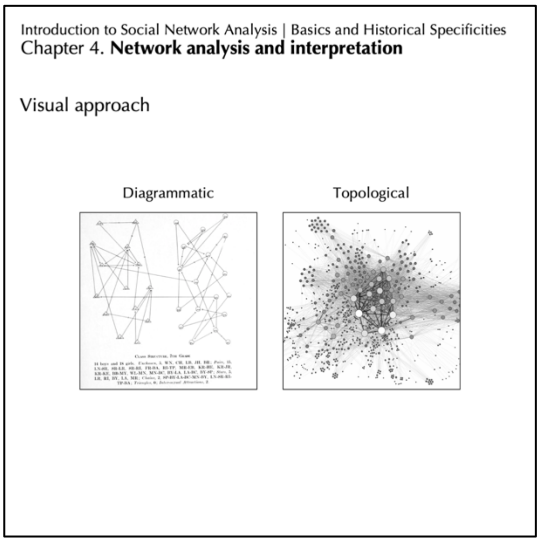
Chapter 4: Network analysis and interpretation



Now we are going to ask ourselves the question of how to interpret a network analysis result. The language of graph theory is not that of the historical sciences, which is why I propose to speak of a "translation" practice: how do we appropriate the toolbox provided to us in order to derive valid results for the humanities? How do we translate the metrics into our language, for example?

We distinguish three level of analysis, that can be articulated: The **visual analysis** (here top left), that consist at looking at the overall organization; the **global metrics** (bottom right) that measure general characteristics of the network; and the **local metrics** (bottom left) that describe the position of one element among the others.

Visual analysis

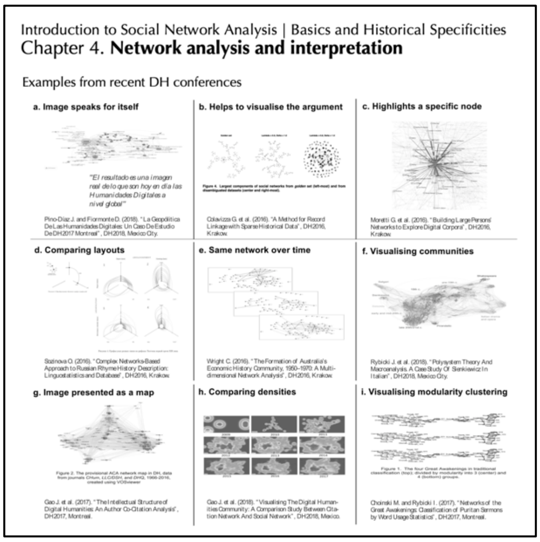


Regarding visual analysis, the way we read networks has changed over time. Historically the question of network readability was asked in terms of aesthetic criteria. For example, Moreno we saw in a previous chapter explicitly sought to avoid edge crossings. Even in the nineties, when giving birth to the modern layout algorithm, Früchterman and Reingold (1991) aimed at "minimizing edge crossings" and "reflecting inherent symmetry".

We call **diagrammatic** the perspective where the network is a diagram that we read by following paths. We do not want the edges to cross, and we use aesthetic criteria to bring clarity. It is still relevant to small networks and local exploration.

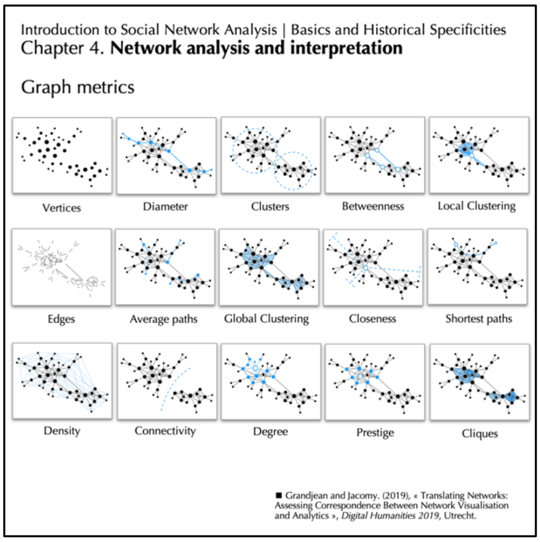
Then we call **topological** the perspective where the network is a structure that we read by detecting patterns. We expect the visualization to help us retrieve structural features like clustering or centralities. It is a common practice in digital history, more holistic and relevant to larger networks.

