27

Structural Cohesion and Cohesive Groups

James Moody and Peter J. Mucha

INTRODUCTION AND BACKGROUND

One of the most interesting features of social networks is they tend to be clumpy – with interactions happening much more often among some subsets of actors than others. Substantively this inhomogeneity reflects natural social groups: in kids' networks these might be peer groups; in organisations, business coalitions; or in national systems, alliances. These sets of actors often capture our most important social activities: these are the people who enforce norms (Axelrod, 1985), influence our behaviour (Friedkin & Cook 1990; Kreager et al., 2011), or otherwise reflect the lived social communities that networks often intend to capture (Freeman, 1992; Friedkin, 2004).

While intuitively common and theoretically salient, social groups are notoriously difficult to identify methodologically, with a multitude of papers describing new approaches for finding groups or communities (for good prior reviews, see, e.g., Porter et al., 2009; Fortunato, 2010; Fortunato & Hric, 2016; Lee & Wilkinson, 2019; Shai et al., 2020). Our intuition is that we face a mismatch between theory and reality. Theoretically, treatments of 'cohesive groups' in social networks tend to be one-dimensional and lead to unrealistic null models. Our theories anticipate sharp boundaries and clear distinctions, but reality is often more nuanced. We think this misfit results from conflating two distinct dimensions that might better be treated separately. On the one hand, groups are characterised by their internal cohesion that captures how difficult it is to separate members of the (sub)network. On the other hand, we expect groups to be socially distinct, implying an implicit boundary maintenance process that delineates 'in' from 'out'. Our theories of groups generally assume high levels of each, when these dimensions might often vary independently.

Even when the theoretical objective is clear(er), however, the methodological problem is non-trivial for at least three reasons. The first is the sheer computational complexity of the task, given the many possible ways to assign nodes to groups; we simply cannot compare every possible solution even with a well-defined metric (indeed, many methods are NP-complete, e.g., Brandes et al., 2008).

Second, for many metrics, there are equivalent solutions that are substantively different, implying we cannot identify a uniquely optimal solution. Consider a bridging node with ties that span two otherwise disconnected cliques: most off-the-shelf methods would require the node to be assigned to

one group or the other, or bring both into a single supercluster. Sometimes this is a signal that the method is inappropriate (perhaps we should not be seeking mutually exclusive solutions, even if it's computationally simpler) but often simply reflects the messiness of the world itself: equivalences are features of the world, not problems of the method.

Finally, notions of cohesion are often scale-dependent, with the appropriateness of a given solution depending on the comparison level for a given analysis. We might find a very clear clustering of a large network into a small number of distinct groups, but on examination discover that each group contains its own fractures. Such hierarchical ordering of groups is a puzzle: sets that seem together on one level (the whole network) are fractured at another (within each initial cluster). This last problem has subtly different variants, on the one hand depending on, for example, the number of groups inferred or the nature of null models captured via a resolution parameter (see below), on the other hand reflecting a substantively different order of collective organisation (hierarchical rather than modular).

Our aim in this chapter is to help clarify these distinctions by delineating the two primary dimensions of groups common in the literature and the tools used to measure them. Figure 27.1 provides a simplified rubric that guides this work.

The first dimension is *connectivity*, which refers substantively to the network being 'well-held-together' (Markovsky & Lawler, 1994). Intuitively, we expect cohesive networks to have many relations connecting many pairs; the collective does not depend on any single node (or small subset of nodes) to control or disrupt it. Information can easily pass between all members of a cohesive network. The natural inversion of a cohesive network is anomic or disintegrated, where nodes have few ties to each other and little chance of sharing collective information or identities across the network.

The second dimension is *boundary salience*. A population with salient boundaries has clear differentiation by categories. The most well recognised are ascribed status characteristics (White, 1966), such as race, gender or caste, though social network researchers are often interested in groups without clear external status indicators, such as 'leading crowds' in schools (Coleman, 1961). When boundaries are salient, even if informal, actors likely recognise them and hold relational expectations for behaviours within and between boundaries: mean girls bully the desperate wannabes; and the wannabes on some level expect it (Waters et al., 2004). Pressures towards social balance (Cartwright & Harary, 1956) tend to create homophily within group boundaries as friends come to like the same activities, in some cases becoming equivalent to the group itself (e.g., jocks playing sports, mathletes[1]).

The intersection of these two dimensions defines a space of network archetypes that can easily be misrecognised in empirical work. Standard intuition is that these two dimensions are positively correlated: settings are either anomic – low

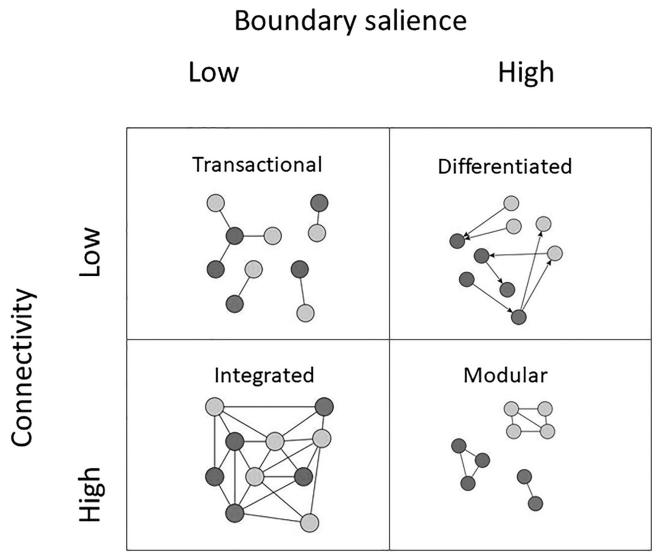

Figure 27.1 Two dimensions to cohesion and clustering in networks

connectivity and low boundary salience – or clustered – highly cohesive with strong boundaries. That is, we expect cases to fall along some continuum defined by the diagonal of Figure 27.1. Our intuitive notion of clustered networks often looks like the lower-right quadrant, with clearly defined sets of nodes and minimal contact between sets.

But social settings admit many other topologies. For example, a social system with high categorical differentiation but low (internal) group cohesion is the defining characteristic of kinship exchange systems (White, 1963; Bearman, 1997), where one clan can only marry across clan boundaries to a specific subset of other clans. This generalised exchange system archetype is referred to as 'differentiated' in Figure 27.1, illustrated by the cyclic relationship between the groups (yellow →red→green→yellow).[2] Opposite differentiated networks we find integrated ones that are highly cohesive but have no salient boundaries. The most well known are core–periphery systems composed of a diffuse core that admits a continuous gradient from central to peripheral (Mani & Moody, 2014).

STRUCTURAL COHESION

Node Connectivity

We have worked above with an intuitive notion of cohesion. Here we sharpen the definition by discussing structural cohesion. Structural cohesion was first introduced by Doug White and Frank Harary (2001) and then expanded in Moody and White (2003, p. 103, emphasis original): a group's structural cohesion is equal to *'the minimum number of actors who, if removed from the group, would disconnect the group'*. This concept is known as node connectivity in graph theory (Harary, 1969). Structural cohesion provides a nice operationalisation of Simmel's notion that the 'supra-individual' nature of a collective is the defining characteristic of social life (Simmel, [1908]1950).

