Community structure and natural kinds

Whit Schonbein
whit.schonbein@gmail.com

February 18, 2018

1 Introduction

According to *property cluster* accounts of natural kinds, a natural kind is a collection or family of properties that, for some reason or another (e.g., causal relations), tend to cluster together (Boyd, 1991; Khalidi, 2015). Because these properties cluster, they ground inductive inferences of the sort characteristic of scientific explanation, e.g., the projectability of predicates. For example, we are warranted in inferring from the fact instances $x_0, x_1, \ldots, x_n$ satisfy predicate P to the conclusion all instances of that type satisfy P because these instances instantiate a collection of properties such that those properties considered independently of P tend to also cause the instantiation of P.

In a recent paper, Khalidi (2015) introduces what appears to be a *graph-theoretic* version of a cluster account, according to which natural kinds are viewed as nodes in a causal network satisfying certain conditions (described below). While this is a promising strategy for elaborating on cluster approaches, Khalidi's analysis fails to capitalize on the benefits of introducing graph theory. In this paper, I appeal to the notion of *community structure* to further develop the graph-theoretic approach, and show how doing so can lead to a more robust understanding of natural kinds.

The structure of this paper is as follows. The graph-theoretic components of Khalidi's account are briefly summarized in section 2. In section 3 I extend this analysis by appealing to other features of graphs beyond those identified by Khalidi. Finally, in section 4, I discuss some of the benefits and limitations of this extension.

2 Khalidi's account

As noted above, the core of Khalidi's approach is to treat natural kinds (NKs) as nodes in a causal network. Focusing on just its graph-theoretic aspects, the account appears to have two primary components. First, empirical science identifies properties and causal relations between them, representable as directed causal graphs. For instance, in figure 1, property p_0 causes properties p_1, p_2 and p_5, and so on (cf. Khalidi, 2015, figure 1).

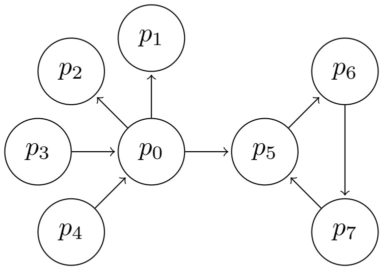

Figure 1

The second component is a graph-theoretic criterion on 'naturalness'. According to Khalidi, "[natural kinds] are represented by those vertices in directed causal graphs from which many edges originate, whether directly or indirectly, leading to other vertices" (Khalidi, 2015, p. XXXX). In graph theory, the number of edges leaving a vertex is known as the 'outdegree' of that vertex; so, the proposal is that the greater the outdegree of a vertex in a causal graph, the stronger a candidate natural kind that predicate is. For example, in figure 1, p_0 has the greatest outdegree (3) and hence is the most natural; this is followed by p_3 through p_7 (1 each); the remaining vertices have outdegrees of zero, so are weak candidates for designating natural kinds under Khalidi's criterion.[1]

There is much more to Khalidi's account, but for present purposes this is sufficient to illustrate at least one reason we might find it problematic. In Khalidi's approach, NKs are identified with *individual* vertices in a graph. However, Khalidi explicitly invokes clusters when introducing his account (p. 1). Therefore, it seems NKs ought themselves to be represented as *subgraphs* of interconnected properties rather than as individual vertices.

One solution to this conundrum is to allow vertices to represent *conjunctive* properties, i.e., a single vertex can stand for a cluster of properties and hence a NK (this seems to be the approach adopted by Khalidi). However, on the one hand, if we expand one of these vertices into its subgraph of causally connected component properties, then the criteria of naturalness – outdegree – applies to each of the vertices in this subgraph, and the putative natural kind no longer attaches to the cluster of properties, but rather to those vertices in the subgraph with the largest outdegree. On the other hand, if we leave the vertex intact, we are left with no explanation of why *these* properties are conjoined rather than some other set of properties, apparently defeating the original purpose of appealing to clusters to help understand NKs.