To illustrate our very heterogeneous relationship to visual network analysis, here are some examples from abstracts of recent DH conferences. Networks appear to have a wide range of usages. Their visualizations are either self-sufficient (a), an optional help to understanding (b) or strongly connected to the text. Some authors use them to highlight the position of a specific node (c). Network visualization can be used to compare layouts (d) or the layout of the same graph in time (e). They may aim at visualizing communities (f). Or they can be used at mapping a general structure, sometimes considered a "map" (g), or tracking density patterns (h) or monitoring algorithms like modularity clustering (i).

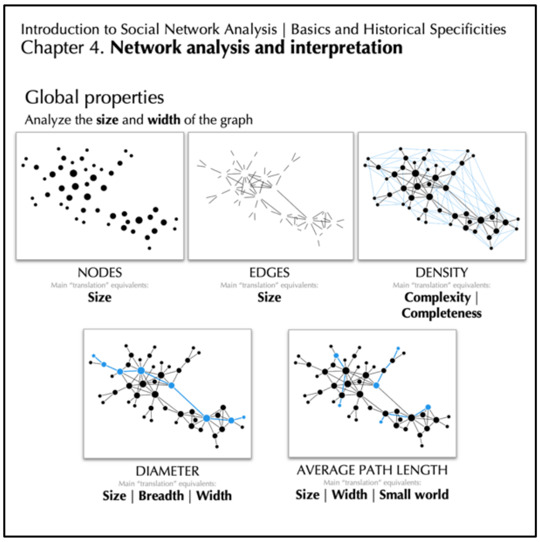


Metrics

Regarding the metrics, we've already seen that they are often opposed to visual interpretation, of which they are supposed to be a more objective and reliable representation. But graph metrics have a history that goes back to more than half a century and it shows that they are not immutable and require constant adaptation to usage.

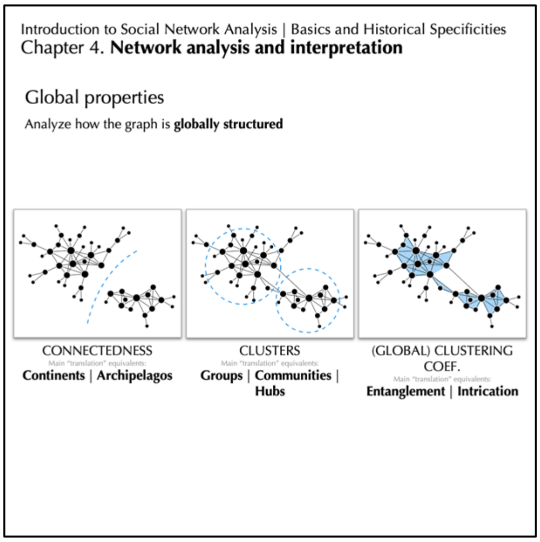


Moreover, Linton Freeman, in 1979, insists on the fact that the notion of "centrality" is the result of several intuitive conceptions. To remind that these metrics are based on "intuition" means to recognize that they have no meaning in themselves and that their interpretation must be rediscussed - and therefore translated - according to the context. In the following slides, I will therefore list the most commonly used metrics and suggest ways of translating them to show where and how they are generally interpreted in the historical sciences (and in the humanities in general).



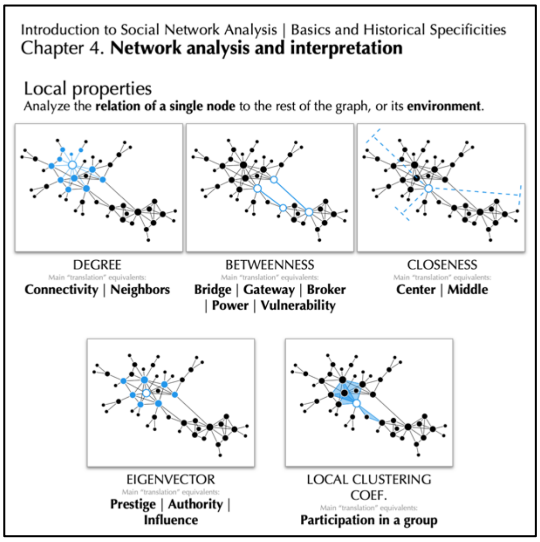
Statistical analysis allows for comparing networks across multiple dimensions at once. For instance, comparing the **number of nodes and edges** of different graphs of the same type can be a ranking tool that is directly translatable into natural language. The relationship between these two very simple measures, the **density**, is a very good indicator when comparing large networks in which it is not possible to count all the elements visually. The diameter can be used to describe how the density is distributed: complex networks are often characterized by a small diameter while high diameter is frequent in geographical networks. It can be translated as the width of the network. The average path length could serve as a complement to diameter because the latter can be influenced by a few nodes that are very far from the main component of the graph. Like the previous one, it can be used to describe the size, breadth or width of the network. But it can also be translated into an indicator of a small world situation.