A second and equivalent definition of cohesion turns on the relationships between node connectivity and the number of non-overlapping paths connecting pairs in a network. A path in a network is a sequence of adjacent nodes and edges starting with one node and ending with another that does not cross any node/edge multiple times. Two paths are node-independent if they share their starting and ending nodes but no others – they represent alternative routes between source and target. Menger's theorem equates the number of node-independent paths between a pair of nodes to the minimum number of nodes that have to be removed to disconnect the pair. This means we can define an equivalent version of structural cohesion: *a group's structural cohesion is equal to the minimum number of node-independent paths linking each pair of actors in the group.* This is sociologically interesting as it implies that groups with many unique paths connecting everyone in the group are more cohesive, matching our intuition about the ability to share information quickly and robustly among group members. This ability to communicate should lead the group to develop shared ideas and generalised understandings. Figure 27.2 illustrates these points.

Panel 1 and 2 present two networks of equivalent size – eight nodes and ten edges – in which every node is connected to every other node by at least one path. However, removing node 'e' in panel 1 would disconnect the graph, leaving two groups on the 'top' and the 'bottom' of the diagram. That is, the graph is 1-node-connected (note we could have removed 'f' as well, separating out 'd' as an isolate). In panel 2, there is no single node whose removal would disconnect the graph. Instead, we would have to strategically remove pairs ({e f}, {c b} etc.) to disconnect the graph, so the graph is 2-node-connected or a 'biconnected component'.

Note as well that while we can trace a path from node 'a' to 'h' in both cases, in panel 1 all such paths must go through node 'e' – making it a 'cutnode' in this graph – while the other network contains at least one path that avoids 'e': for example, we can highlight paths {a b d f h} and {a c e g h}. These two paths are node-independent; they overlap only on their starting and ending nodes.

In general, a network that has minimum node connectivity k is said to be k-node-connected. Node connectivity is bounded by minimum degree: one cannot be more connected than one has ties since, regardless of where the path would go, independent paths are limited by the number of first-steps on such paths, and thus this is the upper limit on node-independent paths. For sets of nodes, this means that every k-connected set must be at least a k-core (i.e., have degree ≥ k in the subgraph).

Node connectivity is a network-level property – it describes the path structure of the ties within the full graph under consideration. However, we could partition the network by some other feature and subsequently measure the node connectivity of the induced subset. This means that we can use node connectivity as a comparative characteristic for different networks (one school vs another, for example) or sub-parts of a single network.

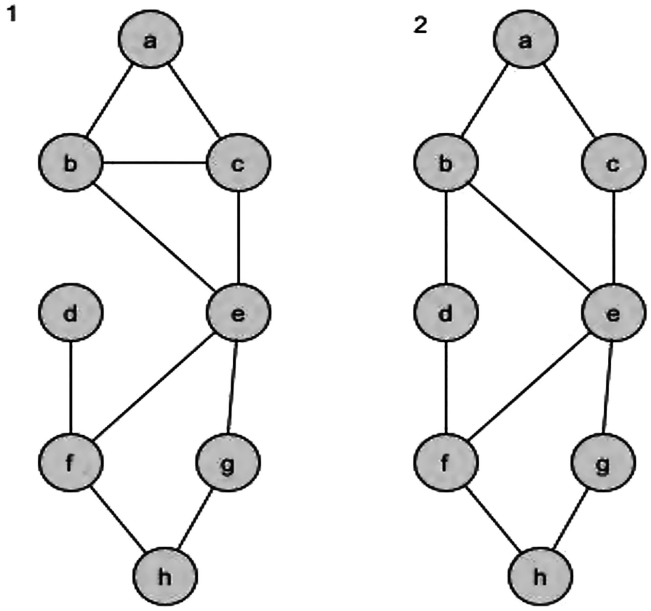

Figure 27.2 Equal volume networks with different node connectivity

Cohesive Blocking

Empirical networks usually admit to internal variation in node connectivity – some parts of a network may be more cohesive than others. We can better describe the cohesive structure of a network by recursively enumerating all k-connected sets and describing how they relate to each other. To do so, one first enumerates all minimum size cutsets and then removes all of the cutsets, assigning each to the relevant sides of their cut. If the induced subgraph is interesting[3] – that is, neither complete nor simply strings of cutnodes – then repeat the procedure on the resulting subgraphs, continuing until no further cutting can be done because you have reached a complete clique or have only isolates left. This procedure ensures that any (k + l)-connected set embedded within the network will be identified. Since each step takes us deeper into the network, removing the most weakly connected nodes and leaving stronger, more connected sets, we uncover the nested cohesion structure of the network. As an example, consider Figure 27.3.

Figure 27.3 provides a full 'cohesive blocking' of a small example network. The image at the left encircles the induced subsets of the network while the tree at the right enumerates each set and its relation to all the others. The network is 1-connected, with nodes 4 and 13 being the minimum-sized cutsets (since k = 1 these are 'cutnodes'). Removing nodes 4 and 13 results in one 'singleton' cut (node 14) and two bicomponents. We typically ignore the singleton cut and proceed on the bicomponents. The small bicomponent (right branch of the cut enumeration tree) has a size-4 three-clique embedded within it. Since this clique cannot be cut further, this branch ends here. On the other side, the size-12 bicomponent has multiple two cuts along the {9-15-16-17-2} path. Removing non-adjacent pairs of those nodes will disconnect the graph, though most of the cuts result in uninteresting singleton partitions (i.e., removing {15 17} gives two graphs, one of size 3 ({15 16 17}) and one of size 11 (all but 16). After enumerating all of these cuts and examining the resulting induced graphs, one substantive subgraph of size 9 is left. Note this subgraph is also 2-connected, so it has the same node connectivity as the parent graph it was induced from.

The cohesive blocking procedure described above can induce subgraphs that are less connected than their parent subgraphs (indicated in Figure 27.4 by dashed lines on the induced graph). The defining feature is the nested nature of the subgraph – to get to that subset, one had to go through the weakest cuts at the level above. The number of recursions needed to get to a particular subset is its *nestedness* depth, which Moody and White (2003) argue is a reasonable measure of *social*

