1

Temporal and spatial networks

Our primary concern is *understanding* both large temporal and large spatial networks in ways going beyond simple general descriptions of their structures. For the former, doing this amounts to discerning the structure(s) of such networks as they develop over time, and grasping the social forces driving these changes. For the latter, it involves understanding spatial social patterns and the processes by which they were generated. For both network types, these two broad tasks – delineating structures and understanding their formation – go hand in hand: doing one without the other leaves our understanding of these networks incomplete. However, in order to understand the impact of social forces, it is necessary to know the structure(s) of networks. We focus, initially, on outlining foundational network concepts in Chapter 2. A detailed presentation of methods for analyzing citation networks is included in Chapter 3. In the remaining chapters, we study how temporal networks change and social phenomena are distributed over spatial networks. We provide substantively based interpretations of the results we obtain. As is usually the case, for us, creating these understandings was an iterative process where empirical results led to substantive understandings which, in turn, triggered further analyses. We report results of these analytic sequences but without reporting the iterations.

1.1 Modern social network analysis

Freeman (2004) argued that four features define *modern* social network analysis (SNA). In a slightly expanded form they are:

1. SNA is founded on a ‘structural intuition’ regarding social ties linking social actors. This motivates the study of the social networks formed by these social ties when they form *coherent* wholes.
2. ‘It is grounded in *systematic* empirical data (emphasis added).’ Implicitly, network data must be meaningful for studying social networks: not all social network data sets are useful.

Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution, First Edition. Vladimir Batagelj, Patrick Doreian, Anuška Ferligoj and Nataša Kejžar.

3. 'It draws heavily on graphical imagery' to represent these social networks and their salient features in useful ways. Visualization of these features is useful both for displaying results and for suggesting further avenues of inquiry.

4. 'It relies on the use of mathematical and/or computational models.' This dual reliance has grown even stronger since 2004.

We add the following three items:

1. Fully *understanding* social networks in time and across space requires a concern with substance.

2. When studying the operation of social *processes* creating, sustaining, and dissolving social networks, Doreian and Stokman (1997), the relevant network data *must* be temporal. Intuitively, a temporal network has units and relational ties distributed through time.

3. Given that most social networks are conditioned by the contexts within which they exist, ignoring these contexts imposes major constraints on understanding network phenomena. One contextual feature is the geographic space within which these networks are located. Spatial networks have units and relational ties distributed across geographical space.

While substance can never be ignored safely, we note that many networks have been studied without considering time. Other networks were studied while ignoring space. Quite often, neither time nor space had relevance for analyzing network data. This has changed dramatically in recent years with considerable attention being devoted to both space and time when studying social networks. Consistent with this new emphasis, the networks we consider here involve time or involve space and, occasionally, both. In the main, we focus on temporal networks.

Building upon the above seven items, our study of temporal networks and spatial networks is informed by four working assumptions:

1. Social networks form through the operation of *social* processes. These processes have direct relevance for studying networks, implying that substantive ideas really matter. In turn, the contexts within which social networks are generated are crucial for understanding network creation and the consequences they have for the people, groups, organizations, states, and nations located in them.

2. As Freeman noted, computation has been crucial. However, *practical* and *sound* computational methods are required for detecting useful structural patterns in networks. Developing these methods is necessary, even mandatory. Ideally, computational methods are informed by substantive concerns. However, we have no objection to developing methods for their own sake. Even so, the use of methods developed in this fashion requires some justification in terms of both substance and relevance, at least as far as understanding social network processes is concerned. Methods are more useful when coupled to the substantive issues for which analyses are performed.

3. Temporal network data have to be *meaningful* in terms of both social substance and social contexts. This implies that temporal network data need to be selected carefully in order to be relevant substantively. The same arguments hold for spatial networks.

4. Coupling substance, context, methods, and data is most effective when these items are combined into a single coherent framework.

1.2 Network sizes

We define the terms small, large, and huge for network sizes in Section 2.3. The networks discussed by Freeman (2004) are small. Indeed, for many decades, social network analysis dealt primarily with small or very small networks. This was driven by traditional ways of collecting data and by the technical constraints on the collection and analysis of social network data. However, since the 1990s, large networks (having from thousands to many millions of units) have become abundant, for which information technology (IT) has been particularly important in assembling these data. This development has serious implications both for visualizing networks and for implementing computational models. Many of the earlier traditional computational methods useful for studying small networks are now completely impractical for analyzing large networks. As a result, developing new practical computation methods has become essential. We focus on some newly designed computationally feasible methods for handling these networks and present the results of using them while being attentive to substantive concerns.

1.3 Substantive concerns

At face value, the only feature common to the networks studied here is being large. While the book title emphasizes this, its most important word is ‘understanding.’ The datasets we use were selected to cover different substantive domains, to have different sizes, and to be characterized by different structures. We have no single ‘cookie cutter’ method suitable for all temporal networks because of these differences. However, two methods used repeatedly in our analyses are line islands (used in Chapters 4–6) and clustering symbolic data (used in Chapters 5 and 8). We focus briefly on substantive concerns here and more fully in Chapters 4–9 where different combinations of methods, based on the foundations laid out in Chapter 2, and developed in Chapter 3, are used. Methods employed only in single chapters are presented therein.

1.3.1 Citation networks

Chapters 4–6 consider three distinct citation networks. While citation networks may seem to be the ‘same’ in their general structure, with later units (documents) citing earlier units, these three networks differ greatly with regard to their sizes and network structures in addition to substance.