An alternative solution is to eschew logical connectives in favor of formal properties of the graphs themselves. In the remainder of this paper, I elaborate on this alternative.

[1] Khalidi leaves the notion of an 'indirect' outdegree undefined, so I bracket indirectness for the remainder of this paper.

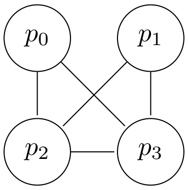

Figure 2

3 Community structure

Khalidi introduces graph theory to invoke outdegree as a measure of natural kindness. This opens the door to considering other graph-theoretic features for the same purpose. Most notably, some of these features specifically address the identification of *clusters* of properties (a.k.a., 'community structure'), and hence may be capable of shedding light where outdegree fails.

One possibility is to consider the *clustering coefficient* of subgraphs of a graph. A triple of vertices is most tightly coupled when each vertex connects to the others (as opposed to, e.g., forming a linear sequence). The clustering coefficient of a graph, then, is the ratio of closed triples of vertices to unclosed triples. The more tightly coupled a graph (i.e., the closer it comes to every vertex being connected to every other vertex), the higher its clustering coefficient. On this measurement, then, the higher the clustering coefficient for a collection of properties, the more natural that cluster is.

If this strategy is to be used as an account of NKs, then there are technical issues that must be addressed. For example, clustering coefficients are independent of the number of vertices in the graph: if a graph contains 20 fully interconnected vertices, then the coefficient for that graph is one, which is also the coefficient for each of its subgraphs of size three or more. In this case, all clusters are equally natural, so the coefficient provides no reason for preferring any particular subset over another – does the graph represent a single NK, $\binom{20}{3}$ NKs, or some number in between?

One way to address this issue is to include size as an additional criterion: we prefer the 20 vertex graph with a coefficient of one to its subgraphs with the same coefficient because the former is more inclusive. However, it seems theoretically possible that a natural kind cluster might not be fully interconnected. For example, perhaps the cluster depicted in figure 2 is properly viewed as comprising a single NK; according to the clustering-coefficient-plus-maximum-size criterion, the graph is restricted to *two* NKs: $\{p_0, p_2, p_3\}$ and $\{p_1, p_2, p_3\}$.

An alternative method for identifying clustering in networks is *modularity maximization* (Clauset, Newman, and Moore, 2004). The insight behind this method is that a random network can have vertices connecting to any other vertex, while a network with clusters will have vertices whose edges tend to lead to other vertices in the same cluster rather than to those outside. The algorithm discovers clusters by identifying a partition of vertices whose interconnections are furthest away from the random case. An example is given in figure 3, with

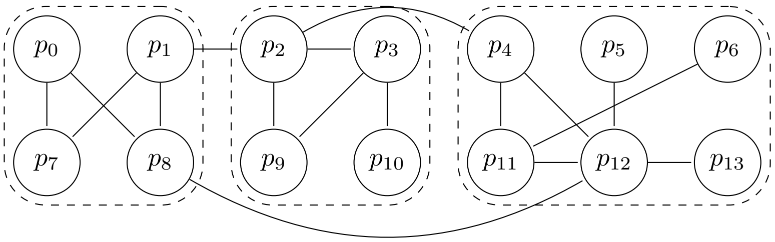

Figure 3

the clusters identified by modularity maximization indicated by dashed boxes.

This approach to community structure has the benefit of allowing clusters that are not fully interconnected; indeed, under modularity maximization, it is possible for collections with clustering coefficients of zero to nonetheless be viewed as forming clusters (e.g., the leftmost group in figure 3). As with clustering coefficients, the method can be applied hierarchically, so subsubgraphs can form clusters. However, the approach also has its own set of technical challenges. For example, while it is possible for the method to yield partitions with a single member, in many cases every vertex is grouped with at least one other, i.e., the method may be too liberal in its identification of NKs. Likewise, in the non-hierarchical case, partitions are mutually exclusive, so no property partakes in more than one possible NK. One possible strategy for dealing with this limitation is to allow the same property to appear multiple times in a graph; however, doing so means the graph fails to capture all known causal links, thereby leading to potential errors.