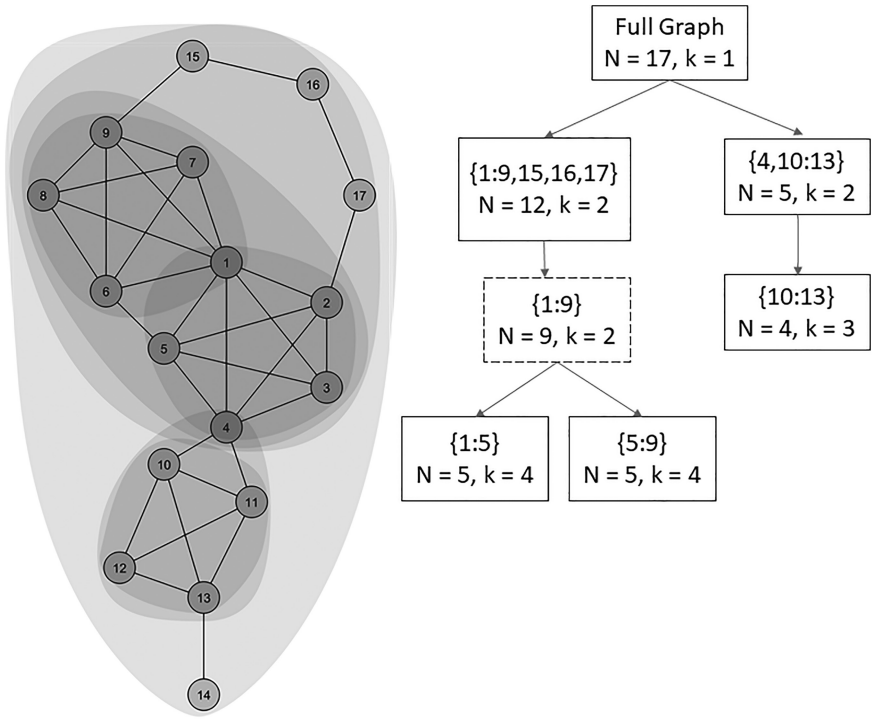

Figure 27.3 Example of cohesive blocking on a network

embeddedness (Granovetter, 1985). The choice to include weaker nested sets reflects the cutting process and thus the 'disconnect' aspect of social cohesion; but it is not unreasonable to focus only on induced graphs that increase *k-connectivity*, in which case the dashed-line box would be removed and one would go directly to the two four components embedded within the overall larger (n = 12) bicomponent, and nestedness is then equivalent to the highest node connectivity set that a node is a member of.

The cohesive blocking routine generally results in two types of induced subgraphs. The first are sets of nodes that calve away from the rest of the graph – separate distinct subsets. These are branching points in the blocking tree and generally correspond to the sorts of subgroups we traditionally think of as 'modular' in Figure 27.1. The other sort are long chains in the blocking structure, representing increasingly cohesive sets nested within each other like Russian dolls – one group is not necessarily distinct, but rather more deeply embedded than the other.

Cohesive blocking provides a full summary of the network connectivity structure, but is difficult to use in subsequent analysis. One solution is to assign nodes to a group based on shared membership in a nested *k* component. In this sense, for example, we might say that all members of either large bicomponent in Figure 27.3 are two distinct groups, with node 4 belonging to both groups. The analyst then needs to determine where in the nestedness tree to break the network, noting that such breaks can result in both overlapping groups (if at a branching point) or a non-exhaustive assignment (i.e., node 14 belongs to neither bicomponent).

Alternatively, one can characterise the network as a whole based on summaries of pairwise cohesion levels. For example, one can determine the highest *k-connected* set all pairs of nodes belong to. In Figure 27.3, for example, the overall for all pairs is 1.9. One can apply a similar summary to individuals (average, min, max) relative to all others (for individual level analysis) or to subgroups induced based on other features. Which summary statistic (mean, median, upper quartile, etc.) best captures the social process at hand is an open question. Any single summary score will almost always average over important divisions – a bow-tie network with a single node connecting two large cliques can have a high average score even

though the graph as a whole is 1-connected, so there are obvious trade-offs to this approach.

Extensions and Observations

The main advantage of the structural cohesion approach is the clear link between the exact path structure of a network and the sociological notion of cohesion: if our ties bind us together, then structurally cohesive networks are more tightly bound than non-cohesive networks. There are a number of related ideas that have somewhat different implications and are worth spelling out clearly. Most of these alternatives take a feature that is maximised in a graph-theoretic clique and use that dimension as the foundation for the cohesion metric. In most cases, however, these metrics fall victim to a base asymmetry: structurally cohesive groups have these features, but maximising these features does not ensure cohesion. The archetypical case are centralised graphs with single cutnodes (like node e in Figure 27.2) where you have low node connectivity even if other features are high.

For example, consider the difference between node and edge connectivity. Edge connectivity is the number of edges one would have to remove to disconnect a network. This fails as a general feature, when networks are highly centralised, as many ties funnel through a single cutnode. In any social process where nodes control the flow of resources through the network, highly centralised networks introduce unequal control of the flow in the network, leading to bottlenecks, fragmentation and inequality. All k-node-connected networks are at least k-edge-connected; but not all k-edge-connected networks are k-node-connected. Of course, if the social process at play is known to depend more on edges than nodes, then we may be better off focusing on edge connectivity instead. In those cases, we refer the reader to the long literature on min-cut, max-flow problems.

Other alternative metrics aimed at capturing how well the network is 'held together' by the social relations suffer much the same problem. The most common cohesion metric is probably simple density – the proportion of pairs in a network who are connected (simple) or the average strength of relations across pairs (if relation is weighted). Density focuses on volume – all else equal we would expect that having more ties among a set of nodes will result in greater connectivity and ease of resource flow. But this assumes that the network ties are somewhat evenly distributed, as otherwise highly clustered or centralised networks can suffer the sorts of problems illustrated in panel b of Figure 27.2.

A second common measure is average social distance: in a cohesive network each person is socially close to all others (at the limit, directly adjacent as in a clique). This, again, is subject to the dominance effects of high centralisation: networks with very central nodes will have short overall distances between most pairs, but that central node has a dominant role over flow in the network. In some settings, this sort of centralisation is effective – there are clear efficiencies to having centralised nodes that act as routers for distributing information or basic control purposes, for example – but we would argue these are best thought of as *efficient* structures not *cohesive* structures.

A third common measure for cohesion is the proportion of triangles in a network that are closed. Intuitively, this builds on the notion that collections of strong ties form the basis of small groups (Freeman, 1992), and we'd expect that the network as a whole is held together because each connected pair is jointly connected to common thirds. In practice, triadic closure tends to limit group size based on degree as the redundancy implied by common close ties means that relations tend to 'turn in' on others in the same group, creating a direct linkage between the 'held together' and 'distinct' dimensions discussed in Figure 27.1. Cohesion based on transitivity shares the same issues as density and k-cores discussed above: highly cohesive networks tend to be transitive, but highly transitive networks are not necessarily structurally cohesive. Moreover, because there is no distance requirement for node-connectivity, it is possible to have comparatively high k-node-cohesion (k = 3, 4) in very large groups with minimal transitivity. Whether this is a feature or a bug is context dependent. On the one hand, if the social process under investigation allows resources to flow long distances (reputation, information, viruses), then structural cohesion can provide greater efficiency re-linking wider populations with fewer ties. On the other hand, if the setting requires face-to-face reinforcement to be effective, then the distance-limiting features of transitive closure may take precedence.

The main pragmatic barrier to using structural cohesion is computational. The approach provided by Moody and White (2003) combines algorithms by Kanevsky (1990, 1993) and the key step of identifying cutsets runs in $O(2^k V^3)$ (available in SAS/IML and, with some modifications, in iGraph for R). This is often prohibitively computationally expensive for large networks. Sinkovitz et al. (2017) provide a targeted search approach that exploits k-cores (which are fast to find), but forgoes the full blocking approach implied by identifying all cutsets discussed above. Their

routine allows enumeration of all k-connected sets on networks with hundreds of thousands of nodes.