1.3.1.1 Scientific citation networks

Hummon et al. (1990) noted three features of science: 1) ‘Science is a cumulative venture where each new discovery or development depends on some prior work;’ 2) ‘The products are generated at the research fronts of specialty fields’ where the products are scientific productions; and 3) ‘The resulting written record, in the form of citation networks, left as research fronts move on, contains valuable information for understanding the processes of science.’ The small citation network they studied concerned the centrality literature between the initial task-oriented group experiments of Bavelas (1948) through Freeman's (Freeman, 1979) systematic statement of three distinct operationalizations of centrality. This network, while complete, was small with only 119 scientific productions. Since then, centrality became accepted widely as one of the most important concepts in social network analysis. Certainly, it is one of the most frequently used ideas: in short, this literature has exploded. We were curious about the extent to which this literature grew and the structure of the resulting citation network.

Although, as documented by Freeman (2004), SNA has been in existence for over a century, physicists have developed a recent interest in networks within the rubric of what has become known as 'network science.' See, for example, Watts and Strogatz (1998), Newman (2001), Barabási (2003), and Newman et al. (2006). This interest was triggered, in part, by the availability of large networks obtained readily via electronic methods. Despite some notable exceptions, physicists have tended to ignore most of the prior social network literature while claiming the creation of a 'new' field resulting from their endeavors. In the memorable phrase of Bonacich (2004), this was (seen as) 'the invasion of the physicists.' We were curious as to the structure of the citation network of the SNA citation network following this 'invasion.' Some results of our analyses are laid out in Chapter 4 together with a narrative concerning the ways citation network structures are developed and some of the institutional forces involved in this process.

We study two scientific citation networks. One is restricted to the centrality literature while the other, more broadly, is the SNA literature. In the main, the former is located within the latter. However, as centrality was initially a very narrowly focused concept, there is interest value in looking at it especially as the concept has been applied in many substantive domains. Of additional interest is that we have learned (see Chapter 4) that the concept has been formulated and found valuable in areas remote from SNA. Indeed, tracking how a technical concept has been applied in different substantive areas adds to the value of looking at the citation network for centrality. One implication of the results reported in Chapter 4 is that the centrality literature in no longer located fully within the SNA literature.

We were interested also in the nature of the linkages between the traditional SNA literature, as seen by social network analysts, and the network science literature. To the extent that different fields merged in pursuit of studying social networks, it is reasonable to expect flows of ideas between them. A rival expectation is that the fields partially diverged while cleaving to their own conceptual frameworks, network interests, and methods.

As described below, both the patent citation network and the Supreme Court citation network were 'cleaned' explicitly (for patents) or implicitly (for the Supreme Court). This cannot be done fully with academic citation networks. Once publications are in the literature, these works remain. Of course, never-cited works can be removed as a part of a data analytic process removing them. Even so, apparent dead-end lines of work remain, a topic we do not pursue here.[1]

1.3.1.2 Patent citation networks

Patents are legal devices attempting to confer some protection of intellectual property rights for the inventors of new technological items. They have been seen as particularly interesting

[1] It remains an open question whether approaches simply dying out (for example, 'functional theories' in sociology), or because of scientific revolutions in their fields (Kuhn 1970), can be tracked in citation networks.

regarding their role in the study of innovation and technological change. Economists have long sought to link technological innovation to economic change and development. Examining patents and their role in triggering technological change has been a part of this effort. However, our concern differed: we sought an understanding of the temporal patterns of patents citing earlier patents. This includes when and how inventions protected by earlier patents become useful for later inventions and patent applications. More precisely, we sought an understanding of the *time scales* regarding how earlier inventions became useful for later inventions. We learned that there are at least four distinctive patterns to these time scales. Also of interest is the influence patterns between broad technological areas (they are significant), how they vary over time, and how specific technologies decline temporally, to be replaced by others in specific eras. We document some of these changes in Chapter 5.

We learned that inventors, or their proxies, applying for patents enter complex technical *and* legal arenas. The technological arena alone is intellectually complex as is documented in Chapter 5. As a result, applications for patents to protect inventions trigger a stringent review process performed by experts in the United States Patent and Trademark Office (USPTO), at least for patents issued in the USA.[2] The result – in addition to the granting of patents for inventions or deciding specific inventions cannot be patented – is a complete citation network that is 'efficient' in the sense of citations being made *only* to *all relevant* earlier patents. The institutionalized review process of the USPTO is the first phase for creating a cleaner citation network.

In addition, when existing patents are revoked, they and all citations to them, as well as citations from them, are removed from the patent citation database by the USPTO. As a result, only genuinely useful patents remain in the available data. This is particularly important because, in essence, the boundary problem (Laumann et al., 1979) – one posing problems for many social network studies – is solved completely: these patent data are the cleanest citation data we examine. Intuitively, the boundary problem for networks is simple to state and has two basic components. One is the exclusion of relevant data points (units) and all of the (real) network ties involving them. The other is the inclusion of data points that are not part of the network being studied but are included in the network along with their network ties. Both types of errors have great potential for distorting networks and analyses of them.[3]

However, more than technological issues are involved when considering patents. Many economic actors seek to capitalize on the patents they have by creating production processes and services, new physical products and new substances for economic gain. Conflict over them leads to legal issues involving both apparent and real patent infringements when other economic actors produce similar or identical products for sale. The parties involved in these disputes either attempt to protect inventions covered by patents they hold or challenge the legitimacy of patents already granted to others.

