**Note to reader: This paper, which recently appeared in *Synthese* (May 6, 2025) is a later version of an earlier paper with the same title which was added to the philsic archive in 2023. This version differs significantly from the earlier version. Please quote from the published version**.

**Networks, Dynamics and Explanation**

**James Woodward**

**1. Introduction**

Networks are used for many different purposes in science -- for example, to provide diagnoses, as when differences in functional connectivity in brains, represented by networks, are used to distinguish subjects with mental disorders from those who are healthy. Networks are also used to classify diseases on the basis of patterns of co-occurrence or to group proteins together on the basis of whether they interact with other proteins. But in addition, networks are also used to explain. The main focus of this paper is one very common form that such explanations can take. Put informally these are cases in which what is of interest is the spread of some process[[1]](#footnote-1) along a network and the explanandum is some *overall* behavior of the system characterized by the network (as opposed, say, to why some particular node in the network is in a certain state)[[2]](#footnote-2). Examples include the spread of a disease along a contact network (explaining why the disease spreads rapidly or becomes an epidemic), the propagation of neural signals along structural or anatomical connections in a brain (explaining why these are associated with the presence of some mental disorder), the spread of beliefs along social networks (explaining the presence of extensive misinformation), and movement (of cars, planes etc.) along transportation networks (explaining patterns of traffic congestion). Examples of this sort represent important uses of networks and are extensively discussed in both the scientific and philosophical literature. Although, as noted below, they are not the only explanatory uses of networks, the examples share the following feature: they appeal *both* to network structure and to independent dynamical assumptions governing the processes occurring along the network (the dynamics of the propagation of the disease etc.). I will use "independent dynamics explanations" (IDEs) as a label to cover these but I caution that in using this label I do *not* mean that such cases are disjoint from or have nothing in common with explanations that appeal only to network structure. (See below).

 One of my main claims will be that many standard cases in which networks are used to explain are cases in which the explanation does not work by appealing to network or topological structure alone but rather by the combination of this structure and dynamics working together[[3]](#footnote-3). This is one of several reasons why at least in many cases these explanations should not (as sometimes claimed) be viewed as "autonomous” or "stand alone" in the sense that they are independent of causal information or dynamical assumptions.

 That said, I will also emphasize that the extent to which dynamics matter is a matter of degree in the following sense: In some cases, explananda of interest may be very sensitive to the details of the dynamics-- given a fixed network structure, explananda or outcomes (or some of these) will vary considerably depending on changes in the governing dynamics. In other cases, for a fixed network structure, variations in the dynamics within a certain range may make little or no difference to outcomes, although variations outside this range may make a big difference.

 Next a note on terminology: Networks can figure in other sorts of explanations besides those on which I focus in this paper. For example, network structure can figure in explanations of why various systems are robust or fragile under deletions of various nodes or edges. My view is that in these cases too the resulting explanations require additional claims that go beyond the network structure itself, so that these also are *not* cases in which only network structure does the explaining[[4]](#footnote-4). However, for reasons of space, I will not argue for this conclusion here and the claims that follow are not intended to apply to examples of this sort. Thus, when I talk of "network explanations" in what follows I will mean explanations like those considered below, in which either (i) network structure and dynamics both play an explanatory role or (ii) only network structure plays an explanatory role.

 The philosophical literature often talks about distinctively topological explanations (DTEs) where often it seems to be assumed that these are cases in which network structure alone does the explanatory work, with dynamics playing no role. If DTEs are understood in this way and IDEs are understood as cases in which both network structure and dynamics play a non-trivial role, then by definition IDEs and DTEs represent distinct, indeed disjoint categories. This distinction by stipulation would make sense if the networks that figure in IDEs were very different in important respects from the networks that figure in DTEs and made very different explanatory contribution to outcomes in the two cases. However, in my view, this is not the case. If the account presented in Section 2 is correct, the information contained in a network representation and its explanatory contribution are the same whether the network figures in a IDE or a DTE (understood in the restricted way above) -- in both cases the information has to do with the possibility of causal influence between nodes.

