1

Introduction

Patrick Doreian[3,4], Vladimir Batagelj[1,2,5], and Anuška Ferligoj[3,5]
[1]IMFM Ljubljana
[2]IAM, University of Primorska, Koper
[3]FDV, University of Ljubljana
[4]University of Pittsburgh
[5]NRU HSE Moscow

This book focuses on network clustering regardless of the disciplines within which a network was established. In the initial conception for the book, our attention was driven primarily by concerns regarding blockmodeling and community detection as they applied to social networks. But as we looked further into this general topic to invite potential contributors, we realized that the domain was much broader. The wide variety of approaches contained in this volume exemplifies this diversity. For us, as we assembled this volume, this was an exciting learning experience, one we hope will be experienced by readers of this book.

There is no single best approach to network clustering. Put differently, there is no cookie-cutter approach fitting all such data sets. Yes, there are adherents of one (their) approach who think this is the case. As shown in the chapters that follow, none of the authors of the contributed chapters share this very narrow view. This is a wide-open realm with multiple exciting approaches. We reflect further on this in the concluding chapter. Here, we describe briefly the contents of the following chapters merely as an introduction to them. In our view, each chapter merits close attention.

1.1 On the Chapters

As the book is concerned with network clustering, Chapter 2 offers an extended examination of the network clustering literature. Identifying the citation network for this literature turned out to be a complex task. Methods for doing this are described in detail. The initial search used the Web

Advances in Network Clustering and Blockmodeling, First Edition.
Edited by Patrick Doreian, Vladimir Batagelj, and Anuška Ferligoj.

of Science (WoS) to identify documents using search terms. There were multiple searches, with the first being for 2015 and the final network was obtained for February 2017. This literature expanded dramatically and in a complex fashion.

While a citation network is composed of links between works treated as nodes, there is more to consider when other types of units are included. These include authors, journals, and keywords. As part of a more general strategy, one starting from the citation network but using additional information, multiple two-mode networks were constructed. This included an authorship network with works × authors, a journal network featuring works × journals, and a keyword network with works × keywords. They help give more insight into the one-mode citation network having only scientific productions.

Chapter 2 lists the most cited works, the most citing works, the most used keywords, and a discussion of the "boundary problem" as it relates to citation networks. One message from this chapter is that the way the boundary of a citation network is established affects greatly the data analytic results that follow. This implies using *great care* in constructing citation networks – not all citation networks will suffice for meaningful analyses.

Chapter 2 presents a wide range of analyses of the citation network for the network clustering literature. Components of the network were established. Both critical project main paths and key-route paths were identified. In the analysis of this network, a clear transition between the blockmodeling and the community detection literatures was revealed. Another technique used in Chapter 2 is the identification of link islands which have higher levels of internal cohesion as a way identifying some important subnetworks of this literature. The largest of these link islands featured works from the blockmodeling and community detection literatures. Since the transition from the former to the latter, it seems the two have developed independently. This seems problematic given our belief in the utility of useful ideas flowing between fields and sub-fields.

The other islands discussed in Chapter 2 come from the fields of engineering geology, geophysics, as well as electromagnetic fields and their impacts on humans. One of the searches used in the searches of the WoS database included the terms "block model" and "block". The latter crops up in other literatures, a surprise for us. In considering these other link islands, another surprise awaited the authors of this chapter. We are used to debates in the social network literature regarding the difference between static and temporal approaches to studying networks. This divide is present also in the natural sciences.

Chapter 2 also contains an examination of authors and measures of productivity within research groups, collaboration, and an examination of citations among authors contributing to the network clustering literature. Again, the stark division between the community detection and blockmodeling literatures was clear. Examined also are citations between journals publishing articles in the broad area of the network clustering literature. The methodological details of doing this, as outlined in this chapter, merit further attention.

Bibliographic coupling, which occurs when two works both cite a third work in their bibliographies, was also examined. This included a sustained assessment as to how this coupling is measured. These tools were applied to the network clustering literature, especially for the largest identified island. This included an examination of the most frequently used keywords in the social networks literature and the physicist-driven approach to studying social networks. While some keywords were the same, there were considerable differences, again illustrating the different concerns in these two literatures.

Chapter 3 provides an overview of "classical" clustering including both the clustering of networks and clustering in networks. The clustering problem is considered as a discrete optimization problem which turns to be, in most cases, NP hard. Therefore, local optimization or greedy methods are usually used for solving it. These methods can be adapted also for clustering in networks (or clustering with relational constraints). The hierarchical agglomerative clustering method can be extended for efficiently clustering large sparse networks. This is illustrated with an analysis of normalized author citation networks from the network clustering literature from Chapter 2.