Conflicts over inventions and patents often lead to court cases. Some reach the US Supreme Court, where the Justices on this court weigh in on technological matters and the patentability of inventions. We sought an understanding of patent citation phenomena in technological contexts defined by the USPTO. The primary tools used were line islands (see Section 2.9) and clustering symbolic data (described in Section 3.10). The results of

[2] No doubt, this is true elsewhere, but our data are for US patents.

[3] There is also the separate measurement problem, even if the boundary problem has been solved, of erroneously excluding network ties, erroneously including them or recording ties inaccurately. These problems are solved also for the patent data.

using these methods for the patent data are presented in Chapter 5. The involvement of the Supreme Court in these technological matters added an unexpected (for us) connection between Chapters 5 and 6. The technique of identifying line islands was important for considering the role of the Supreme Court in evaluating rival claims over patents.

1.3.1.3 A US Supreme Court citation network

Fowler and Jeon (2008) compiled an extensive citation database for US Supreme Court (SC) decisions citing earlier SC decisions based on the content of their majority opinions and the citations they contain. At face value, this data set is complete also in including all decisions in a specified and very long period (1789–2002).[4] These citation data differ from the patent citation data in at least one important respect. There are no constraints on Supreme Court Justices writing opinions regarding which earlier decisions they cite as precedent nor on which earlier decisions they choose to ignore as precedent. This is a luxury unavailable in patent applications where, as noted above, relevant, and only relevant, citations to earlier patents are permitted. This adds an intriguing wrinkle for understanding this temporal network.

Fowler and Jeon's primary focus was the evolution of precedent as a legal concept: earlier decisions inform (and therefore constrain) later decisions. For analyzing precedent, the frequency of decisions being cited, especially by other salient decisions, takes centre stage. We take their results as a given.

Our interest takes a complementary form: we focused on subsets of Supreme Court decisions forming coherent *parts* within the overall citation network. In terms of methods, we did not focus on computing measures for single decisions but examined sets of decisions instead. We approached this by considering the extent to which earlier decisions are *co-cited* by later decisions. The rationale behind this interest was driven by a key intuition: decisions cited *together* have import by having common substantive or legal principles (or both) holding them together. To qualify for this additional closer scrutiny, earlier decisions have to be co-cited *frequently* (by pairs of subsequent decisions).[5] The primary method used in studying these coherent parts of the Supreme Court citation network was identifying line islands, a procedure described in Section 2.9.

By definition, never cited earlier decisions – of which there are very many – can never be co-cited. Similarly, decisions citing no earlier decisions cannot contribute much of interest regarding precedent nor for considering co-citating. Decisions neither citing other decisions nor receiving citations are easily removed. In effect, doing this helps 'clean' the US Supreme Court citation network to achieve implicitly an effect similar to the results due to the process enforced by USPTO's review process. Removing these isolated decisions, having no historical relevance, prunes the citation network. While this serves our purposes very well, this may affect the fitted distributions of measures computed for single decisions.

[4] On looking closer, and examining many decisions, we learned that some decisions were omitted from the Fowler and Jeon dataset. While we have inserted the missed decisions that we located, there is no guarantee that the list of decisions we studied is complete. Indeed, all large datasets contain errors as Fowler and Jeon note for the dataset they created. However, it is very close to being complete for *relevant* decisions. Never cited decisions are the most likely to be overlooked. These omissions are unimportant because they have no relevance in this citation network. A potentially much more serious data recording problem was unearthed, and this is examined in Section 6.6.

[5] It is possible to examine pairs of decisions in terms of frequently *co-citing* earlier decisions, a line of inquiry not pursued here.

The Supreme Court network can be looked at as a stand-alone entity to be studied by itself. However, the Supreme Court is only one of three ‘top’ branches for governing the US. The other branches are the President and the US Congress (made up of the House of Representatives and the Senate). In principle, the three branches were created as independent entities designed to constrain each other through the much discussed ‘checks and balances.’ As we show in Chapter 6, there were periods when the three branches acted in concert while being in sharp conflict at other times. Both of these contextual conditions have relevance for understanding the actions of the Supreme Court. The same applies more generally with regard to historical contexts for all social networks.

At face value, the task of the Supreme Court is simple: its members use the Constitution as the foundation for establishing the legitimacy (or not) of laws and the appropriateness (or not) of decisions rendered by lower courts. Unfortunately for this naive view, the Constitution is profoundly ambiguous, a built-in feature reflecting the greatly divergent positions of the rival parties and political interests represented by those involved in drafting this document. These deep conflicts have never been resolved. Indeed, it seems impossible to resolve them, and the vague (ambiguous) language of this constitutional document papered over these differences in order to get enough signatures to it. The deep political conflicts did not end with the signing of the Constitution, and the subsequent ratification process in the separate states was deeply conflictual within them. As a result, the whole judicial system, and the Supreme Court in particular, as authorized by the Constitution, was ‘political’ from its inception and has remained so. This alone implies caution in taking the Constitution as an ‘objective’ document free of biases and contradictions and viewing the resulting citation network as a simple record of the processes leading to its decisions. It cannot be studied as a network detached from broader social contexts.

Further, this citation network covers by far the longest time interval (more than two centuries) of the citation networks we consider. In this long period, the USA experienced great economic, political, and social change. In short, the context within which the Supreme Court operated changed dramatically over time. This suggests that the resulting citation network cannot be studied solely as if it were simply just another citation network. The changing contexts within which decisions were made matters greatly for understanding it. Some results using line islands in the Supreme Court citation network are presented in Chapter 6. With these identified islands, close attention was paid to the changing history of the USA and the Court in understanding both the citation network’s structure and the more important actions of the Court. This included attention to ‘accidents’ (sudden deaths of Justices and Presidents, plus unanticipated electoral outcomes) and changes in the composition of the Court over time.