Other methods for identifying clusters allow both vertices with no cluster membership and overlapping structure. A well-known approach is the *maximal clique* method (Palla et al., 2005). A clique is a graph where every vertex is connected to every other vertex, and its size is simply the number of vertices it contains. The maximal clique method identifies the largest cliques in a graph; vertices that are not members of any of these cliques are not part of any cluster. Furthermore, since a single vertex can participate in multiple cliques, the approach acknowledges the possibility that natural kind clusters may overlap. For example, vertex p_4 in figure 4 is a member of two cliques (marked with dashed boxes), indicating the property it represents may participate in multiple NKs, while p_9 is excluded from any cluster.

As expected, treating clusters as maximal cliques also has shortcomings, most notably that by definition the approach requires every member of a cluster stand in direct causal relation with every other member. So, like the clustering coefficient approach, maximal cliques exclude the possibility of clusters where some properties only influence some members indirectly, as in the case of causal cycles.

Table 1 summarizes some of the strengths and weakness of the various ap-

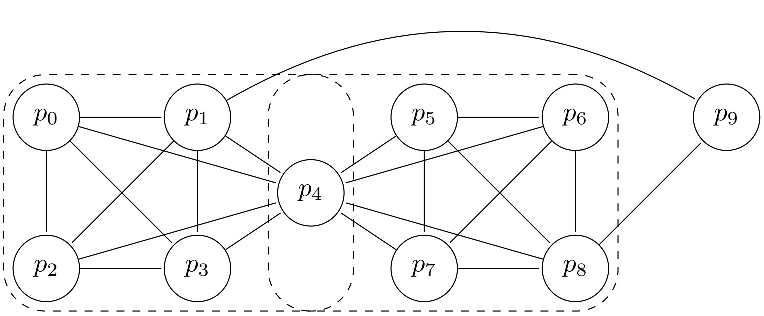

Figure 4

Method	Hierarchical	Overlapping structure	Tolerates disconnections	Allows outliers
Clustering coefficient	✓	✓	x	✓
Modularity maximization	✓	x	✓	x
Maximal clique	x	✓	x	✓

Table 1

proaches touched on in this section.[2] As is clear from this summary, none of the methods mentioned here appear to meet all of the intuitive desiderata for an account of NK clusters. However, these examples nonetheless prove the point of this paper: graphs contain potentially useful structural features beyond mere outdegree.

4 Discussion

Khalidi (2015) identifies some of the benefits of pursuing a graph-theoretic approach to NKs based on outdegree. I believe the present proposal preserves these benefits while (perhaps) adding others. For example, recognizing that graphs contain clusters defined by relations between vertices can help us understand why some causal connections are constitutive of a NK while others are not. For instance, in figure 3, p_8 and p_{12} are not in the same cluster despite being connected. The reason is the graph is *less ordered* when those two vertices are included in the same cluster as opposed to those indicated in the figure. If NKs are supposed to reflect islands of organization in a sea of noisy interaction,

[2] Notes: (1) as noted above, the clustering coefficient method allows disconnections, but doing so decreases the coefficient of the supergraph in comparison to (some) of its subgraphs. So, if naturalness is a function of clustering coefficient, this has the effect of pushing the potential NKs to the level of the subgraphs. (2) To simplify exposition, I've assumed unweighted, undirected graphs with no self-loops. As far as I know, all of the methods surveyed in this section can be adapted to handle weighted (e.g., probabilistic) edges, directed edges, and vertices with edges pointing to themselves.