Chapter 4 describes different approaches to community detection. The authors ask very useful questions which led them to provide helpful guidelines for researchers contemplating network clustering within a community detection framework. It starts with a bold claim that there is no precise definition of a community. We concur. Their focused review of community detection methods makes it clear that researchers need to have a clear idea as to why any method is selected prior to its use. The authors point to the problem that the same term can have different meanings in different subfields, reflecting what was found in Chapter 2. They argue "community detection should not be viewed as a well-defined problem but rather an umbrella term with many facets." This delightful image is equally applicable to blockmodeling within social network analysis!

Four different approaches to community detection are outlined. The first uses the cut-based perspective. The second is a clustering approach maximizing the internal density of clusters and the third is the stochastic equivalence perspective. Finally, there is a dynamical perspective focusing on the impact of communities and dynamic processes to establish a dynamically relevant coarse-grained partitions of network structure. Four sections follow which provide precise descriptions of the fundamental properties of these approaches and the results stemming from adopting them.

Their discussion makes it clear that there is no single "best" community detection algorithm and that there can be multiple equally valid partitions of a network depending on which of the four considered approaches is used. Again, this sentence holds fully when "blockmodeling" is used instead of "community detection". Of course, this applies to all the network clustering approaches presented in this volume.

Chapter 5 provides an extensive discussion of label propagation as a heuristic method initially proposed for community detection. There is natural segue between Chapters 4 and 5. Label propagation is a partially supervised machine learning algorithm assigning labels to previously unlabeled data points. At the start of the algorithm, subsets of nodes have labels, which amounts to a clustering of them. These labels are propagated to the unlabeled points throughout the course of the algorithm. Nodes carry a label denoting the community to which they belong. Membership in a community changes based on the labels that the neighboring nodes possess as the labels diffuse through the network.

The author is clear that while it is not the most accurate or the most robust clustering method, a label propagation algorithm is simple to implement and is exceptionally fast. Networks with hundreds of millions of nodes can be analyzed readily. The early work on this approach is described with the basic ideas presented for simple undirected networks. However, the author points out that it can be used for many more types of networks, including those with multiple edges between nodes, two-mode networks, and signed networks. It can be used also to identify and delineate overlapping clusters: it is not restricted to establishing only partitions of nodes, a useful property. Nested hierarchies of groups of nodes can be identified also.

Issues regarding the number of clusters are discussed along with updating labels, which can be done with either synchronous propagation or asynchronous propagation. Depending on the structure of the network, each can produce undesirable outcomes, which are discussed in detail. Reaching an equilibrium with nodes having stable labels is critical and ways for achieving this are discussed.

Advances in label propagation methods are described. They include adding constraints to the objective function to prevent trivial solutions, using preferences to adjust the propagation strength of nodes, and improving algorithmic performance by promoting its stability and reducing its complexity. Simple examples with planted communities are provided and used to show the subtlety of the method and the choices made when using it. A connection is made with the blockmodeling literature when using structural equivalence, consistent with the general conception motivating the book to apply and connect different methods for establishing clusters of network nodes.

Chapter 6 marks a transition from community detection issues to blockmodeling concerns. There is now an abundant amount of valued network data being collected and made available. As the initial work on blockmodeling dealt with binary networks, extending and adapting this approach to handle valued networks is important. The authors show that creating this extension is far from straightforward because subtle issues arise as to how valued network data can be treated prior to blockmodeling them. The authors discuss the difference between the traditional indirect approach and the direct approach to blockmodeling. In the former, networks are transformed into arrays expressing the similarity, or dissimilarity, of the nodes. These measures are then used to cluster the nodes. Their Figure 6.1 lays out the relevant decision points. In contrast, the direct approach eschews such transformations. Within the rubric of generalized blockmodeling, network data are analyzed directly. For valued networks, the authors cleave to the latter approach by making strategic adaptations to handle valued data in useful ways. This includes homogeneity blockmodeling championed by one of the authors and deviation generalized blockmodeling promoted by the other contributing author.

Two well-known empirical data sets were selected for a detailed examination of the issues involved in blockmodeling valued data. The simplest of the two is a friendship network. The second involves trade flows between nations, one raising the issue of relational capacity. This has major implications for the nature of discerning the relevant useful transformations. As with the previous chapters in this volume, close attention is paid to the choices that must be made regarding the appropriateness of methods given the data being analyzed and the criteria for making these choices. Their Figure 6.2 is particularly important as a way of guiding researchers to make appropriate decisions. It could be generalized more broadly and adapted for the other network clustering approaches presented in this volume.