Connectedness is quite a simple global property, and I already mentioned the importance of distinguishing networks where all components are connected from disconnected networks. On the contrary, clusters or community detection are widely used. It is especially useful for exploration. But be careful, it is tempting to take the result of a cluster calculation as a given. In some cases, it is interesting to compare these clusters with previously known groups (categories that do not depend on the structure obtained). In terms of translation, this notion of community is very directly related to the way in which the social sciences and humanities use the metaphor of the "network". The global or average clustering coefficients are complementary to the community detection. They give an idea of the entanglement/intrication and the presence of a more localized density.



With regard to local measures, the **degree** (number of neighboring nodes) is the simplest **centrality**, and the only one systematically used between the late 1950s and early 1970s, before the development of more diversified metrics. Its simplicity allows for a transparent translation: in a literary network, for example, it counts the number of times one character speaks to another.

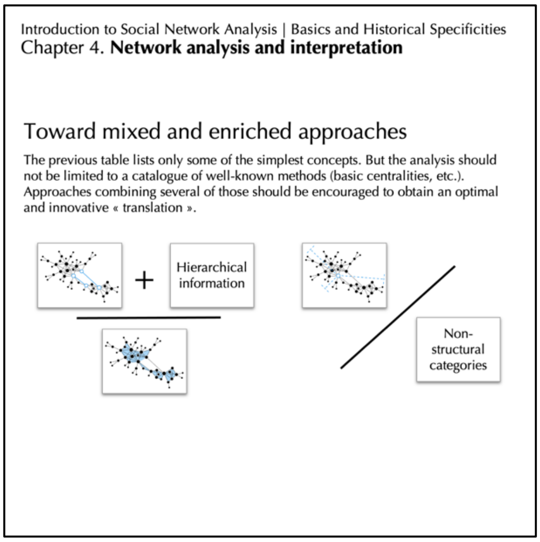
The notion of **betweenness centrality** disrupts the conception of what the “center” of a network may consist of. Its ability to reveal structural elements bridging large, immediately visible clusters makes it popular in the social sciences since the emergence of Granovetter’s concept of “weak ties. Betweenness is very closely linked to the notion of circulation: it counts the shortest paths to detect intermediate “bridges” or “key passages” capable of opening or locking certain parts of the network. Depending on applications, these are therefore both positions of power and vulnerable places.



The **closeness centrality** allows to highlight the “geographical” middle of the graph. In networks of a certain density and when they are not divided into several distinct communities, the closeness is generally fairly evenly distributed and allows a good translation of the notions of “center” and “periphery”.

For its part, the **eigenvector centrality** is quite complicated to translate since it works iteratively and is very much dependent on the structural context at short and medium range around a node. It was named “power centrality” by its author, it’s an indicator of “prestige” or “influence”, it qualifies a node’s environment while operating in cascade: a well-connected node gives its neighbors a part of its authority capital, and so on. It is therefore particularly useful when trying to analyze the hierarchy of the nodes in a graph.

The local **clustering coefficient** is also a metric making possible to analyze relationships at the collective level: it can be translated as an indicator of participation in a group (or, on the opposite, loneliness, solitude).