 For these reasons, the position adopted in this essay is *not* that explanations that appeal to both network structure and dynamics and those that appeal only to network structure are disjoint and unrelated; instead, the latter are best thought of as a special or limiting case of the former (IDEs). The latter are special cases in the sense that they are cases in which a parameter or structure specifying the dynamics in the more generic case in which both network structure and dynamics matter is (metaphorically) set to zero or removed . This picture is also what is naturally suggested by my observation above that the extent to which the dynamics matters in network explanations is one of degree. Thus, I will understand IDEs to encompass a general category that includes explanation in which network structure matters and in which dynamics may or may not matter and may matter to varying degrees.

 We can capture this picture by thinking of an IDE as an ordered triple <*N, D,* *E*> in which there is a slot, parameter or argument specifying the network structure *N*, a slot, parameter or argument specifying a dynamics *D* and a specification of an explanandum *E*. When the dynamics plays no role in the sense that *E* does not depend at all on *D*, think of this as a special case in which the *D* parameter is set to zero or the <*N, D, E*> structure is projected on to a lower dimensional <*N, E*> structure.[[5]](#footnote-5) I emphasize that this is just a way of representing things; the *motivation* for adopting this way of looking at things is what I describe in the preceding paragraph-- in both cases in which the dynamics plays a non-trivial role and cases in which it does not, network structure seems to function in the same way and the importance of the dynamics is one of degree.

 It is true that DTEs, understood in the restrictive sense, have some distinctive features-- for example, they may be (perhaps often are) explanations of impossibilities[[6]](#footnote-6), while these are not the most typical explananda of IDEs. On the other hand, according to the account that I will present there is nothing mysterious about this and it does not warrant the claim that DTEs belong in a separate category. As argued in Section 5, the fact that network structures can sometimes explain impossibilities follows directly from my claims in Section 2 about their representing possibilities of interaction. Thus (I emphasize) my view is *not* that explanations that appeal to network structure alone do not explain-- they do explain and provide complete explanations when it is true that their explananda depend only on network structure and not the dynamics. However, as many of the examples discussed in Section 4 illustrate, many explanations that appeal to networks also make use of dynamical assumptions. This is true of examples that are often presented in the philosophical literature as appealing to network structure alone.

 The rest of this paper is organized as follows. Section 2 introduces the idea that in many characteristic applications networks represent constraintsconcerning which components of a system can interact with others. Section 3 argues that such constraints are supported by physical structures. Section 4 provides further detail in support of the idea that in an important range of cases networks are combined with an independent dynamics to explain system behavior, a claim which is further illustrated in an Appendix. Sections 5 contrasts cases in which both dynamics and network structure are important with a case in which dynamical considerations are absent or at least non-salient. Section 6 draws out some implications of the independent roles of network structure and dynamics in many explanations in which networks figure. Section 7 proposes a general schema for such explanations and contrasts this with some alternative proposals. Section 8 criticizes the claim that network explanations appeal to properties (so-called topological properties) that are non-physical, non-empirical or non-causal. Section 9 explores the extent to which network explanations are "autonomous" or "stand alone". Section 10 takes up issues having to with the "direction" of network explanations, arguing that my earlier claims about the independence of network structure and dynamics helps us to understand these. Section 11 explores the relatively neglected issue of why and when network explanations are fruitful, and Section 12 concludes.

**2. Networks as Representing Constraints[[7]](#footnote-7)**

 Network representations are most fruitful when applied to systems with many (but not too many) components (represented by nodes) whose possibilities of interaction are restricted and structured in various ways-- that is, each component can interact with certain other components but not with others. Edges between nodes represent such possibilities of interaction -- they represent constraints on what can influence what, telling us which interactions are possible or impossible. Moreover, although networks represent heterogeneity in what can interact with what via the presence or absence of edges, the nodes and edges themselves are typically not further differentiated[[8]](#footnote-8)-- differences in their internal structure are not represented.

 Consider a light connected to conducting wires in a certain configuration and a battery. A network description of the circuit configuration describes constraints on the possible current flow and other outcomes-- current can only flow through the wires, the light will only go on if it is connected via the wires to the battery and so on. However, this network structure merely constrains or limits some of the possible behaviors of the system without fixing all such behaviors-- the actual current flow through the circuit will depend not just on the circuit configuration but also on the battery voltage and the resistance in the circuit, as captured by relations like Ohm's law. The latter information has to do with what I will call dynamics, discussed below.