1.3.2 Other types of large networks

The US patent network is acyclic (a concept defined in Section 3.1). Both the Supreme Court network and the scientific citation network lack this feature which creates technical problems for analyzing them. Fortunately, such networks can be transformed easily to an acyclic form by using the methods described in Chapter 3. In order to obtain completely different types of temporal networks, we focused on different substantive issues where the techniques useful for citation networks were not relevant. These other networks are not acyclic networks and *cannot* be studied as such. Including them was important because ‘large temporal networks’ can take many forms, including some we do not consider here.

1.3.2.1 The movement of football players across the globe

For the first alternative type of network, we focused on the movement of football (soccer) players as they traveled across the globe to play football. Football players, almost from the formal inception of the game, following its codification, moved between clubs and countries. In doing so, they created club-to-club networks because clubs had to agree on the movements of players between them. We conceive of football players' careers fundamentally as *movements between clubs*.[6] These movements define and create basic social networks linking clubs and, secondarily, countries. While this is our primary interest, we use these movements also to examine some common presumptions about 'the beautiful game.' Alas, no single systematic nor reliable dataset comprising these movements exists. So, we constructed one, albeit with a geographically restricted focus. The details of this database are described fully in Appendix A.4, along with the many difficulties encountered while constructing it. These data are unique. We recognize that this claim can be made for any dataset but this one was created by combining information from over a thousand different sources.

Recognizing that the game is played in over 200 countries, we restricted attention to just one country because one of us (Doreian) has had a long-term interest in football played in England. Our data were defined by *all* of the football players who appeared in the English Premier League (EPL) during its first 15 seasons (1992/3 through 2006/7).[7] This temporal restriction was, primarily, a practical issue – although we did track these players through to the end of 2012. To set the broad background for the analyses that we pursue, we describe football in England in Chapter 7 as a local institutionalized representation of the so-called world game. Additional reasons for focusing on football in England, beyond familiarity, stem from its unique history, one having major impacts on player movements within and to England. This is part of the context for these player-induced networks.

English officials administering the game were, and continue to be, acutely aware of the game having been invented in their land. Of course, the game quickly spread to many other places but this diffusion was ignored largely by these administrators for close to a century. As a result of the assumed historical primacy of their legacy and a presumption of English 'superiority' regarding 'their' game, these officials assumed that they had little to learn from developments in football at other places on the globe. Indeed, they attempted, with great success, to keep (most, but not all) 'foreigners' out of the game as well as 'foreign' conceptions of strategy, tactics, and styles for playing football.

This fundamental restrictive control of the game in England was shattered by a series of court cases. One was indigenous to England (resulting in the Eastman decision as described in Chapter 7). Far more importantly, the European Union (EU) ruled on labor practices for all its members in the Bosman Decision, also described in Chapter 7. The so-called 'transfer and control' system by which football clubs controlled their players did not come close

[6] Other conceptions of careers focus on player and club performances on the field in terms of goals scored, defensive plays made, appearances for national teams, club victories, and trophies won. This conception of careers in terms of success has little interest for us beyond operationalizing temporal sequences of club success (overall ranks) to characterize player careers as sequences of moves between the clubs, in nationally and internationally stratified systems, for which they play. This is described more fully in Chapter 8.

[7] For this study period, the EPL is an accurate label. Much more recently, Swansea City and Cardiff City, both located in Wales, were promoted to this top league. The term Premiership is now used frequently rather than EPL as a label. An alternative label is BPL, presumably for 'British Premier League,' although this usage ignores the Scottish Premiership.

to conforming to EU rules, especially with regard to people moving between its members to find employment. These court decisions changed forever the relations between football players and the clubs for which they played. Suddenly, players had greater (but not complete) freedom to move between clubs (and between countries). Some of this changes are described in Chapter 7. We were curious about the resulting network *patterns* of player movements and what this implied for the organization of football.

Much has been written about player movements *to* England, the nature of football in England, and the impact of TV money flooding into the game since the late 1980s. These include arguments about the impact of this flow of foreign football players on football in the place where it was invented. Some concern the widely held belief about the EPL being the 'best' league on the world and how beneficial this has been locally. Other arguments claim that these flows were disastrous for *English* football, especially at the national level. Clearly, skepticism is merited regarding these arguments. Kuper and Szymanski (2009) exemplify this skepticism, a stance which prompted some of our analyses. As conventional (assumed) wisdom often rests on ignoring relevant information, our interest was piqued by these rival claims.

We examine some claims about modern football using the network of players moves that we constructed, together with some ancillary data. Our results and conclusions are reported in Chapters 7 and 8. Some of the hypotheses were confirmed and some failed while others turned out to be untestable.

1.3.2.2 A large US spatial network

As an example of a large spatial network, we examined the network defined for all US counties in the Continental USA.[8] We had two motivations for considering these data. One was substantive while the other was methodological. The Continental USA has 3111 counties. Pairs of counties are linked through sharing a common border. This adjacency in geographical space defines an unambiguous spatial relation linking counties. The Continental USA is divided also into 48 states each made up of counties. Each state has its 'own' history. In these histories, events and outcomes are described frequently as being unique to the 'proud' history of each state. Yet, on the ground, the boundaries between many pairs of states are evident only by signs marking them.[9] Certainly, social processes operate across the boundaries between these large aggregates. Attempts to understand these broad social processes need to move beyond state boundaries.[10]

There have been two broad approaches to characterizing the spatial distribution of the large social, economic, and political diversity within the USA. One attempts to map broad contiguous areas of the landscape within which greater homogeneity is thought to exist. Two examples of doing this are Garreau (1981) who defined and delineated Nine Nations covering the USA, Canada, Mexico, and the Caribbean Islands, and Woodard (2011) who argued for there being eleven such nations. Their general argument has appeal, with both authors assembling considerable qualitative evidence in support of their theses. While there

[8] Hawaii and Alaska were excluded, for obvious reasons.