Conclusion

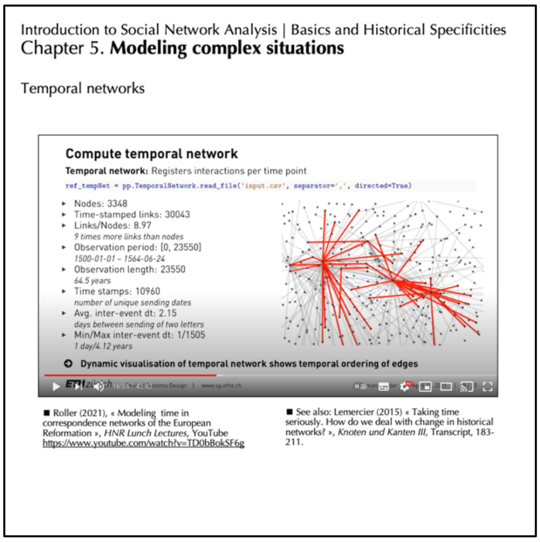
In this chapter, I have only mentioned a few possible interpretations, a few ways of translating these measures and observations into the language of historical analysis. But each of these approaches could be discussed in detail, whether it be visual analysis or metrics. Of course, the analysis should not be limited to a catalogue of well-known methods (basic centralities, etc.). In fact, approaches combining several of those should be encouraged to obtain an optimal and innovative “translation”. In this way, we could compare metrics or combine them to establish rankings. Furthermore, the enrichment of the networks by means of categories that are not dependent on the structure, like the gender of individuals in a social network or the discipline of projects in a scientometric analysis, allows to test translation and interpretation hypotheses by avoiding the blind approach of testing all possible graph metrics.

Chapter 5: Modeling complex situations

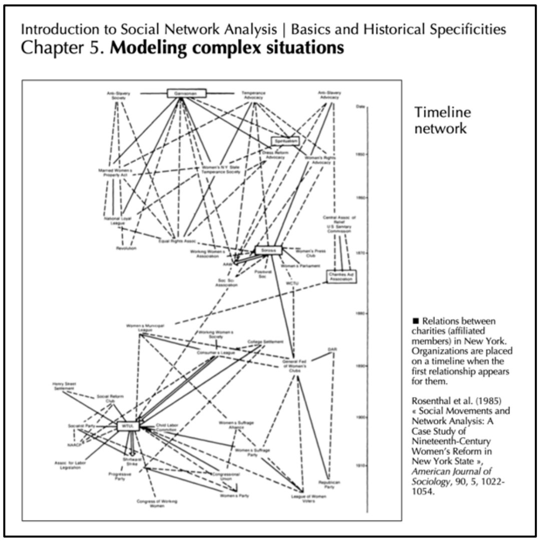
This last chapter of our introduction to network analysis applied to history is not a conclusion. Rather, it is an opening to some of the challenges that currently occupy our community. Even more than the previous chapters, it can only be a very brief evocation, as these questions are of crucial importance for the development of the discipline. I therefore want to evoke in particular the notion of temporality as well as that of multilayer system. First of all, the study of temporality is the foundation of historical studies; without the flow of time, there is no object of study. Secondly, the question of multilayer and scale is one the greatest promise of digital history: to be able to go from the local to the global and vice versa, to be able to take into account the effects of what happens down below on what happens above, etc.

Temporality

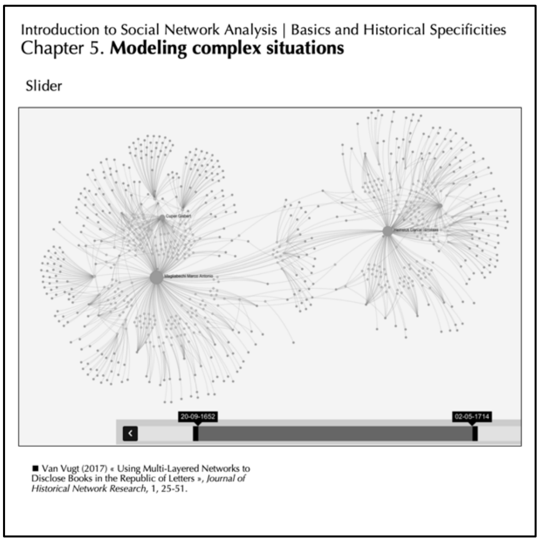
Taking temporality into account in networks is above all a question of visualization and exploration. Indeed, it is not conceptually complicated to imagine that relationships can have a beginning and an end, or a unique moment of existence. It is therefore quite natural to model networks that change over time, that take into account what happened before or what will happen later. This task is often facilitated by the rigor with which historians are accustomed: all information is supposed to be dated, etc. It is obviously more complicated when we are talking about interpersonal social networks that are not based on historical sources that can be formalized, but solutions are found. When does a friendship begin and end, for example?