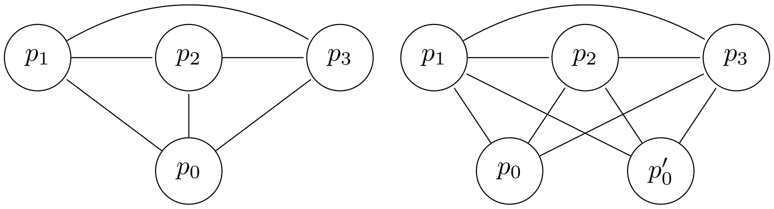

Figure 5

then this is the sort of explanation we want.[3]

Another benefit of introducing community structure is that doing so can help us grasp how our understanding of NKs can change over time. As noted in section 2, graphs are representations of our current state of empirical knowledge. As this knowledge changes, the graph changes, impacting community structure. Consequently, the very same criteria applied to the new graph may result in modified boundaries for NK clusters.

For example, a graph might change through a process of 'property fission', where some predicate used by empirical science is subsequently determined to designate two distinct properties (the classic example being 'jade'). Assuming for purposes of illustration that the relevant notion of a cluster is that of a clique, figure 5 illustrates how property fission can lead to kind fission. In the left graph, property p_0 participates in a clique with the other vertices, so the entire graph represents a candidate NK cluster. However, it is decided that the predicate taken to designate p_0 actually designates two distinct properties, p_0 and p'_0. These properties nonetheless still stand in the same causal relations with the original properties in the cluster (right figure). Given the clique criterion for clustering, the result of the property fission is two candidate NK clusters, $\{p_0, p_1, p_2, p_3\}$ and $\{p'_0, p_1, p_2, p_3\}$. As this example illustrates, by appealing to the notion of community structure, we may be able to explicate what it means for kinds to undergo bifurcation during the process of empirical inquiry.[4]

Regardless of these benefits, someone might object that by focusing on graph-theoretic criteria for community structure I've implicitly (and unacceptably) changed the terms of the discussion. Consider the network depicted in figure 6. On Khalidi's outdegree criterion, vertices p_0, p_4, and p_{12} are maximally natural (outdegree = 8) and hence are stronger candidate NKs than any of the others (outdegree = 1). In contrast, since there are zero cliques or closed triangles in the graph, the maximal clique and clustering coefficient methods identifiy zero candidate NKs, while modularity

[3] The maximal clique approach also provides an explanation for why some properties are included while others are not, even when there exists a causal connection: any property not part of a clique is not part of a NK.

[4] Another type change that can influence community structure is property elimination (i.e., a vertex that was supposed to represent a 'real' property actually represents nothing at all (Churchland, 1981)). This case is handled in the same way by a community structure approach.

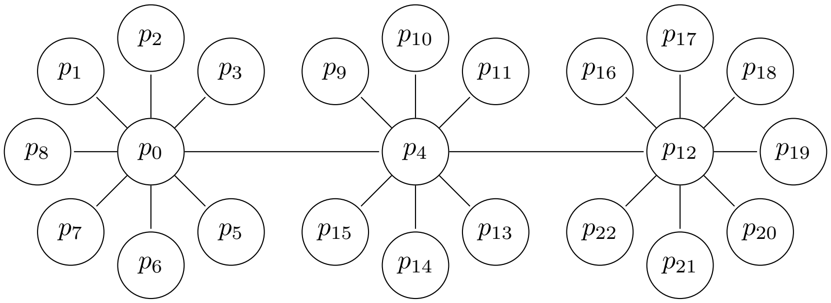

Figure 6

maximization forms three clusters centered on those vertices with the maximal outdegrees: $\{p_0, p_1, p_2, p_3, p_5, p_6, p_7, p_8\}$, $\{p_4, p_9, p_{10}, p_{11}, p_{13}, p_{14}, p_{15}\}$, and $\{p_{12}, p_{16}, p_{17}, p_{18}, p_{19}, p_{20}, p_{21}, p_{22}\}$. The objection, then, is that these differing results indicate the discussion of community structure offered in this paper is off-topic: Khalidi's account is not concerned with *clusters*, but rather with how important a vertex is with respect to connecting other vertices, i.e., with *centrality*.