 As another illustration consider a network representing relations of close spatial contact within a population. This may be used to describe limits or constraints on how a disease that can only be transmitted by contact can spread in the population-- even if *X* is infected, *X* cannot infect *Y* if *X* and *Y* are not in spatial contact, as represented by an edge connecting *X* and *Y*. If there is no path (that is, a sequence of edges) from an infected individual to *Y*, *Y* cannot be infected at all. However, even if *X* and *Y* are in spatial contact, whether *X* in fact infects *Y* will typically depend on other considerations not represented by the contact network: the transmission probability for the disease, whether those who get the disease and recover can become reinfected and so on. (See Section 4 and Appendix.) In this respect, what the edge from *X* to *Y* represents is something like the possibility of transmission.

 Systems in which there are constraints on which components can affect others contrast with systems in which each component can interact with any other, as with the molecules comprising a dilute gas. Even apart from the astronomically large number of edges and nodes that would be required, a network representation in this case will not be useful because there are no constraints on which components can interact with others-- the representation would require a fully connected network. A similar point holds for a "fully mixed" susceptible/infected model in disease epidemiology in which each individual in a population has an equal probability of contacting (and serving as a source of infection for) any other. Network representations are not used in this case because there are no constraints that distinguish individuals as to the possibility of their being involved in disease transmission. Instead, network models are employed in epidemiology when diseases spread by contact and there are restrictions on which people are in contact with others[[9]](#footnote-9).

 Not all networks represent constraints or (im) possibilities of interaction. For example, in a network representing *functional* connectivity in the brain, an edge just corresponds to the presence of a temporal correlation in activity between different brain areas, without there necessarily being any causal influence or interaction between these areas. (The correlation might arise because both areas are influenced by some third area.)

**3. Constraints are Supported by Physical Structures**

 When an edge represents such constraints, often -- perhaps always-- there is a corresponding physical structure associated with the edge or at least some set of empirical facts that, so to speak, support the constraint in question[[10]](#footnote-10). In the case of a brain, these might be an appropriate set of anatomical or "structural" connections-- e.g., neurons or white matter tracts linking two brain regions. If we are interested in traffic flow, one network of interest will likely be the road network, which is a straightforwardly physical structure which constrains how traffic can move[[11]](#footnote-11). In the case of a network representing friendship relations, nothing as concrete as a neuron or a road will be present, but there will be some underlying physical facts bearing on the construction of the network-- friends will exchange messages, likely will be in spatial proximity on occasion and so on. Information about this allows researchers to "estimate" (that is, discover) a friendship network.

 However, although there are physical structures that typically support the constraints associated with a network, it is important to keep in mind that (as far as the network goes) these physical structures are relevant only to the extent that they afford such constraints. For example, anatomical connections in brains are relevant to the extent that these constrain how neural signals propagate-- there are lots of anatomical, spatial or structural relations in brains that are irrelevant to signal propagation and would not be appropriately represented in a network involved in neural processing. For this reason, it is usually misleading to think of edges in networks as representing purely spatial or geometrical connections. Instead, it is more appropriate to think of the edges as capturing or encoding restrictions on how nodes can influence one another, but where material facts about the network restrict which influences are possible.

 In describing edges as encoding information about which nodes can influence others, I've been employing causal language since "influence" (and "interaction" etc.) are causal notions. But the causal commitments carried by this characterization are very minimal-- what is conveyed by the presence of an edge is simply that one node can or will influence another in some way or other, with no further commitments about the nature of that influence or exactly what is transmitted along the edge or, for that matter, the detailed internal structure of either the node or the edge. Use of networks allows us to "blackbox" these details. For example, a network representing relations among friends conveys that there are interactions among the friends but what form this takes is not further specified-- it may involve the exchange of messages or information via the internet, face to face conversations or the transmission of an infectious diseases.