[9] Rivers are one of the exceptions when they form clear boundaries between states. Occasionally lakes do this.

[10] The same argument can be made with regard to counties. However, as we claim in Chapter 9, counties represent a reasonable compromise between large heterogeneous areas like states and very small potentially more homogenous local areas for which systematic data do not exist.

are some commonalities to the two sets of nations they defined, there are also considerable differences. This alone merits a closer examination of their detailed delineation of nations within North America.

A second broad approach is exemplified by Chinni and Gimpel (2010) who eschewed geography during their detailed data analysis. After assembling statistical data for counties, Chinni and Gimpel clustered them using these constructed variables. They then plotted these clusters of counties in geographical space to describe a ‘patchwork’ nation with very different patches distributed across the nation and within states.

It seemed reasonable to seek a middle ground between focusing solely in large contiguous regions and focusing solely on the attributes of the units (counties) located in geographical space. The general problem is one of clustering units based on measured variables while being attentive to relations among the units. Although it was not proposed initially for dealing with spatially distributed data, one method for doing this – clustering with relational constraints – was proposed by Ferligoj and Batagelj (1982, 1983). It clusters units based on a set of measured variables, consistent with the approach of Chinni and Gimpel (2010), while constraining cluster memberships according a relation linking the units being clustered. The obvious relation in the US context is the spatial adjacency of counties. However, the method, as initially formulated, is impractical for any large network, especially for one as large as this spatial network. The technical concern motivating our analysis was establishing a practical computational method for networks of this size while remaining faithful to the core conception of clustering with relational constraints. The newly developed algorithms and the results of applying them are described in Chapter 9.

1.4 Computational methods

We develop extensive formal foundations for the methods we use in Chapter 2. It serves as a preliminary introduction to graph theoretical representations of networks. We then extend this systematically to deal with temporal networks, as defined in Section 2.2 and detailed in Chapter 3. Our focus on large networks is driven, primarily, by the intriguing computational difficulties of handling them efficiently. The notion of a ‘large network’ is defined in Section 2.3 to include networks with many millions of units. In terms of computation, the central workhorse for the empirical results we present is Pajek (Batagelj and Mrvar, 1998; de Nooy et al., 2012). Indeed, Pajek was designed *explicitly* for analyzing large networks efficiently.

We do not claim that Pajek is the only useful software for this purpose: it was simply the one we chose for our computational efforts when analyzing large networks. It served our purposes well. Doreian (2006) noted that Pajek is not a ‘one button’ set of routines. Instead, the results obtained from most of the analyses we present were completed by *combining sets of commands*. This design feature of Pajek facilitates great flexibility. However, it also requires users to understand the program’s logical structure. Given this, we include Pajek commands wherever they are appropriate so that readers can do the analyses leading to our results for themselves if they wish – either on the data we used or with data of their own.

Figure 1.1 shows the primary (initial) dialogue box for Pajek. There are two distinct listings of objects. The general concept of a network is that it is composed of vertices (representing units) and lines representing relational ties between units. These terms are defined fully in Chapter 2. On the left (reading from the top), is a column listing objects: Networks (described by vertices and lines); Partitions (assigning values to units to split them into clusters); Vectors (assigning numerical values to units); Permutations (to rearrange the order of units as they are

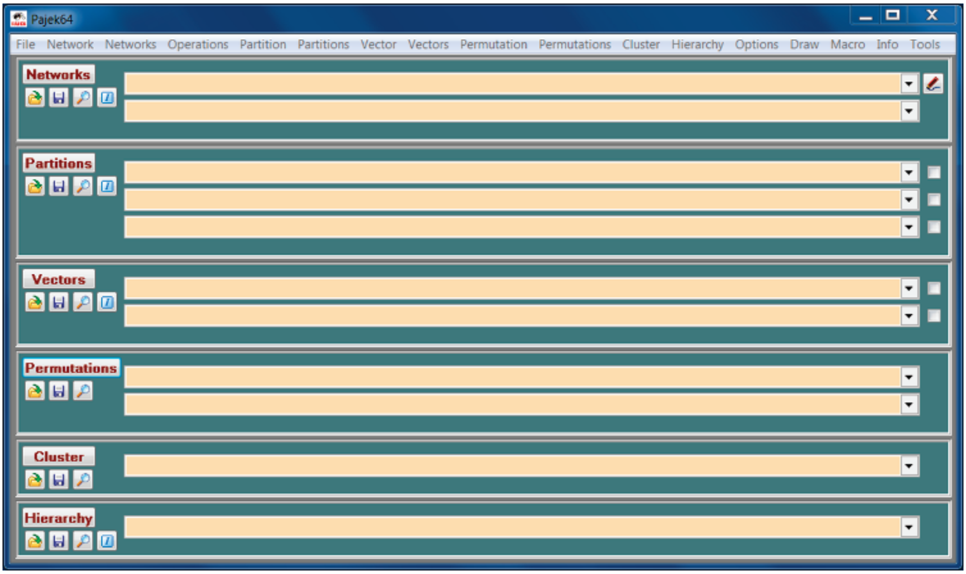

Figure 1.1 The Pajek main dialogue box.

stored in files); Cluster; and Hierarchy. The icons for each of these objects are used (reading from the left) for reading objects saved as Pajek files, saving objects as Pajek files, examining the contents of objects, and getting summary information about these objects.