By presenting detailed analyses of the selected empirical networks using different approaches and a variety of transformations, the authors show, and examine, the different outcomes resulting from making different strategic choices. The results have interest value in their own right, and lead to a set of useful recommendations about these choices. Some open problems are discussed briefly.

Chapter 7 continues the consideration of blockmodeling but tackles a very different issue, namely, measurement error. The premise for blockmodeling is that using these methods reveals the structural features of networks at both the macro and micro levels. The presence of measurement error complicates these analyses. In the worst case scenario it can render blockmodeling results useless. Three types of measurement errors are discussed. One takes the form of having errors in the recorded ties. The second is item non-response and the third is actor non-response.

Actor non-response is the primary focus of this chapter. Typically, and regrettably, within social network analysis, the standard response to this problem is to discard all information about the non-respondents, including information about the ties directed to them by respondents. This discarded information can be used to recover (most of) the network. The authors contend that this must be done and provide strong evidence supporting this claim. The question that follows is simple to state: how is this done?

The authors present seven ways for using such data as imputation methods for recovering the network from the ravages of actor non-response: reconstruction, imputation using the mean of the incoming ties, imputation with modal values of the incoming ties, reconstruction combined with using the incoming modal values, imputation of the total mean, imputations using the median of the three nearest neighbors based on incoming ties, and null tie imputation.

The authors examine the relative merits of these ways of recovering network data using four known empirical networks. Five steps are involved. The first establishes a partition of the known network within the indirect blockmodeling approach. The second step creates "observed" networks by randomly removing some actors (at various levels ranging from 1% to over 45%). Step three involves using each of the imputation methods to generate recovered networks. The fourth step is the clustering of each recovered network using the exact same clustering method as in the first step. The final step compares the partitions for all pairs of known and recovered networks. Two criteria were used in the comparisons. One is the Adjusted Rand Index for comparing two partitions. The other is the proportion of correctly identified blocks by position in the blockmodel. These criteria are stringent and there are clear differences regarding the adequacy of imputation methods. As with the foregoing chapters, recommendations are made regarding the best ways for recovering network data given the presence of actor non-response.

Chapter 8 addresses the clustering of signed networks, a topic that has garnered considerable attention within both the blockmodeling and community detection literatures. The authors adopt a formal approach and start within the structural balance perspective, which is a substantively driven approach to studying signed networks. The basics of this approach are reviewed. Some of the early theorems are restated with some new proofs provided. One critical feature of structural balance theory states that a signed network is balanced if all its cycles are balanced. But cycles can vary in length, a feature that complicates algorithms used for determining the extent to which graphs are imbalanced. The authors use the concept of chords, which allows cycles to have two subcycles which simplifies computing the sign of a cycle.

One prominent feature of the balance theoretic approach centers on what has come to be called the "structure theorems". Initially, if a signed graph was balanced, its nodes could be partitioned into two clusters such that all the positive ties were in one cluster and all the negative ties went between the two clusters. Later, this was extended to any number of clusters having this property. Here, the authors call the former strong structural balance and the latter weak structural balance. One interesting theorem in this chapter states that signed graphs are weakly structurally balanced if and only if all chordless cycles are weakly structurally balanced.

The authors then turn to consider the clustering of signed networks in a more general fashion. For strong structural balance, they couple this approach to spectral theory. They rework an early concept of switching (from a 1958 paper) to prove theorems relating to balance in signed networks using this concept. For weak structural balance, one allowing for more than two clusters, additional ideas have emerged. The structure theorems point to a blockmodel where diagonal blocks are positive (with primarily positive ties) and non-diagonal blocks are negative (with primarily negative ties). Yet empirical networks can come in the form of having off-diagonal positive blocks and, far more rarely, on-diagonal negative blocks. As the authors note, more

research is required regarding the distribution of signed blocks in a blockmodel. They consider also community detection issues for signed networks and note that one of the main concepts of this approach, modularity, has problems when there are negative links in the network. They provide ways of addressing this problem.

The authors address the critical problem of studying temporal networks in a dynamic context and, as an empirical example, study the international system with signed relation between nations. They display timeline graphs showing variations in the levels of imbalance in the international system using different methods. They confirm that signed networks move both towards balance and away from balance depending on contexts. Their results add to the solid argument *against* the early presumption of structural balance theorists that signed networks always move towards balance. Also displayed are partitions of nations for various time points, which are interpreted in interesting ways.