 The fact that networks usually don't incorporate highly detailed causal information is connected to the fact that we often have ways of identifying their structure independently of such information and independently of the dynamics occurring along the network. For example, anatomical investigation, tracer methods and diffuse tensor imaging can identify structural connectivity networks in the brain independently of detailed information about how neural signals are propagated along such connections. Similarly, we can identify a contact network from information about spatial proximity among people without having detailed information about how disease is spread along that network (beyond the fact that it spreads by spatial contact). Particularly if there is reason to think that, given a network structure, outcomes will be largely the same for a range of different "reasonable" dynamics or, alternatively, if what we are interested in just how, for a given dynamics, variations in network structure affect outcomes, the fact that we have independent ways of discovering networks allows investigation to proceed in the absence of detailed causal or dynamical information. This is part of the power of network representations.

**4. Cases in which Network Structure is Distinct from Dynamics on the Network**

 In cases involving IDEs it is important to distinguish the (i) constraints and corresponding physical structures just described, represented by a network from (ii) the processes or interactions that occur along the network and the rules or equations that describe these. It is these rules or equations (and the processes they describe) that, following a very common usage, I have called the *dynamics* governing these processes or interactions. It is important that often a dynamics can be specified independently of and can change independently of the network structure itself.

 As a simple illustration consider the well-known Watts and Strogatz paper (1998) on small world networks. In this paper, the authors observe that many phenomena such as the spread of disease may be affected by the network of contacts within the population. In particular, under plausible assumptions, a disease will spread more quickly and completely in a small world network[[12]](#footnote-12) than in a random network. What is less frequently observed, at least by philosophers, is that in reaching this conclusion Watts and Strogatz adopt certain dynamical assumptions about the disease spread --in particular, they adopt the SIR model[[13]](#footnote-13). Thus, Watts and Strogatz’s model is *not* one in which disease spread depends on the network structure alone, independently of dynamical considerations[[14]](#footnote-14).

 The SIR model has three variables: *S* representing the number of susceptible individuals-- those who become infected with some specified probability when they come in contact with an infected individual, *I* representing the number of infected individuals and *R* representing the number of individuals who either recover or are removed (e.g., die). The dynamics of the system is then represented by the following coupled differential equations.

(4.1) *dS/dt= - bIS/N*

*(*4.2*) dI/dt = bIS/N-gI*

(4.3*) dR/dt= gI*

here *b* is the average number of contacts per person per unit time multiplied by the probability of transmission when a susceptible individual is in contact with an infected person, *g* is the recovery rate and *N= S+ I +R.*

 These dynamical assumptions are obviously distinct from and independent of assumptions about network structure-- variables characterizing the network do not enter into the above equations (or vice-versa) and these assumptions can be combined with different assumptions about network structure, as Strogatz and Watts's analysis illustrates. Moreover, as described in more detail in the Appendix, not only can the combination of a particular set of dynamical assumptions about disease behavior with different assumptions about network structure lead to different outcomes, the reverse is also true-- different dynamical assumptions can lead to different results about spread on the same network.

 As another example, networks can be used to represent patterns of structural connectivity in brains but much of the behavior of the propagation of neural signals along these connections requires dynamical assumptions of some kind[[15]](#footnote-15). These are often very complex and will describe processes occurring at many different scales-- ingredients may include processes described by the Hodgkin Huxley model or more complex descendants such as the Wilson - Cowan equations. (See Schiff, 2012.) Again, these dynamical assumptions are distinct as a matter of mathematics from the network structure-- the dynamics is specified by a set of coupled differential equations while the network structure is specified by a graph. But more importantly, the two are also distinct in the sense that the same signaling processes described by the same dynamics can occur on different networks (with different overall outcomes resulting) and, when a network has the appropriate character, different signaling processes can occur along the same network, again with different outcomes..

 As yet another illustration (cf. Newman, 2010, 152ff) consider a simple diffusion model on a network. Such models are commonly used to describe the flow or spread of some "substance" (an idea, a disease etc.) along a network. Suppose that the amount of substance at node *i* is *Si*. Assume that this moves or flows along the edges of the network from vertex *i* to adjacent vertex *j* at rate *C (Sj-Si)* with *C* as the diffusion constant. Thus

(4.4) In a small interval of time the flow from *i* to *j* is *C ((Sj -Si)dt.*

(4.4) is another example of a dynamical assumption-- it says that the flow between *i* and *j* depends in a linear way on the difference between the substance amounts at these nodes. The rate of change of *Si* is then given by