Node-connectivity can naturally extend to connectivity through nodes-of-a-class, such as would be found in bipartite networks. Cornwell and Burchard (2019) provide a detailed examination of how graphs are connected through one mode or another and define two-mode node connectivity as the number of nodes in the first mode that would have to be removed to disconnect the second mode. Extending this to the general multilevel network case could be informative, but remains unexplored to the best of our knowledge. Another way one might extend structural cohesion would be to consider paths of particular lengths or that sum to particular weights. These sorts of extensions would lead to somewhat messy and localised versions of a 'cohesive horizon' around focal nodes, but maximal sets that are node-and-distance connected might capture social processes with only localised flow.

PEER GROUPS AND COMMUNITIES

Background

Structural cohesion provides a sociologically principled approach to understanding the connectivity dimension of Figure 27.1, but does not address boundary saliency. Our traditional notions of peer groups (Freeman, 1992) are simultaneously cohesive *and distinct*. When high boundary salience and internal connectivity are combined, we get modular groups: sets of nodes with many ties to each other within groups and few(er) ties between groups. Sociological peer groups are seen as the primary site for social action. Peers share information with each other, react to status updates, enforce norms and generally provide the primary social context within which people orient their behaviour. These groups can be recognised and named within a community formally (youth gangs would be an example) or informally ('the cool kids') but often they are not named, though they might be recognised (Coleman, 1961).

The methodological conundrum of peer groups is that outsiders generally, and data collectors in particular, often lack the information necessary to know which groups are which. This has come to be known generally as the 'community detection'[4] problem, where the goal is to partition the nodes of a network into subsets in a way that optimises a target measure of joint cohesion/distinction.

The ideal-typical modular communities are similar to those in Figure 27.4, which shows four clearly distinct groups, each with very few ties to any other group and many ties within the group. In a standard sociogram, these sorts of groups jump out clearly (left panel); if we represent them as a mixing matrix (right panel) we see that the network is nearly block diagonal, meaning that the blocks on the diagonal have much higher weight than off-diagonal blocks – that is, there is more relational volume within groups than between groups.

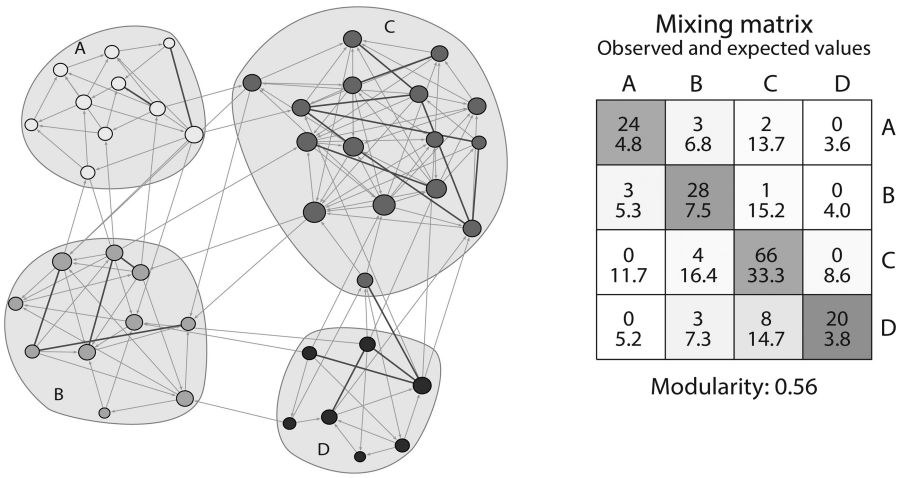

Figure 27.4 Example of a clearly modular network

In most realistic settings, however, the case is not nearly so simple. Consider Figure 27.5 below, which presents eight common clustering algorithms (discussed in detail below) applied to the Gagnon prison network (MacRae, 1960).

This multiplicity of solutions to the community detection problem is expected, analogous to the number of different available methods for unsupervised clustering of point-cloud data. In the presence of unknown mechanisms driving only weakly and potentially overlapping clusters, different algorithmic approaches will often identify different clusters. Some cluster/community labels might be more or less useful than others in a particular application: researchers should ask how particular solutions match their theoretical concerns to evaluate and select them.

Community Detection Methods

Most contemporary strategies for community detection focus on three basic strategies.[5] The first (and most common) is to use some algorithmic

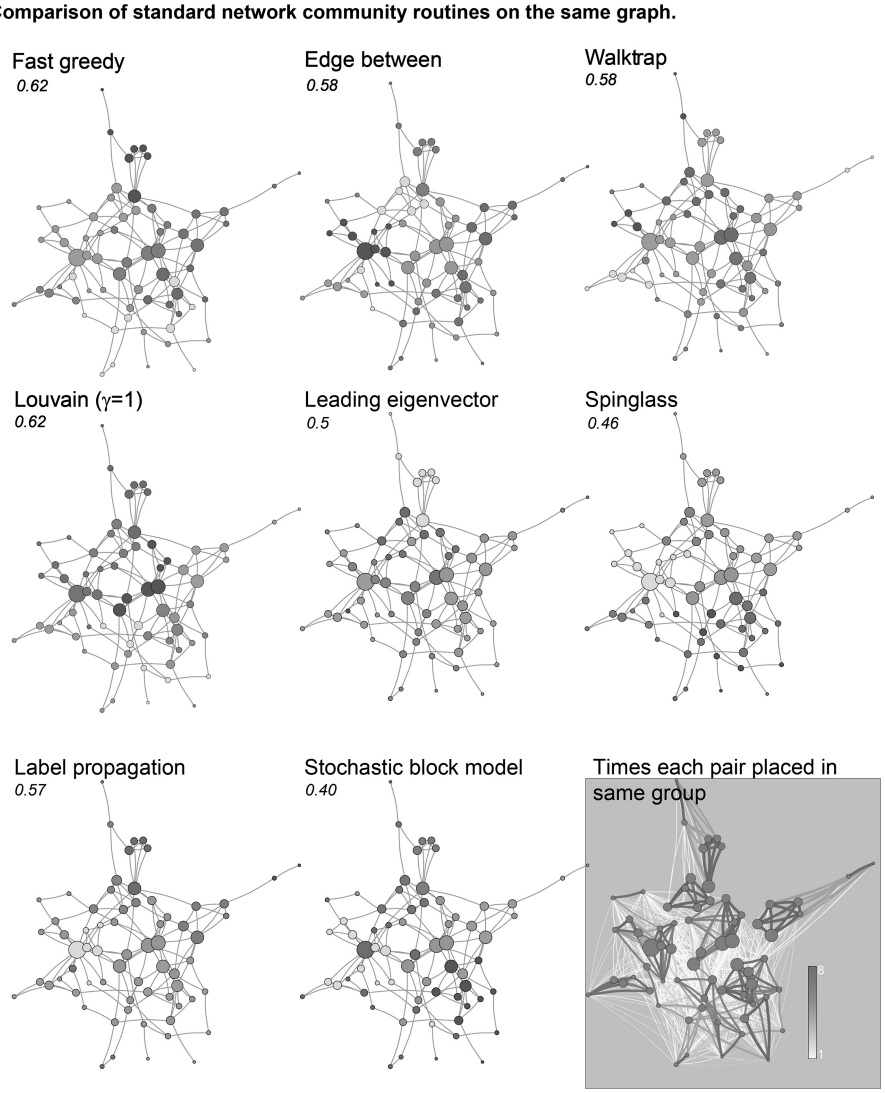

Figure 27.5 Comparison of eight clustering methods on the same real network: colours within panels denote same-community membership

process as a heuristic for optimising a community-relevant descriptive score. The second is to exploit a strong correspondence between the eigenstructure of a network and groups using variants of singular value decomposition. The third considers generative network models to recast community detection as a statistical problem where community membership influences tie probability. A full review of all such methods is beyond the scope of this chapter, but we describe exemplars and highlight some best practices.