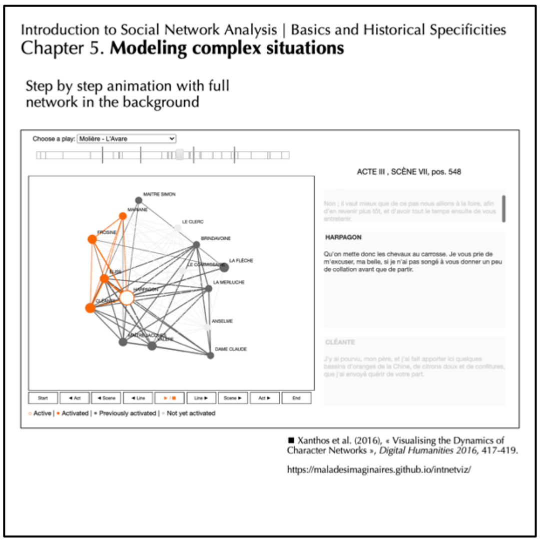
But technically, the implementation of these good ideas is really a problem. We lack the tools - and probably the creativity - to invent solutions that allow us to explore temporal networks efficiently. In fact, we are limited by our own ability to read a changing graphical representation, or to understand the evolution of a structure in a complex data set. I don't necessarily go so far as to say that this is a definitive cognitive limitation, but it probably has a lot to do with the fact that we are not used to seeing a network evolve. We'll look at a few ways to do this, but since an HNR Lunch lecture was held very recently on the subject, I'll refer you to this talk by Ramona Roller. I also refer you to Claire Lemercier's article "Taking time seriously".



The search for solutions to represent temporality in networks is not a recent phenomenon, as this 3-decade-old example shows. While we usually try to date the relations, this visualization shows that we can also date the vertices of the graph. Here we see the proximity of charities in New York at the end of the 19th century. The vertices are placed on a vertical timeline at the moment their first relationship appears. This does not account for the fact that most of these organizations are active for decades, but if what we are interested in is the triggering moment when they affiliate members of other organizations, this does the job. In fact, it is a kind of family tree, an elegant way to take into account the temporality without making the graph vary. At the cost of expensive visualization sacrifices, of course.

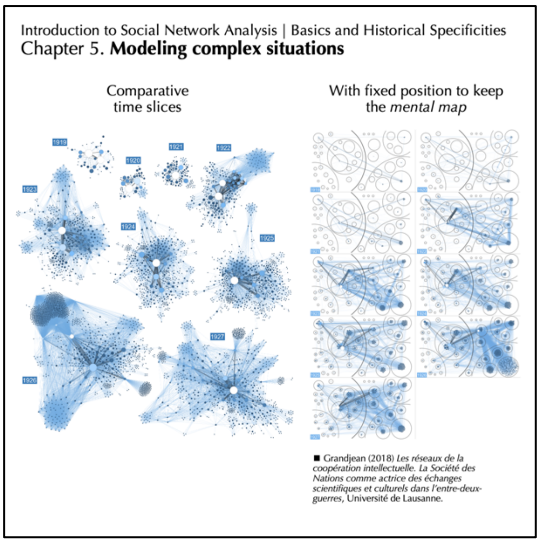


One way made relatively common by many visualization software is the addition of a slider that allows you to choose the portion of time you want to display. Efficient for exploration, it is less obvious to use for the average user: if you let the slider move by itself, as these systems sometimes propose by default, you are faced with a network that moves in all directions, jumps from one corner to another of the screen, like a living cell under the microscope. We therefore lose part of the overall view, and it is difficult to follow the trajectory of a particular vertex. Note that the question of maintaining the so-called "mental map" of the network is a very current concern of the visual studies community, there is a lot of work on this subject at the moment.



Another way to proceed is to compute the total network and then maintain the positions to create a shaded version on which we superimpose only the vertices and edges that are actually activated at the chosen time. This works quite well if, as in this example, we visualize metadata or textual annotations. But it is mainly a way to make this temporality accessible to a public, not really an exploration and research tool.