Across the top of the main dialogue box, the listed objects are (reading from the left) File, Network, Networks (for handling more than one network), Operations (using different combinations of networks, partitions, vectors, permutations, and hierarchies), Partition, Partitions (for using multiple partitions), Vector, Vectors (for using more than one vector), Permutation, Permutations (for utilizing multiple permutations), Cluster and Hierarchy. Clicking on these icons produces drop-down menus with more detailed data analytic options. These icons on the top row are followed by Options, Draw (for visualizing networks), Macro, Info, and Tools (for exporting selected information to other programs, including R and SPSS, for supplementary analyses) which are concerned also with mobilizing procedures. Clicking the Macro icon presents a list of prepared and stored sets of commands presented in Pajek. Users can define and save their own macros for combinations of commands they use often enough to merit the construction of macros. Clicking on each of these opens a dialogue box for working with, and using, objects, pairs of objects, or triples of objects. When we present commands for using Pajek, we use primarily the items across the top of this dialogue box followed by the relevant options.

In Section 2.5, we distinguish statistical summaries of network features and summaries formed through network analytic methods. While they differ in the analyses performed, these methods are most effective when coupled. In the main, for the former, we used R.[11] Where necessary, we provide the R code used for some of our analyses, as Pajek permits easy transitions to analyses using network outputs within R.

[11] See http://www.r-project.org/.

Given our focus on large citation networks, Chapter 3 builds on Chapter 2 to lay out formal tools designed for examining such temporal networks. In general, methods used in multiple chapters are described in these two chapters. As noted above, methods specific to single substantive chapters are presented in those chapters, especially for the patent citation network (Chapter 5), the EPL football player movement network (Chapter 8), and the US spatial network (Chapter 10).

We are fully aware of a wider literature discussing the topology of large temporal networks, the importance of which we do not deny. See, for example, all of the contributions brought together in Newman et al. (2006) and the many scientific productions built on these foundations. Our goal here is *not* to provide a broad comparison with all of the methods used in other literatures.[12] Instead, in the spirit of 'letting many flowers bloom' we lay out another complete framework for studying large temporal networks. If the results of using this approach do have value, then comparisons between different approaches will have considerable merit. It is simply too early to impose a single approach for studying large networks on the study of all such networks. Of course, we are not proposing the methods introduced here as the only appropriate ones for studying large networks.

1.5 Data for large temporal networks

The datasets we use fall into two categories. The first contains data defined by the substantive interests outlined in Section 1.3. These data are used for the analysis and results presented in Chapters 4–9. The other (secondary) category[13] has data used for *illustrating* concepts and methods introduced in Chapters 2 and 3. We know that the term 'interesting' (when it is not used as cover for not expressing an opinion one way or another) is in the eye of the beholder. The distinction between main (primary) and secondary datasets is not intended as an evaluative statement about their relative merits even though we do insist that the data considered here need to be relevant for specific substantive concerns. The secondary data sets have different substantive interests and technical issues in mind.

1.5.1 The main datasets

We describe briefly these main datasets, each driven by substantive interests, and present their dimensions here. Appendix A contains detailed descriptions of them, including how they were obtained and the data processing for getting them into the form we use. Their initial[14] dimensions are provided in Table 1.1. Some[15] of these data are freely available at Pajek datasets (see http://vlado.fmf.uni-lj.si/pub/networks/data/sport/football.htm).

[12] We have noticed in submitted manuscripts involving blockmodeling (see Doreian et al. (2005)), reviewers often *demand* a *full coverage* of the community detection literature (created mainly by physicists) ideas even when community detection ideas are tangential. While there is, at face value, some commonality between these approaches, the differences are quite marked and rather subtle. Such broad summaries often are distractions – and, when complied with, can affect citation networks.

[13] With a few exceptions, we maintain this distinction to have our substantively relevant results remain within single chapters.

[14] For some analyses, not all of these data were used. For other analyses various subsets were used and the results combined. (See Section 2.4 for a description of the 'divide and conquer' strategy that we employ for simplifying large networks.)

[15] The exception is the football data because we intend to explore them further before making them available publicly.

Table 1.1 Dimensions of the datasets used in Chapters 4–10.

Substantive network	Number of units	Number of lines	Pajek data-set name
Patent	> 3.2 million	> 32 million	`patent.net`
Supreme Court	30,288	216,758	`allcitesV4.net`
Centrality	995,783	1,856,102	`Cite.net`
SNA	193,376	324,616	`SN5cite.net`
Football	2355	40,246	not available
Spatial USA	3111	7101	`UScounties.zip`

The patent citation network (for patents citing earlier granted patents) features patents issued in the USA. The time period is relatively short, covering 1976–2006, a mere 30 years. However, this network is the largest dataset we consider for the substantive chapters, having more than 3.2 million patents linked by over 32 million citation links. The US Supreme Court citation network, in contrast, is much smaller with more than 30,000 units and over 216,000 citation links. However, it covers more than 200 years; this is by far the longest time span of all of the networks we study.