Heuristic Sorting Approaches

Heuristic approaches to community detection sort nodes into groups based either on an index of 'groupedness' or a simulated social process that is sensitive to an underlying group structure. Given the complexity of the assignment task, such sorting approaches may not truly maximise the objective function, but instead reflect different strategies to overcome the necessary trade-offs between completeness and computational feasibility in a way that finds a reasonably good partition in a reasonable amount of time. Each of the algorithms we discuss below takes a slightly different approach to this trade-off.

The first step in direct sorting models is to define a score that reflects the group structure of the network which can then be used to judge the fit of any proposed partition. Historically a number of such metrics were developed (e.g., the Freeman segregation index (1972) or the Segregation Matrix Index (Fershtman, 1997)), which intuitively shared the notion that substantive groups are composed of people who have most of their (strong) ties within the group rather than between.

In contemporary work, the *modularity score* is the most common metric used to identify communities.[6] Modularity (Newman & Girvan, 2004) with resolution parameter γ (Reichardt & Bornholdt, 2006) counts the total edge weight within communities relative to a multiple of the expected weight in a random graph model with the same expected degree sequence:

$$\frac{1}{2m}\sum_{ij}\ A_{ij} - \gamma\frac{k_i k_j}{2m}\ \delta\left(C_i,\ C_j\right)$$

where m is the total weight of edges; k_i is the strength (weighted degree) of node i; A_{ij} is the edge weight between nodes i and j; and δ is an indicator function giving 1 if the pair is assigned to the same community ($C_i = C_j$) and 0 otherwise, thus limiting the sum to within-group node pairs.

The difference within the parentheses (A_{ij}-$\gamma(k_i k_j)/2m$) compares the weight of observed ties in the group compared to that (up to a factor of γ) expected under random mixing. In the original default γ = 1 formulation, the modularity of a network with only one community is precisely zero by construction, while the modularity of putting each node into its own community (i.e., N communities) gives negative values, and the maximum modularity will be found for some intermediate number of communities. This means that Q can serve as an objective function for partition selection, which was difficult to do with earlier measures. The extent to which one considers the total within-community edge weight relative to the expected random ties is governed by the resolution parameter γ: larger γ values increase the penalty in Q paid by putting node pairs in the same group, typically leading to larger numbers of smaller communities.

Because of the 1/2m normalisation, modularity cannot be larger than one. Then, by comparing the total within-group edge weight against that for a corresponding random graph with the given community labels, one expects Q≈0 for γ = 1 in the case where the community labels are randomly assigned. As such, it is a common misconception to assume some threshold of Q to immediately imply 'good' clustering into communities. However, it is essential to keep in mind that these special modularity values are relative to a given set of community labels. Importantly, even a random graph without any *a priori* defined community structure may have some partition of nodes into communities with a high value of modularity (Guimerà et al., 2004). Modularity only provides a measure for comparing one set of community labels to another set for the same network; any comparison of the maximum modularity obtained between different networks, or to assess that modularity is 'high' for a given network and thus claim it is strongly clustered, must be performed carefully. For example, one might compare the maximum modularity obtained for a given network against suitable permutations of the network – for example, double edge swaps to explore the space of the associated configuration model (see Fosdick et al., 2018) – and the distribution of maximum modularities computed for each network so obtained.

Early uses of modularity as a score ignored the resolution parameter (which is equivalent to setting it to 1) and much of the initial excitement over modularity was that it 'solved' the number of clusters problem. This, it turns out, is illusory: the number of clusters varies with the resolution parameter, and there is not necessarily a reason to pick γ = 1 over another value. As an empirical

feature, many social networks do admit to scale effects – some small set of resolution values for which there is a largely stable set of returned communities. For example, if one were to cluster faculty in a university based on shared graduate student committees, a coarse division between natural sciences, social sciences and humanities would probably appear for a wide range of small resolution parameter values. But, turning the resolution knob up a bit, one would probably see a stable set of solutions at a higher value reflecting university departments. Importantly, neither of these solutions is wrong – they simply capture sociality at different levels of interaction. We want to emphasise that this is a feature easily *revealed* by tuning the resolution parameter in modularity but is in no way an *artefact* of using the modularity score: many real network settings are naturally clustered at different scales.

A common approach to identifying a good resolution parameter is to calculate communities at multiple different resolution parameters and select points in the parameter space where the identified communities appear to be robust to modest change (see, e.g., Fenn et al., 2009). Figure 27.6 shows the number of communities found by the R-igraph cluster_leiden() function (Traag et al., 2019) for three different networks, demonstrating robust results (plateaus) occur at values above default $\gamma = 1$.[7] The need to undertake some systematic approach becomes even more pressing for community detection in multilayer networks (see Kivela et al., 2014) because of the introduction of at least one additional parameter (typically notated ω) to set the coupling between layers (Mucha et al., 2010). Newman (2016) addressed resolution parameter selection by identifying a fundamental equivalence between modularity maximisation and stochastic blockmodel inference in the special case of a degree-corrected planted partition model, leading to an iterative procedure for finding the γ that self-consistently maximises both modularity and the corresponding likelihood. Pamfil et al. (2019) extended Newman's iterative approach to a variety of multilayer settings. In a different approach, Weir et al. (2017) developed an efficient post-processor to find the convex hull of admissible modularity partitions (CHAMP) for an input collection of community-label partitions, however obtained, to quickly pick out the domains in γ where each partition maximises modularity, thus making it easier to find community features that are robust to changes in γ. The bottom right panel of Figure 27.6 demonstrates this post-processing for the unweighted karate club, finding the community-label partitions corresponding to the line segments along the upper envelope of modularity $Q(\gamma)$, where each line in the diagram corresponds to a single partition of nodes into communities obtained by cluster_leiden(). Recently, Gibson and Mucha (2022) combined the SBM equivalence approach with CHAMP to eliminate the possibility of stochastic fluctuations causing fixed points of the iterative process to go unstable, thus making it easier to find appropriate resolution (and interlayer coupling) parameters.

Once one has a score to optimise, how best to perform the sorting into groups? That is, given a computationally complex optimisation problem, much of the development in community detection has involved coming up with new heuristic search procedures. The earliest models were based on variants of simulated annealing (e.g., UCINET's Factions, de Amorim et al., 1990). Most contemporary approaches use a more targeted search procedure using either a divisive or agglomerative approach. In a divisive approach, one starts with the full network considered as one 'root' community, which is then split by cutting it at its weakest point(s). A commonly used example is Girvan and Newman's (2002) 'edge-betweenness' approach. Edge betweenness is the number of shortest paths between node pairs that cross over that edge. Edges with high betweenness scores generally link parts of the network that are otherwise less linked – that is, they are global bridges. By cutting the network at these points, natural subgroups will fall out. Carried out to its full extent, the process of repeated betweenness calculation and cutting will result in a tree with the whole graph at the root and each node individually at the lowest possible cut level. One way to return a single solution from such a tree is to select the level with highest modularity (Newman & Girvan, 2004).