(4.5) *dSi/dt= C ΣAij (Sj -Si)*

where *Aij* is the adjacency matrix for the network. This matrix encodes the connectivity of the network, and we see from (4.4) and (4.5) that this can vary as the above dynamics remains the same-- *Aij* is not a function of *C (Sj -Si*). Similarly, the expression (4.4) for the dynamics does not make reference to the adjacency matrix. Thus, we might use other assumptions about the dynamics in conjunction with the same adjacency matrix. This is thus another illustration of how the dynamics governing the processes occurring along a network is independent of the connectivity of the network. I emphasize again that what happens as the substance diffuses along the network is a function of *both* the network topology *Aij* and the assumed dynamics, rather than just one of these alone-- indeed, this is what (4.5) tells us.

 So far, I've been emphasizing the idea that network structure and dynamics are often independent. Strictly speaking, however, if we have in mind *causal* independence, at least in many cases this is only an approximation, although often a very good one. What *is* true is that networks are often (or at least assumed to be) governed by a different timescale than the timescale characterizing the dynamics. So, although the network may be affected by the processes occurring along the network and their dynamics, this typically happens relatively slowly in comparison with the faster timescale associated with the dynamics. For many purposes, this justifies treating the network as static or unchanging. For example, a system of roads may be represented by a network and processes that occur along this network (the traffic) and the dynamics governing it will certainly influence features of the road network over the long run-- if left unrepaired, a road may become impassible because of the causal influence of the cars that pass over it. But in the shorter run, if we are interested in traffic flow over a day or a week, it is entirely reasonable to treat the road network as fixed and uninfluenced by the traffic that passes over it (as well as by dynamical assumptions about that traffic). Similarly, the continued propagation of signals over a structural neural network may alter some features of the connectivity of that network-- e.g., it may lead to formation of new synapses -- but this is assumed to take place more slowly than the time course of propagation of any individual signal. Moreover, other features of the network such as neuronal growth or death, will typically change considerably more slowly if at all. Thus, on a short time scale one can think of this network as relatively fixed in structure, in comparison to the processes occurring on it.

 Finally, for purposes of contrast, consider the use of networks in which edges represent correlations of various sorts among nodes but but without any assumption that these correlations directly represent causal relations. (Functionalconnectivity which describes temporal correlations among patterns of brain activity is one illustration.) In such cases, in contrast to the examples considered above, the edges do not represent constraints on possible interactions. Thus, in such a case it will not make sense to specify an independent dynamics describing a process occurring along the network[[16]](#footnote-16). In this case there is nothing more to what the network represents than a pattern of correlation and this does not amount to an activity or process along the network for the dynamics to describe.

 My emphasis on the role of dynamical assumptions in many explanations that appeal to networks stands in contrast to a great deal of recent philosophical discussion, which tends to focus on the network structure alone. There certainly are cases in which this sort of focus is justified-- in addition to cases in which only network structure matters, there are also, as discussed below, cases in which any "reasonable" dynamics will yield largely the same outcomes, given a fixed network structure. But even in cases of this latter sort, dynamical assumptions are not completely irrelevant, since (i) one needs *some* dynamical assumptions if the problem of calculating outcomes is to be well-defined and (ii) there usually will be some alternative dynamical assumptions that will yield different outcomes than those that follow from more reasonable assumptions.[[17]](#footnote-17)

 Let me add that, in addition to the examples just discussed, the following simple consideration motivates the need for dynamical assumptions in many cases: often what we want to explain by using networks involves a change of some kind-- a disease spreads from a few infected individuals to an entire population, a change in effective connectivity leads to a mental disorder, an episode of congestion occurs on a road network. In such cases the networks themselves are, as we have seen, assumed to be relatively static and unchanging. A plausible general principle is that an unchanging *X* is by itself insufficient to explain a changing *Y*. So when a change occurs in the overall behavior of the system characterized by the network, there must be something else that is changing (the disease spreads, the traffic becomes more dense) and it is the behavior of this that is characterized by the dynamics.