Chapter 9 presents a summary of work on multimode network clustering and illustrates the results of using different methods on a single well-known early two-mode network. One of the conceptions behind this volume is bringing together ideas from multiple disciplines. Actually, the authors of this chapter engage in this process explicitly. While they focus on two-mode networks in the form of actors × events as a bipartite network, the implications of the materials in this chapter extend much further. Although such two-mode networks have been considered in earlier chapters, especially Chapter 2, having an integrated discussion of the ways in which they can be analyzed is particularly useful. They note that both binary and valued two-mode networks can be analyzed within a common rubric. Multiple such methods are discussed in the chapter.

The authors establish a conceptual link to community detection, make some definitions to help link the two literatures, and note that community detection is a special case of blockmodeling. The same point was made in Chapter 4. Some authors focused on community detection methods might disagree! Here, the core community detection notion of modularity is provided and, more importantly, the authors extend this to two-mode networks. Their first presented partition (of actors) concerns group assignment maximizing modularity.

Given a two-mode network, denoted A, it is straightforward to create two projections for actors, using AA', and events, using $A'A$. They challenge the presumption that evidence is lost in this dual projection. Without doubt, as the authors note, this holds when projections are dichotomized or if only one projection is used. However, they challenge the claim that dual projection loses information even when both projections are used in their undichotomized forms. Elsewhere, they have provided strong evidence that this is not the case and have promoted what they call the dual-projection approach. They present further partitions of the actors of the considered empirical network using dual-projection, using core-periphery notions and for dual-projection community detection. The latter led to two more partitions, one with two clusters and one with four clusters of actors.

In their spirit of integration, the authors include a consideration of signed two-mode networks and a consideration of spectral methods. Two more partitions of the actors are presented. As with previous chapters, it seems reasonable to have multiple valid partitions of a network. The authors finish with suggestions regarding the analyses of more complex data structure involving more modes and extend this to temporal evolution of two-mode networks.

Chapter 10 is devoted to blockmodeling linked networks and provides another segue in this volume. This time it is from Chapter 9. The term "linked networks" features a set of one-mode networks where the nodes from the one-mode networks are linked through two-mode networks.

This can be done in a variety of ways, including the coupling of networks linked over multiple time points. In dealing with these configurations, the author distinguishes analyses of separate networks, using a conversion approach under which all the one-mode networks are converted to a single level by joining them through the two-mode networks, and using a genuine multilevel approach. The results of examining both the conversion to a single level and using the multilevel approach are presented. Comparisons of them demonstrate clearly how the linked blockmodeling approach has greater potential value.

Two empirical examples are used. One concerns a coauthorship network at two time points while the other features participants in a fair-trade exchange for TV programs. In both examples, results of different partitions are presented with insightful comparisons of the results. As with all methods, parameter settings require consideration. For these analyses, this concerns the weighting of null and complete blocks for the scientific citation network. The final reported results provide a very coherent result with strikingly clear differences in the partitions for two distinct time points, which provide useful interpretations of the dynamics of scientific collaboration. The empirical results resulting from using the genuine multilevel method for the trade fair are equally compelling. As with other chapters, the author provides a provisional agenda for future work.

Chapter 11 provides a self-contained introduction to using Bayesian inference to extract the large-scale modular structure from network data. In terms of the foregoing content of this chapter, the modules are clusters (or groups) identified in the network. Rather than focus on deterministic blockmodeling, Chapter 11 deals with Bayesian stochastic blockmodeling. A major focus is on estimating probabilistic models to shed light on the network mechanisms generating the observed network(s). An overarching feature is to distinguish genuine structure from randomness.

In this context, Figure 11.1 is especially provocative, with three displays of a randomly generated network having three separate orderings of the nodes. Two of them appear to show clear – but different – blockmodel structures, exactly the sort that those using deterministic blockmodeling would take as evidence of structure. While these blockmodels could be accepted as "real" and could be "interpreted", this would reveal nothing about the generative process creating the network. This might rattle the cages of some social network blockmodelers. The author invites readers to think probabilistically and couple two ideas. One is to think about mechanisms that could generate networks. The other is to use the network data to discern which mechanism was the most likely to have generated the network. This leads directly to notion of stochastic blockmodels within which known (prescribed) modular structures are generated according to probabilistic rules. Then, given network data, Bayesian inference is used to infer the modular structure of observed networks.