Agglomerative approaches start with each node assigned to its own community and joins them if doing so improves the objective (e.g., modularity). One generally starts by assigning nodes to some other node they are connected to (with a strong edge weight, if relevant), then joining pairs of nodes that share many neighbours, growing groups until doing so no longer improves fit. Unlike edge betweenness, there is no fixed order to test the joins – particularly early in the process, there are often many equally good assignments so some choices have to be made at random. Much of the effort in these sorts of models is in identifying reasonably good ways to change group assignment, pick which groups to merge and ensuring that doing so doesn't inadvertently walk one into a poor solution.[8] For example, the 'fast greedy' method (Clauset et al., 2004) rapidly increases modularity by merging communities; the Louvain method (Blondel et al., 2008) continues to move one random node at a time between existing

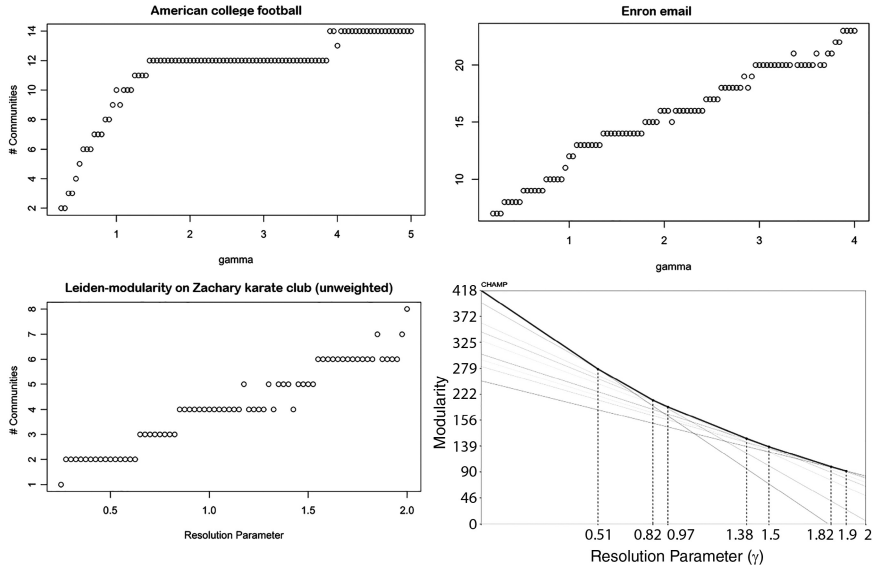

Figure 27.6 Modularity maximisation at different γ resolution parameter values

communities until no further improvement can be found by a single such move and then lifts hierarchically to a smaller graph with nodes corresponding to the communities and repeats; and the recent improvements moving from Louvain to Leiden (Traag et al., 2019) include new checks to ensure that each identified community is a connected component. Ken Frank's KliqueFinder (1995) uses a similar strategy, but optimises the in-group odds ratio for a shared tie.

An alternative heuristic strategy is to allow some process to operate on the network that would itself reveal groups, and then use that result as a proxy for finding groups directly. For example, Moody (2001) reasoned that a peer-influence process should reveal cohesive groups, since within-group ties would push members to hold similar ideas that are distinct from other groups (Friedkin, 1998). This model simulates a peer-influence process over multiple random variables, then clusters nodes based on their resulting scores.[9] Richards' (1995) Negopy program used a similar sorting process based on node IDs as the initialisation stage in its search process, which was then improved on using rule-based sorting techniques. Contemporary methods using this tactic have focused either on a simulated communication process (packet passing methods, such as label propagation (Raghavan et al., 2007)) by following the intuition that peer groups will share more internally to the group than externally, or by simulating random walks on the network, assuming that a random walker will tend to get 'stuck' within communities (Pons & Latapy, 2005).

Eigenvector-based Approaches

Many researchers recognised the qualitative similarity between finding communities in a network and the problem of finding lower-dimensional representations of high-dimensional data objects in general. Intuitively, just as items in a scale should all be more similar to each other than not, peers in a community should all be close to each other (in some sense) in the network. Early approaches (MacRae, 1960; Cairns & Cairns, 1995) used this analogy directly and applied PCA to (some transformation of) the (valued) adjacency matrix. The main advantage of this model was ease of use, but an additional feature is that groups are non-exclusive – nodes with strong ties to multiple groups will have significant loadings on both groups.

This initial simple model has fallen out of favour in recent work (though it still appears sometimes as a starting point for heuristic improvement approaches) in favour of eigenvector-based approaches with more principled mathematical foundations (see, e.g., Chung, 1997). Fiedler (1973) introduced spectral partitioning in terms of the eigenvectors of the (combinatorial) Laplacian matrix.[10] Recently, Priebe et al. (2019) compared and contrasted the types of communities

identified via eigendecomposition of the normalised Laplacian matrix versus those identified in the adjacency matrix. In the simplest versions, the elements of a single eigenvector of the corresponding matrix are used to split the network into two parts, and then the process is repeated recursively on each part. Alternatively, a higher-dimensional decomposition may be used to directly partition into multiple parts (as in Priebe et al., 2019). The two obvious advantages of these approaches are the quick performance on large networks and use of a clear mathematical framework that allows developers to build on the large body of work in related spectral methods. Practically, these models (like all methods) work very well if clustering is clear and seem to work well enough when the problem is data reduction on very large networks, where a node misassignment here or there is not too critical. Because the models work on lower-dimensional summaries of the full network, however, some oddities can occur that are easier to control in the node-sorting approaches discussed above – such as ensuring that each community is a connected component.

Stochastic Block Models

The final common approach leverages advances in the statistical models of networks to identify communities. Stochastic block models (SBMs) are generative models, describing a probabilistic process to generate a network with the characteristics defined by the model (Holland et al., 1983; Wang & Wong, 1987; Snijders & Nowicki, 1997; Lee & Wilkinson, 2019). The simplest SBM with K blocks is given by

$$A_{ij} \left| U_i, U_j \sim Bern\left(U_i^T Q U_j\right) \right.$$

where A is the adjacency matrix, U is an NxK matrix of block membership where $U_{ij} = 1$ if node *i* is a member of block *j* and 0 otherwise, and Q is a KxK matrix defining the probability of ties within and between blocks. If Q is diagonally dominant – for example, Q_{ii} is much greater than Q_{ij} ($i \neq j$) – then blocks are ('assortative') communities. While this is the most common use of stochastic blockmodels, one can specify Q in any configuration. Moreover, one need not assume hard boundaries on block membership: mixed membership SBMs allow nodes to have partial membership in multiple groups that sum to 1 (elements of U are continuous rather than binary) or overlapping models where nodes can be in multiple groups simultaneously (Airoldi et al., 2008). In practice, we usually know A but not U, so the goal is to assign nodes to labels in U in a way that maximises the match with the observed data (see Lee and Wilkinson, 2019, for review).