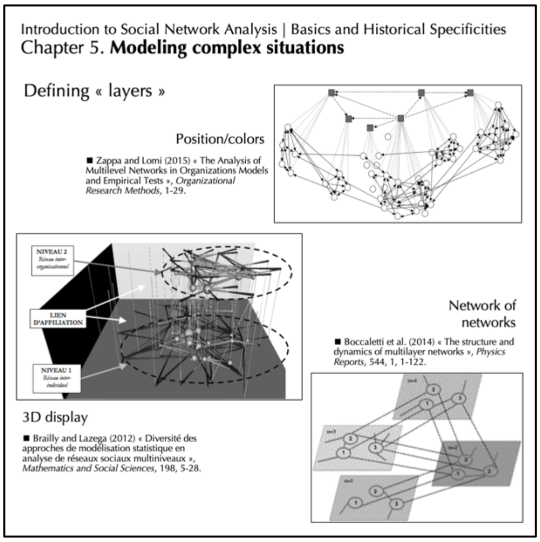
In many cases, we are left with static time slices. This is not always a very satisfactory solution, and it is often dictated by our means of publication: if we publish an article in a journal in paper or PDF format, we cannot include interactive visualization. But often, it is also a way to be sure that the graphical representation on which we base our interpretation is the same as the one we make available to our reader, which we can never be sure of with an interactive visualization. In this case, it is clear that these 1-year time slices do not allow us to "read" the network other than in terms of its overall structure. We understand the increase in quantity, we see clusters forming and deforming as well as certain vertices emerging from chaos, but it is impossible to delve into them more precisely.

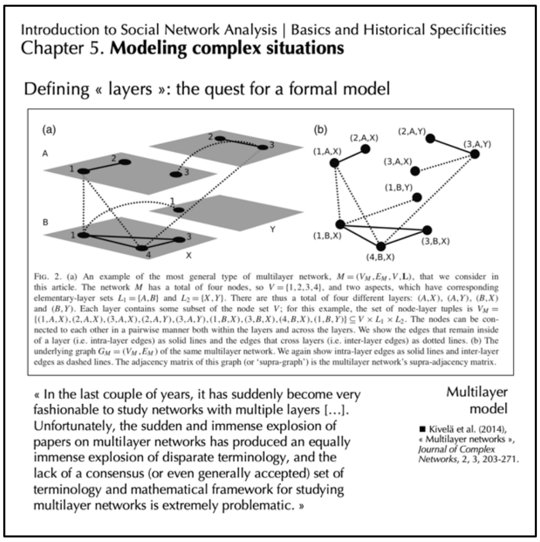


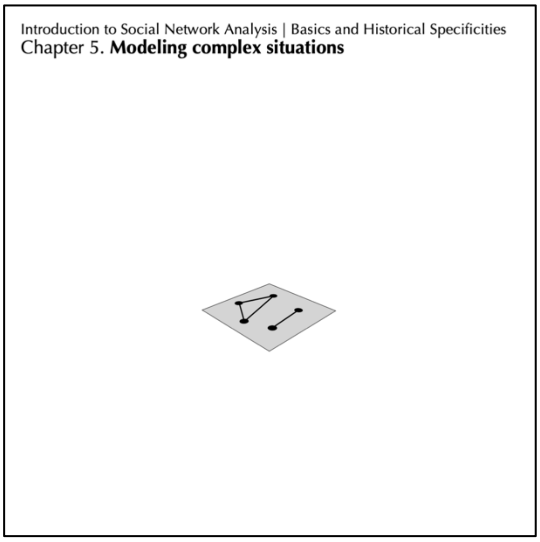
This is why an alternative visualization is proposed, on the right, a representation in which the position of the vertices is fixed according to a prosopographical classification, in order to observe only the evolution of the

edges. This makes it possible to maintain this "mental map" and, even if the density of the network leaves no room for a diagrammatic reading anyway, it is easy to see which group is activated at which moment, and in connection with which other group.
But still, even if playing with several spatializations of the same dataset allows to compare an object from several angles, the rendering of temporality is problematic.