For example, a popular measurement of the centrality of a vertex v_i is *betweenness*, defined as the ratio of (i) the number of shortest paths between any other two vertices v_j and v_k passing through v_i to (ii) all shortest paths between v_j and v_k. The closer this value is to one, the the stronger the role played by v_i in facilitating indirect relations between vertices to which it is connected. On this measurement of centrality, as with outdegree, vertices p_0, p_4, and p_{12} play the most important role in facilitating interactions between other vertices in the graph, so are the primary candidates for NKs. Khalidi's approach thus appears to focus on centrality rather than community structure as a criterion for naturalness.

There are several possible responses to this disagreement. First, just as there are multiple criteria for clustering, centrality does not provide a unified account of NKs. For instance, in figure 7, vertices p_0, p_4, and p_{12} have the highest outdegree, but p_3 is most central as measured by betweenness. So, just to be clear, even if Khalidi's account is indeed centrality-based rather than cluster-based, it faces similar challenges.

Second, community structure seems to do a better job of accounting for the projectability of predicates than centrality. For example, on a clustering approach we can infer from the fact instances $x_1, x_2, \ldots, x_n$ of type T satisfy predicate P to the conclusion all instances of T satisfy P because the properties by which we categorize these entities as members of T form a causal cluster, and this cluster includes the property designated by P. Likewise, we can infer from the fact an instance of T satisfies P to the conclusion it also satisfies Q because the properties designated by P and Q (along with others used to categorize the instance) form a causal cluster. In contrast, it is unclear how the centrality of a property (measured by outdegree or by betweenness) is supposed to generate

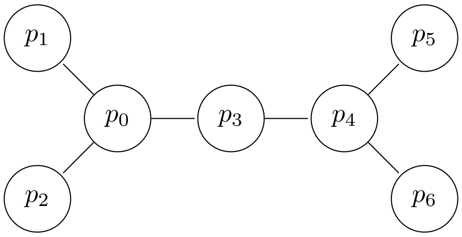

Figure 7

the same intuitive explanations of projectablity.

Third, centrality and community structure are not mutually exclusive; there is no reason we couldn't combine these metrics to articulate a 'two-dimensional' graph-theoretic account of NKs. For instance, one might deploy a maximal clique analysis to identify candidate clusters, followed by a betweeness analysis to provide a centrality assessment of those clusters. On this basis, non-maximal cliques might be judged 'more natural' because of their centrality, while highly-central vertices are discounted because they do not participate in any cliques. In short, by appealing to graph theory, it may be possible to aspire to the best of both worlds when articulating an account of natural kinds.

References

Boyd, Richard (1991). “Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds”. In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 61.1/2, pp. 127–148. ISSN: 0031-8116.

Churchland, Paul M. (1981). “Eliminative materialism and the propositional attitudes”. In: *Journal of Philosophy* 78, pp. 67–90.

Clauset, Aaron, M. E. J. Newman, and Cristopher Moore (2004). “Finding community structure in very large networks”. In: *Physical Review E* 70.6. arXiv: cond-mat/0408187. ISSN: 1539-3755, 1550-2376. DOI: `10.1103/PhysRevE.70.066111`. URL: `http://arxiv.org/abs/cond-mat/0408187`.

Khalidi, Muhammad Ali (2015). “Natural kinds as nodes in causal networks”. en. In: *Synthese*, pp. 1–18. ISSN: 0039-7857, 1573-0964. DOI: `10.1007/s11229-015-0841-y`.

Palla, Gergely et al. (2005). “Uncovering the overlapping community structure of complex networks in nature and society”. en. In: *Nature* 435.7043, pp. 814–818. ISSN: 0028-0836. DOI: `10.1038/nature03607`.