There are multiple implementations of stochastic blockmodels, which can differ both in formal aspects of the model (assumed distributional properties, degree corrections, hard clustering vs soft, inclusion of observed or latent covariates, for example) as well as the search initialisation (often using methods like k-means or spectral clustering for initial group assignments) and model maximisation routines (MCMC, EM likelihood maximisation, etc.). Practically, models often require users to pre-specify the number of clusters searched for (but see Leger, 2016, which maximises within a user-specified range). Importantly, recent advances have improved the computational performance in fitting SBMs (Peixoto, 2014). Substantively, results are often similar to eigenvector approaches and similarly carry no guarantee of connected clusters, unless that is included in a post-processing step. The canonical form of the model does not have any resolution parameter, so users would have to build that in through appropriate information criteria or, more likely in practice, explore stability by forcing lower or higher numbers of communities.

The key advantage of the SBM approach is the core integration with generative statistical modelling frameworks for networks. This integration allows one to build on the formal insights from that literature, including the ability to test graphs against fundamental limits on community detectability (Abbe, 2018), to test alternative models, and include other data-generating features simultaneously. For example, latent network models (Krivitsky et al. 2009) provide the ability to fit latent-space models with clustering, which allows one to better capture natural heterogeneity in latent-space models of networks.

Each of the three approaches to community detection discussed here – heuristic sorts, eigenvalue decomposition and stochastic blockmodels – have deep roots that could each occupy a review in their own right. We emphasise that real-world data is typically too complex to be conclusively modelled with a single technique. Indeed, we want to re-emphasise our analogy to unsupervised clustering of point-cloud data. When data is inherently well clustered, many different algorithms will find the same or highly similar cluster structures. But many real-world cases are less structured, resulting in different algorithms finding differing results. These general observations are intimately connected to *no free lunch theorems* that identify how the improvement of an algorithm across part of an overall problem domain is balanced by diminished performance

elsewhere (Peel et al., 2017). That is, in practice, some algorithms might be more or less useful in a specific application.

CONCLUSION

With this chapter we have briefly discussed some of the varied approaches that are available for identifying cohesion and communities in network data, using the two-dimensional rubric of connectivity and boundary salience. We advise users to consider the many different available techniques as a collection of exploratory tools that can be used to try and better understand their data,[11] and to think carefully about how the problem a researcher is asking fits into the two dimensions of Figure 27.1. Is your question primarily about connectivity? If so, cohesive blocking or pairwise connectivity scoring might be most appropriate or techniques that allow for overlapping cluster membership. Is your question primarily about boundary salience? Then you should think carefully about what constitutes boundaries in your setting and attempt to identify the socially active variables that define the boundaries. If your problem is simultaneously about boundaries and connectivity, or if boundaries admit to no clear external label, then consider an ensemble of different community detection routines and explore resolution parameter limits thoroughly. Fundamentally, it is up to researchers to ask themselves how specific community labels might match or contradict their theoretical concerns – only then are we able to substantively understand, evaluate and benefit from these approaches.

ACKNOWLEDGEMENTS

Thanks to Ava Scharfstein for assistance generating figures, Gabriel Varela for detailed comments on earlier drafts and Tom Wolff for help implementing comparative clustering routines. This work is partially supported by the NSF through HNDS-I (BCS-2024271, BCS-2140024) and RAPID (SES-2029790) and support from the James S. McDonnell Foundation.

Notes

1 "You can't join Mathletes, it's social suicide!" (Waters et al., 2004).

2 Colour figures online. We note that there are myriad sorts of generalised structures that can be layered over categories, including many types of hierarchies and chain-generalised exchange systems.

3 Moody and White's (2003) original procedure ignored induced graphs that would result in single nodes or strings of cutsets, as these tend to be uninteresting substantively – i.e., there is little value in inducing a subgraph and further examining it if you already know that its composed solely of cutnodes from the parent graph. Similarly, complete cliques are often substantively interesting but there's no reason to go any further.

4 Terms differ significantly across research traditions here. In sociology and social network analysis 'cohesive peer groups' or 'social cliques' or 'crowds' are commonly used to define socially salient network subgroups. But 'group' and 'clique' have specific different meanings in mathematics and computer science, where much of the recent work in this field has originated, and has instead come under the heading of 'community detection'. Here we generally use the terms 'peer group' and 'community' interchangeably unless the context requires greater specificity.

5 We have collected a set of comparisons in table form here: people.duke.edu/~jmoody77/ClusterComparisons.pdf

6 Numerous other such scores exist to be used as objective functions for community detection (see, e.g., the different options available in the leidenalg python package, github.com/vtraag/leidenalg).

7 Football example from Girvan and Newman, 2002, available at www-personal.umich.edu/~mejn/netdata/; Enron email example from R-igraphdata.

8 These sorts of assignment changes and checks are deceptively complicated. One might think it simple to sort nodes to the set where most of their peers are: if ego has mode of peers in group j, make ego's assignment 'j'. But, in practice, assigning one node affects the fit of every other node they are connected to, which leads to a sequence of changes implied by every assignment; a sorting that thus convergences on 'best' is not trivial.

9 In practice, this particular algorithm ends up being quite close to leading eigenvector approaches, so it's probably best to use that instead.

10 The Laplacian matrix is defined as: $L = D-A$; where D is a matrix with (weighted) degrees on the diagonal and 0s elsewhere and A is the adjacency matrix.

11 At the same time, we acknowledge that our perspective is not universally agreed with (see, e.g., Peixoto, 2022, for a very different opinion).

REFERENCES

Abbe, E. (2018). Community detection and stochastic block models: recent developments. *Journal of Machine Learning and Research*, *18*, 1–86.

Airoldi, E.M., Blei, D.M., Fienberg, S.E., & Xing, E.P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning and Research*, *9*, 1981–2014.

Axelrod, R. (1985). *The evolution of cooperation*. Basic Books.

Bearman, P. (1997). Generalized exchange. *American Journal of Sociology*, *102*, 1383–1415.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*, P10008.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., & Wagner, D. (2008). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering* 20:172–88.

Cairns, R.B., & Cairns, B.D. (1995). *Lifelines and risks: pathways of youth in our time*. Cambridge University Press.

Cartwright, D., & Harary, F. (1956). Structural balance: a generalization of Heider's theory. *Psychological Review*, *63*, 277–293.

Chung, F. (1997). *Spectral graph theory*. Vol. 92. CBMS Regional Conference Series in Mathematics. American Mathematical Society. doi.org/10.1090/cbms/092

Clauset, A., Newman, M.E.J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*, 066111.

Coleman, J. (1961). *The adolescent society*. Free Press.

Cornwell, B., & Burchard, J. (2019). Structural cohesion and embeddedness in two-mode networks. *Journal of Mathematical Sociology*, *43*, 179–194.

de Amorim, S.G., Barthélemy, J.P., & Ribeiro (1990). Clustering and clique partitioning: simulated annealing and tabu search approaches. Research report from Groups d'études et de recherche en analyse des décisions. HEC Montreal, McGill University.

Fenn, D.J., Porter, M.A., McDonald, M., Williams, S., Johnson, N.F., & Jones, N.S. (2009). Dynamic communities in multichannel data: an application to the foreign exchange market during the 2007–2008 credit crisis. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *19*, 033119–8.

Fershtman, M. (1997). Cohesive group detection in a social network by the Segregation Matrix Index. *Social Networks*, *19*, 193–207.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, *23*, 298.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*, 75–174.

Fortunato, S., & Hric, D. (2016). Community detection in networks: a user guide. *Physics Reports*, *659*, 1–44.