There are intrinsic differences beyond their sizes of these two networks. We noted in Section 1.3.1 the strict constraints on patent citations, in contrast to the freedom that SC Justices have in citing prior decisions. There are many SC decisions that neither made nor received citations. One practical consequence is that the relevant citation network has fewer units than the number listed in Table 1.1. However, the long time span and the depth (defined in Chapter 3) of the SC citation network created technical problems requiring attention before the general methods for acyclic networks presented in Chapter 3 could be used. The patent citation network was acyclic as received. This was not the case for the SC data: some decisions handed down by the same Court in a short period of time do cite each other, a phenomenon present also in the scientific citation data for publications appearing in the same year. Solutions for handling this problem are described in Chapter 3 and mobilized in the analyses of both the centrality and the broader SNA literature, in addition to the Supreme Court network. In analysis of centrality and SNA literature we used also some other bibliometric networks.

The football data that we constructed have a far more complex structure, featuring football players, football clubs, and countries. It was defined by the 3749 football players playing in any of the first 15 seasons of the EPL. These players had 148 nationalities (dual citizenship is precluded for defining the nationality of players). Even though the player network is defined by these players, our primary interest centered on the clubs for which they played. More specifically, the network ties for this network are the links between these clubs as created by players moving between them. The number of clubs involved in their migrations to and from the EPL was 2355. These clubs are located in 152 countries. The total number of player moves between clubs was 40,246. We also used ancillary data (described in Appendix A.4) on clubs and player presence by nationality in other top European leagues for additional analyses.

Our example of a large spatial network features all of the counties of the contiguous USA and was motivated by trying to reconcile two very different approaches to mapping social diversity in geographic space. The substantive problem has intrinsic interest, and the network we study is one of the larger substantively interesting networks we have located.

Table 1.2 Dimensions of the illustrative datasets.

Substantive network	Number of units	Number of lines	Pajek data set name
EAT	23,219	325,589	`eatSR.net`
NBER	174	11755	`NBERwt.zip`
KEDS	325	78,667	`BalkanDays.net`
e-companies	219	631	`krebs.paj`

1.5.2 Secondary datasets

We report our results for the networks described in Section 1.5.1 extensively in the relevant chapters. To avoid repetition of results, we used data from other sources to illustrate our methods in Chapters 2 and 3. These data, the dimensions of which are listed in Table 1.2, came from the following sources.

1.5.2.1 The Edinburgh Associative Thesaurus (EAT)

The primary goal of the EAT project[16] was to understand how words in the English language are coupled. This was done by examining empirical 'associations' between words. The approach taken to obtain these word associations was straightforward. Subjects were shown a word and then asked to provide the first word coming to their minds. The procedure presented batches of words to each subject. The presented words were regarded as stimuli and the words offered by subjects as responses. The established links between stimuli and responses were provided by subjects. There were no imposed rules dictating the nature (appropriateness) of the responses. The pairings of stimuli and responses were simply empirical associations. For each pair of words, they were aggregated across subjects as a way of quantifying these associations. For example, some frequent couplings included 'husband' in response to 'wife' and 'cheddar' in response to 'cheese'.

The resulting Edinburgh Thesaurus association norms were started from a nucleus set of words. Further associations were collected by expanding from the nucleus: initial words were used to obtain further responses, together with additional words. The EAT website reports this cycle was repeated about three times. By then, the number of different responses became so large they could not be reused as stimuli in a systematic fashion. The EAT data collection stopped after 8400 stimulus words were used.[17] The result was a total of 23,219 words in the Thesaurus network linked by 325,589 associations. The database has two files: one is a SR (stimulus–response) file, with the other being a RS (response–stimulus) file. These data are used in Section 2.5.

1.5.2.2 The NBER-United Nations Trade Data, 1962–2000

This network was used for illustrative purposes in Section 2.6. The network ties are trade exchanges (exports and imports) between nations. The data we used came from 1999: there are

[16] See http://www.eat.rl.ac.uk/ for a description of this project.

[17] Each stimulus word was presented to 100 different subjects. Their website reports that the subjects were mostly undergraduates from many British universities whose ages ranged from 17 to 22 with a modal age of 19. The sex distribution was about 64 per cent male and 36 per cent female. The data were collected between June 1968 and May 1971. Any bias in the distribution of associations due to using university students as subjects has no relevance for our *illustrative* purposes regarding methods.

174 vertices and 11755 trade flows linking nations. The weight of the arcs are trade values in $US1000. The source for these data is http://cid.econ.ucdavis.edu/data/undata/undata.html. The complete dataset is available as the zipped Pajek project file listed in Table 1.2.

1.5.2.3 The Kansas Event Data (KEDS)

The data in this resource are the results of a 20-year project, originally based in the Department of Political Science at the University of Kansas. This project and its data were known as the Kansas Event Data System (KEDS), a label we use here. It was moved to the Department of Political Science at Pennsylvania State University in January 2010 (http://eventdata.psu.edu/). The project uses automated coding of English-language news reports from a variety of news resources to generate political event data focusing on the Middle East, the Balkans, and West Africa. These data were designed primarily for use in statistical early warning models to predict political change in these regions with attention given to suggestions and policies for mediating conflicts. The units for this network are nations and organizations. The relations include ties between nations in the form of actions by one nation directed towards another nation, as described by verbs. These actions include visits, seeking information, issuing warnings, and expelling persons. Data for the Balkans for KEDS are used in Section 2.2. The full dataset is available also at the KEDS website.

1.5.2.4 Krebs Internet industry partnerships

Valdis Krebs collected in 2002 (http://www.orgnet.com/netindustry.html) a network of Internet industry partnerships. Two companies are linked with a line if they have announced a joint venture, strategic alliance, or other partnership during the period 1998–2001. The companies are classified into three classes: 1 – content, 2 – infrastructure, 3 – commerce.