Multilayer networks





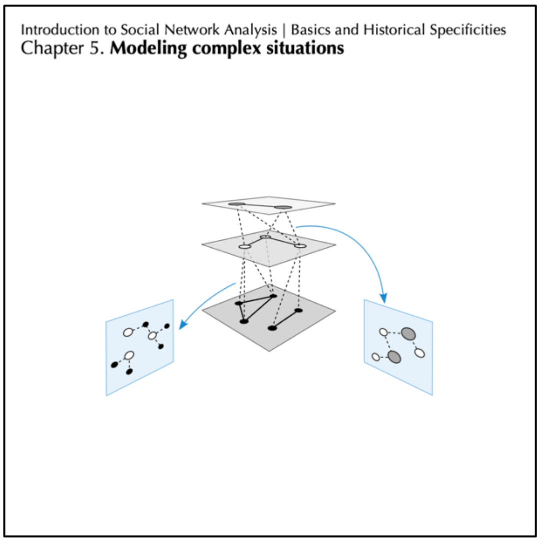


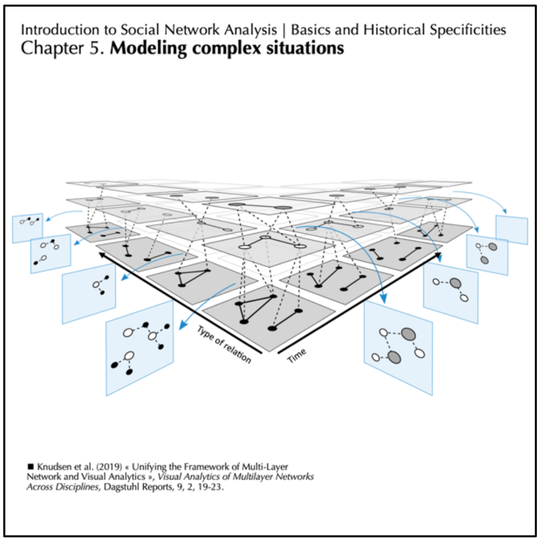
The second complex modeling issue I want to introduce here is the study of networks expressing themselves on several levels. We have seen that affiliation networks, 2-mode networks, already contain a form of verticality. But when we are interested in historical objects, there are rarely only two types of vertices and relations. The problem we have when dealing with issues involving verticality is that we are technically and conceptually limited by the fact that we generally express ourselves in two dimensions and are unable to think in more than three dimensions. As in the examples here, we will therefore set up tricks to try to account for these different layers. This can be done by using different colors, positions or shapes for vertices and edges, for example. But this only works with small networks. Or we can create a false perspective in 3 dimensions, with relief effects to make clear the planes on which the elements are placed. But this false 3D is very quickly unreadable and even an interactive visualization in three dimensions, that we could turn in all directions or even explore with a virtual reality device would not solve our global vision problem. Otherwise, we can say goodbye to the dream of representing everything and use more conceptual visualizations, which allow us to show how our modeling is organized, on well-arranged layers. Then, no more visualization, the data is modeled, and the machine does the analysis work without visual output, or with outputs located on one or more of these layers but not all at the same time, and without the edges between the layers.
By introducing this article by Mikko Kivelä et al, a landmark paper in the formalization of multilayer analysis, I know that I am going far outside the spectrum of an introduction to network analysis, but I think it is important to make you understand that we are in something that is happening now. And it needs to be given importance because our historical topics require complex modeling like this if we are to go beyond the obvious.
Thinking of our historical object as a multi-layered system does not have to be complicated. We can keep things very simple, but it is essential to be aware of potential layers, other facets with which to look at our data set differently. Most of the time, we start from a simple layer,

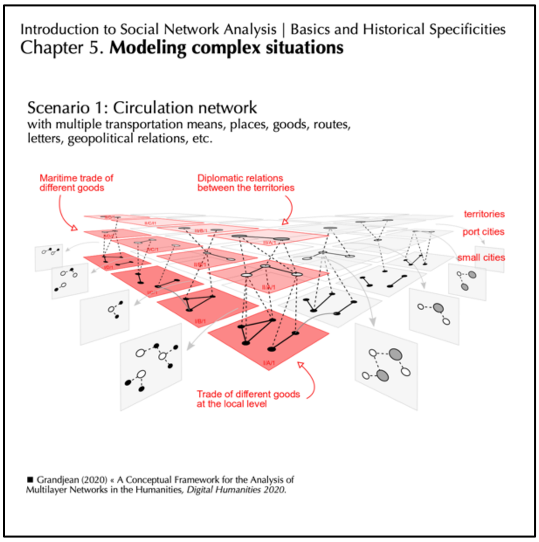
composed for example of individuals and interpersonal relationships. Often, these individuals also have relationships with entities located in another layer. For example, institutions with which they are affiliated. There are relationships on each of these layers, but also between them. In an affiliation network, it would be precisely the 2-mode network that we will project next. But we can imagine other levels, with other layers and relationships between them. In our case, this could be state entities, above the institutions.

And this multi-layered model can vary depending on several parameters. We can for example add a temporal dimension to it and observe the variations of the system along this axis. Or we can look at different types of relationships, or different contexts. For example, the world of social relations, the world of economic relations, etc. In short, even if it looks like a conceptual monster, it is simply the reasoning that we already do when we prepare our network analysis. And let's be clear, the goal is not to create such a representation by mixing dozens of layers but to choose precisely which layer or group of layers we want to study. In fact, this conceptual tool of multilayer network analysis should allow us to develop analysis scenarios.