Fosdick, B.K., Larremore, D.B., Nishimura, J., & Ugander, J. (2018). Configuring random graph models with fixed degree sequences. *SIAM Review*, *60*, 315–355.

Frank, K.A. (1995). Identifying cohesive subgroups. *Social Networks*, *17*, 27–56.

Freeman, L.C. (1972). Segregation in social networks. *Sociological Methods and Research*, *6*, 411–430.

Freeman, L.C. (1992). The sociological concept of 'group': an empirical test of two models. *American Journal of Sociology*, *98*(1), 152–166.

Friedkin, N.E. (1998). *A structural theory of social influence*. Cambridge University Press.

Friedkin, N.E. (2004). Social cohesion. *Annual Review of Sociology*, *30*, 409–425.

Friedkin, N.E., & Cook, K.S. (1990). Peer group influence. *Sociological Methods and Research*, *19*(1), 122–143.

Gibson, R., & Mucha, P.J. (2022). Finite-state parameter space maps for pruning partitions in modularity-based community detection. *Scientific Reports*, *12*(1). doi.org/10.21203/rs.3.rs-1551680/v1

Girvan, M., & Newman, M.E.J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 7821–7826.

Guimerà, R., Sales-Pardo, M., & Nunes Amaral, L.A. (2004). Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, *70*, 025101.

Granovetter, M. (1985). "Economic Action and Social Structure: The Problem of Embeddedness." *American Journal of Sociology*, *91*(3), 481–510. http://www.jstor.org/stable/2780199

Harary, F. (1969). *Graph theory*. Addison-Wesley

Holland, P.W., Laskey, K.B., & Leinhardt, S. (1983). Stochastic blockmodels: first steps. *Social Networks*, *5*, 109–137.

Kanevsky, Arkady. (1990). "On the Number of Minimum Size Separating Vertex Sets in a Graph and How to Find All of Them." *Proceedings of the 1st ACM-SIAM Symposium on Discrete Algorithms* :411–21.

———. (1993). "Finding All Minimum-Size Separating Vertex Sets in a Graph." *Networks* 23:533–41.

Kivela, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., & Porter, M.A. (2014). Multilayer networks. *Journal of Complex Networks*, *2*, 203–271.

Kreager, D.A., Rulison, K., & Moody, J. (2011). Delinquency and the structure of adolescent peer groups. *Criminology*, *49*, 95–127.

Krivitsky, P.N., Handcock, M.S., Raftery, A.E., & Hoff, P.D. (2009). Representing degree distributions,

clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, *31*, 204–213.

Lee, C., & Wilkinson, D.J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, *4*, 122.

Leger, J. (2016). Blockmodels: A R-package for estimating in Latent Block Model and Stochastic Block Model, with various probability functions, with or without covariates. arXive: https://doi.org/10.48550/arXiv.1602.07587

Leger, J.-B., Barbillon, P., & Chiquet, J. (2021). blockmodels: latent and stochastic block model estimation by a 'V-EM' algorithm. *R Package*. cran.r-project.org/package=blockmodels

MacRae, J. (1960). Direct factor analysis of sociometric data. *Sociometry*, *23*, 360–371.

Mani, D., & Moody, J. (2014). Moving beyond stylized economic network models: the hybrid world of the Indian firm ownership network. *American Journal of Sociology*, *119*, 1629–1669.

Markovsky, B., & Lawler, E.J. (1994). A new theory of social solidarity. In B. Markovsky, J. O'Brien & K. Heimer (Eds.), *Advances in group processes*, vol. 11 (pp. 113–137). JAI Press.

Moody, J. (2001). Peer influence groups: Identifying dense clusters in large networks" *Social Networks* 23:261–283.

Moody, J., & White, D.R. (2003). Structural cohesion and embeddedness: a hierarchical concept of social groups. *American Sociological Review*, *68*(1), 103–127.

Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., & Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, *328*, 876–878.

Newman, M.E.J. (2016). Equivalence between modularity optimization and maximum likelihood methods for community detection. *Physical Review E*, *94*(5), 052315.

Newman, M.E.J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*, 026113.

Pamfil, A.R., Howison, S.D., Lambiotte, R., & Porter, M.A. (2019). Relating modularity maximization and stochastic block models in multilayer networks. *SIAM Journal on Mathematics of Data Science*, *1*, 667–698.

Peel, L., Larremore, D.B., & Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances*, *3*, e1602548.

Peixoto, T.P. (2014). The graph-tool python library. figshare. doi: 10.6084/m9.figshare.1164194 [sci-hub, @tor]

Peixoto, T.P. (2022). Descriptive vs inferential community detection: pitfalls, myths and half-truths. *arXiv*, 2112.00183. doi.org/10.48550/arXiv.2112.00183

Pons, P., Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. In: Yolum, p., Güngör, T., Gürgen, F., Özturan, C. (eds) Computer and Information Sciences - ISCIS 2005. ISCIS 2005. Lecture Notes in Computer Science, vol 3733. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11569596_31

Porter, M.A., Onnela, J.-P., & Mucha, P.J. (2009). Communities in networks. *Notices of the AMS*, *56*, 1082–1097 & 1164–1166.

Priebe, C.E., Park, Y., Vogelstein, J.T., Conroy, J.M., Lyzinski, V., Tang, M., Athreya, A., Cape, J., & Bridgeford, E. (2019). On a Two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences*, *116*, 5995–6000.

Raghavan, U.N., Albert, R. & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76, 036106

Reichardt, J., & Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, *74*, 016110.

Richards, W.D. (1995). *NEGOPY 4.30: manual and user's guide*. Simon Fraser University.

Shai, S., Stanley, N., Granell, C., Taylor, D., & Mucha, P.J. (2020). Case studies in network community detection. In R. Light & J. Moody (Eds.), *The Oxford handbook of social networks* (pp. 309–333). Oxford University Press.

Simmel, G. ([1908]1950). The dyad and the triad. In K.H. Wolff (Ed. & Trans.), *The sociology of Georg Simmel*. Free Press.

Sinkovitz, R.S., Moody, J., Oztan, B.T., & White, D.R. (2017). Fast determination of structurally cohesive subgroups in large networks. *Journal of Computational Science*, 62–72.

Snijders, T., & Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, *14*, 75–100.

Traag, V.A., Waltman, L., & van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*, 5233.

Wang, Y.J., & Wong, G.Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, *82*, 8–19

Waters, M.S., Michaels, L., Fey, T., Lohan, L., McAdams, R., Meadows, T., Poehler, A., Gasteyer, A., Chabert, L., & Wiseman, R. (2004). *Mean girls*. Paramount.

Weir, W.H., Emmons, S., Gibson, R., Taylor, D., & Mucha, P.J. (2017). Post-processing partitions to identify domains of modularity optimization. *Algorithms*, *10*, 93.

White, D., & Harary, F. (2001). The cohesiveness of blocks in social networks: node connectivity and conditional density. *Sociological Methodology*, *31*, 305–359.

White, H.C. (1963). *An anatomy of kinship: mathematical models for structures of cumulated roles*. Prentice-Hall.

White, H.C. (1966). Catnets. Notes on the constituents of social structure. Unpublished manuscript.

Zachary, W.W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, *33*, 452–473.