The author provides formal discussions of a wide variety of prior distributions and how data affect them to create posterior distributions. Many empirical applications are used to illustrate the outcomes stemming from using the methods described in formal detail. This includes the subtleties of model selection and the establishment of efficient estimation procedures. The ultimate outcome is the establishment of modular structures that are supported by statistical evidence.

Chapter 12 also focuses on a dynamical perspective. Both this chapter and Chapter 11 are concerned with modular structures having a coarse grain, along with the rich interplay between network structures and network dynamics. However, the authors of Chapter 12 take a very different approach compared with the one contained in Chapter 11. Their concern is centered on the dynamical processes occurring on a fixed network structure. They do not consider, at least

initially, the question of how and why networks occur. Throughout their presentation, they focus on consensus dynamics and diffusion processes both substantively and as guiding examples. Later, they consider diffusion and consensus as dual processes, an important extension.

For modeling dynamics, they use ordinary differential equations in which the actors have attributes that can be changed by the operation of social processes operating over a fixed network. While discrete-time versions could be used, they use continuous-time models throughout their chapter. Also, of great interest, they consider processes having different network time scales under which some variables change slowly while others change more quickly. They illustrate this with a modular network having k modules with strong within coupling and weak between coupling. More complicated structures are examined also.

The authors extend this approach to consider signed networks and use the early work on structural balance theory while restricting their attention to strong balance as described in Chapter 8. The early structural balance literature was fixated on the notion that signed networks always move towards balance. This claim is repeated here. A far more important issue is the examination of how (and why) signed networks move towards balance at some points in time and move way from balance at other times. It would seem that the author's use of using differential equations, as is done in Chapter 12, for signed networks holds immense promise for examining these dynamics, especially with the inclusion of different time scales.

Later in their chapter, the authors turn to using dynamical processes to reveal network structure within the community detection framework. They employ a genetic algorithm framework to do this with a variety of extensions, all of which hold considerable promise. As with previous chapters, some open problems are stated with interesting methodological and substantive implications if they are pursued in a dynamic framework using differential equations.

Chapter 13 is the final contributed chapter, one that examines scientific coauthorship networks. A blockmodeling approach is adopted to understand the structure of these networks with a view to understanding the dynamics of scientific knowledge production. The data used feature collaboration among Slovene scientists using a rich temporal database. While blockmodeling can be used to discern the structure of these networks at multiple points in time, one particularly interesting question is whether these blockmodels, especially the composition of positions (clusters) and the relationships between positions (blocks) are stable over time. The authors of this chapter present a methodology for measuring the stability of such blockmodels over time. Of particular interest is the stability (or not) of cores. For this important task, a variety of indices are proposed for assessing this stability.

Science is dynamic in many ways. Of particular interest in Chapter 13 is the changing relationships between researchers through time. Over the course of their careers, the collaborative behavior of researchers changes as new problems engage their interests and collaborative partners change. Also, some researchers depart while new researchers enter the scientific system. The authors of this chapter consider first one discipline that has a core-periphery structure (with cores, a semi-periphery, and a periphery) at the two time periods they consider. They developed a visualization for transitions between these positions. This is then extended to consider 43 disciplines. For each identified discipline, they identify changes in the number of cores, the average size of them, and the relative sizes of the semi-periphery and the periphery.

The authors use a variety of different indices for measuring the stability of cores for all the studied scientific disciplines. They establish a partition of disciplines into three clusters. The smallest cluster has eight disciplines for which the cores are stable. The next smallest is a cluster of 13 disciplines whose cores are unstable. The largest cluster has 22 disciplines that are located

between these extremes. One implication is that science in not monolithic. While is it obvious that disciplines have different concerns in terms of content, these results reveal clearly how their collaboration structures vary greatly. The authors present results showing why this is the case.

1.2 Looking Forward

Even though the single focus of network clustering defines the impulse behind this volume, the topic has many facets within which many approaches have been adopted. In looking at the contributed chapters, there is great diversity in the topics considered and the approaches taken by the contributing authors. This was expected and was core to constructing this volume. There are many points of consistency across the chapters along with apparent disagreements. The former is great. The latter is not a problem for there will always be diverse views in this literature. We compliment all the contributing authors for their willingness to contribute in an open-minded and engaged fashion. While many academic disciplines have been riven by deep divides across which no compromise is possible, our hope is that the consideration of the network clustering literature presented in this volume will allow us to rise above such foolish divides. The contributed chapters suggest this is very possible.

So, to our readers of this volume, we hope you will enjoy the contributions in each of the contributed chapters. Each has great merit. We will return to some of the general issues raised within each of the chapters, as well joining issues from these chapters, in the concluding chapter.