1.5.2.5 Data archives

There are variety of sources containing many datasets, both large and small, but with a primary focus on large datasets. One is SNAP, the Stanford Large Network Dataset Collection maintained by Jure Leskovec. It is documented at http://snap.stanford.edu/data/. The topics covered include on line social networks, communication networks, citation networks, and collaboration networks. There are also graphs of the internet and physical road systems. Signed networks are included in this archive. KONECT, the Koblenz Network Collection, contains large network datasets assembled at the Institute of Web Science and Technologies at the University of Koblenz-Landau. As stated on its website (http://konect.uni-koblenz.de/): ‘KONECT contains over a hundred network datasets of various types, including directed, undirected, bipartite, weighted, unweighted, signed and rating networks. The networks of KONECT are collected from many diverse areas such as social networks, hyperlink networks, authorship networks, physical networks, interaction networks and communication networks.’

These archives of datasets are used in Section 2.3 when describing the distribution of network sizes in terms of the number of units and relational ties. Networks are sparse when they have roughly the same number of units and relations ties. More specifically, the numbers of these ties are not orders of magnitude larger than the number of units. Networks being sparse is crucial for developing efficient methods for analyzing large networks.

1.6 Induction and deduction

Throughout this book we have been both inductive and deductive in our approach, with a strong bias in favor of being inductive when examining large networks. However, our uses of both induction and deduction were based on substance and driven by curiosity. For the three citation networks, induction reigned. When examining the social network citation networks, we wanted to learn the citation structure following two salient events. One was the formalization of centrality (and network centralization) in the Freeman (1977, 1979) papers. As we have noted, this triggered an explosion of work both extending and using these ideas. The other event was the recent interest of physicists in studying social networks. We stated rival expectations regarding the possible convergence or divergence of the traditional SNA and network science literatures. But, lacking anything beyond broad statements about this question in the literature, we had no foundation for a specific hypothesis.

This was just as well. One of our expectations was that the physicist conception regarding network science supplanted the SNA conception, especially regarding centrality. This was born out but, as noted in Chapter 4, this was not the end of the story. The full disciplinary sequence, for the centrality citation data ending in early 2013, was SNA → network science → neuroscience. We learned that the general concept of 'centrality' has multiple sources. There are parts of the broader centrality literature having *nothing* to do with traditional SNA concerns. Worries about 'the invasion of the physicists' may be a somewhat parochial conception within the older SNA community. This issue is explored more extensively in Chapter 4.

Our exploration of the patent citation network was inductive also. Given the four broad technological domains for 'utility patents' defined by USPTO and described in Section 5.1, we were curious about the flow of ideas between these broad technological areas as reflected by citations between patent applications and how they changed over time. Further, as technologies change over time, specific inventions are likely to have a limited shelf life. One crucial feature related to this is the lag between patents being granted and their ideas being picked up and used fruitfully for later inventions. Our interest centered on the distribution of these lags and their temporal dynamics.

Again, we were inductive in our approach to the Supreme Court citation network. However, there was an implicit hypothesis – about the line islands we identified having coherence – which underlay our analysis. Alternatively put, we gambled on this hypothesis being correct. If the gamble was lost, then this approach would be seen as severely flawed. Fortunately, thus far, *every* line island we have examined has a singular coherence even though the specific nature of their coherence differs by island. Establishing the presence of coherence of decisions being co-cited frequently was a purely inductive, but not surprising, outcome. This coherence among a set of frequently co-cited decisions comes either from the constitutional principles underlying these sets of decisions, the substantive domains of the decisions, or both. Induction of a different sort followed the identification of coherent patches in this citation network. Having identified them, we sought to understand both the decisions and the citations between them in their historical, social, and political contexts. Beyond the line islands considered in Chapter 6, two we considered were technologically driven. One concerned railraids when rail was an emerging technology with great commercial and social implications. Another featured maritime law, first defined over centuries for travel on seas and oceans, and then adapted as internal waterways – rivers and lakes – were used in the USA, especially for commerce.

For the football network, our approach was completely deductive. Based on our reading of the literature regarding football player moves, we formulated an explicit set of hypotheses. We knew that players move in the hope of advancing their careers, while clubs recruited players with the intent of achieving greater success (or avoiding failure) on the field. Coupling the decisions of players and the decisions of clubs is a highly uncertain processes for reasons outlined in Chapter 7. We state 21 hypotheses in Chapter 7 and test them. The results of these tests are reported in Chapters 7 and 8. While some hypotheses were obvious, others were counter to conventional thinking about player movements. Many hypotheses passed muster, some failed miserably, and others, while sounding plausible, turned out to be untestable in the sense of there being both supporting and refuting evidence about them. Not surprisingly, regarding the failed hypotheses and the untestable hypotheses, conventional wisdom about football in England does tend to be supported by selective attention to the evidence.

In essence, we returned to induction for our analysis of the large US spatial network. Indeed, we state no hypotheses. Our intent was to combine two broad– seemingly incompatible– approaches to mapping spatial diversity. The resulting compromise led to results sitting between these two broad approaches. Of course, this does not have surprise value because we were more attentive to *both* network geography (adjacency in space) and also appropriate statistical data. As a result, our results provide the foundations for a deeper characterization of the spatial distribution of diversity in the USA.

The final chapter provides a partial summary of the results provided in Chapters 4–9, together with commentary on the utility of the methods used throughout this book. Also proposed in the final chapter are some suggestions for further work. Despite all that is accomplished here, one salutary implication is that much more needs to be done. Pursuing these issues has immense appeal.