For example, we could work on a network for the circulation of goods and people between port cities. We would be interested in the level of local exchanges, then maritime traffic, then diplomatic relations between territories. The whole thing could be divided into several facets, depending on whether we are interested in different types of goods, etc. To use an example that has been used a lot previously, we could look at an affiliation network. But rather than working on the individual level, we would focus on several institutional layers, one level with organizations and one level with committees involved in these organizations, for example. This would make it possible, for example, to see if the organizations have relationships with each other that correspond to the relational work they do at the meso-organizational level. Or, from a social history perspective, we could model interpersonal relationships that evolve over time. We will then be able to see if the evolution at the personal level is the same if we look at this social microcosm at the level of the groups in which individuals participate.







Conclusion

So, all these questions are not really answered at the moment. But I hope that these few quick introductory chapters, while not giving you the technical and conceptual tools, have at least given you something to think about and encourage you to dig deeper.

References

Anscombe (1973) « Graphs in Statistical Analysis », *The American Statistician*, 27, 1, 17-21.

Boccaletti et al. (2014) « The structure and dynamics of multilayer networks », *Physics Reports*, 544, 1, 1-122.

Brailly and Lazega (2012) « Diversité des approches de modélisation statistique en analyse de réseaux sociaux multiniveaux », *Mathematics and Social Sciences*, 198, 5-28.

Düring et al. (2015), « Interactive Networks for Digital Cutural Heritage Collections – Scoping the future of HistoGraph », *Engineering the Web in the Big Data Era, ICWE 2015*.

Euler (1736) « Solutio Problematis ad Geometriam Situs Pertinentis », *Opera Omnia*, 7, 128-140

Ginalski and Heiniger (2016) « Les réseaux de la réforme sociale à Genève autour des années 1900 », *Histoire et Mesure*, XXXI, 1, 85-121.

Grandjean (2017) « Analisi e visualizzazioni delle reti in storia », *Memoria e Ricerca,* 55, 2, 371-393.

Grandjean (2018) *Les réseaux de la coopération intellectuelle. La Société des Nations comme actrice des échanges scientifiques et culturels dans l'entre-deux-guerres*, Université de Lausanne.

Grandjean and Jacomy. (2019), « Translating Networks: Assessing Correspondence Between Network Visualisation and Analytics », *Digital Humanities 2019*, Utrecht.

Grandjean (2020) « A Conceptual Framework for the Analysis of Multilayer Networks in the Humanities*, Digital Humanities 2020.*

Jacomy (2021) « As a technical mediation, Gephi shifts your goals (and vice-versa) », *Reticular*, https://reticular.hypotheses.org/1866

Kivelä et al. (2014), « Multilayer networks », *Journal of Complex Networks*, 2, 3, 203-271.

Knudsen et al. (2019) « Unifying the Framework of Multi-Layer Network and Visual Analytics », *Visual Analytics of Multilayer Networks Across Disciplines,* Dagstuhl Reports, 9, 2, 19-23.

Lemercier (2015) « Taking time seriously. How do we deal with change in historical networks? », *Knoten und Kanten III,* Transcript, 183-211.

Milgram (1967) « The Small-World Problem », *Psychology Today*, 1, 1, 61-67.

Moreno (1934), *Who Shall Survive? A New Approach to the Problem of Human Interrelations*.

Padgett and Ansell (1993) « Robust Action and the Rise of the Medici, 1400-1434 », *American Journal of Sociology*, 98, 6, 1259-1319.

Roller (2021), « Modeling time in correspondence networks of the European Reformation », *HNR Lunch Lectures*, YouTube https://www.youtube.com/watch?v=TD0bBokSF6g

Rosenthal et al. (1985) « Social Movements and Network Analysis: A Case Study of Nineteenth-Century Women's Reform in New York State », *American Journal of Sociology*, 90, 5, 1022-1054.

Van Vugt (2017) « Using Multi-Layered Networks to Disclose Books in the Republic of Letters », *Journal of Historical Network Research*, 1, 25-51.

Xanthos et al. (2016), « Visualising the Dynamics of Character Networks », *Digital Humanities 2016*, 417-419.

Zappa and Lomi (2015) « The Analysis of Multilevel Networks in Organizations Models and Empirical Tests », *Organizational Research Methods*, 1-29.