**5. Use of Networks to Explain Why Certain Outcomes are Possible or Impossible**

 The examples discussed above differ in at least some respects with a example that is frequently discussed in the literature. This is Euler's analysis of the bridges of Konigsberg (Lange, 2018, Pincock, 2012, Kostic and Khalifa, 2021, Barrantes, 2023). Euler showed that a necessary and sufficient condition for a network to contain an Eulerian path (that is, a path in which each edge is traversed exactly once) is that the network be fully connected and that it contain either two or zero nodes of odd degree. The Konigsberg bridge configuration does not satisfy this condition and it follows that it is impossible to cross each bridge exactly once.

 Does this example make use of dynamical assumptions? It might be argued that it does since by focusing on the existence of an Eulerian path one is assuming what might be described as a walking-type dynamics in which people maintain continuous contact with roads and bridges, rather than one in which people are able to leap over waterways. However, one may also think of this dynamics as encoded in the network itself since if people were able to leap over waterways, a graph with a different connectivity structure or perhaps no graph at all would be appropriate. So the role of dynamics here seems at best minimal and largely implicit. Perhaps we can think of it as presupposed in the statement of the problem rather than a separate independent assumption. Alternatively, we might think of the dynamics as encoded in the network, in contrast to the examples discussed in Section 4. As still another possibility, one might think of this as an example in which dynamics plays no role at all in the explanation which instead rests entirely on an appeal to network structure.

 In any case the role (or non-role) of dynamics in this example seems connected to several other features that distinguish it from many of those considered previously. First, the explanandum in the Konigsberg example is naturally regarded as an *impossibility*[[18]](#footnote-18)*,* while this is not a natural way of thinking about explananda like the spread of disease along a network or the onset of a mental disorder. Second, and relatedly, the Konigsberg explanation does not involve a change or a process that occurs over time.

 Although Euler's conditions for the existence of an Eulerian path are non-obvious, the explanatory use of the network in this case seems to just draw on its ability to represent possibilities and impossibilities. (Recall that, according to the account above, networks represent paths of possible influence or transmission; hence they tell us, among other things, whether influence or transmission is impossible.) In this respect, the Konigsberg case does not seem very different from one in which (i) cars can only travel along roads and (ii) there is no road leading from location A to location B-- so no path in the network representation. Then, given (i) one might appeal to the absence of such a path to explain why it is impossible to drive a car from A to B and to presence of such a path to explain why such a trip is possible[[19]](#footnote-19). Here this impossibility is built into assumptions (i) and (ii) in a transparent way. In Euler's analysis the impossibility result requires more inferential steps, but the underlying logic seems the same[[20]](#footnote-20).

 The moral that I draw from this is that the Konigsberg bridge case and similar examples are importantly different from those in which dynamics plays a more salient role. Because of these differences, the Konigsberg case seems more like one in which the explanation appeals only to network structure, with no or a minimal role for dynamical considerations. In this respect, it is an "edge" case which may not be representative of many other cases in which networks explanations. I return to this example in Section 10.

**6. Some Consequences of the Independence of Dynamics and Network Structure along with some Nuances and Complications**

 The fact that in generic IDE cases the overall outcome -- e.g., the actual pattern of disease spread, of traffic congestion, of neural signaling etc.-- will reflect *both* the structure of the network *and* the dynamical assumptions about the processes operating along the network has a number of important implications. Most obviously it implies that it is a mistake to hold that in such cases these outcomes are explained by the connectivity of the network alone. Moreover, as illustrated by the examples above, the assumed dynamics will typically embody *causal* assumptions[[21]](#footnote-21) (in addition to the causal assumptions, having to do with what can influence what, that are encoded in the network itself). Thus, even if one holds, contrary to what I have suggested, that the specification of a network itself does not require any causal information, the explanation for the overall outcome for what happens on the network will commonly require causal commitments in the form of a dynamics. In addition, as we shall see (Section 10), the fact that, in the general case, both network structure and dynamics are needed for explanation of outcomes is crucial to understanding how such explanations can have directional features-- that is, why we think that network structures can help to explain outcomes on the network rather than vice -versa.

. Although in IDEs outcomes will commonly reflect both network structure and dynamics, it is also true, as intimated above, the relative importance of the dynamics will vary from case to case, depending on empirical details and the problem addressed. In some cases, this can justify a focus mainly on network structure. First, once one specifies or assumes a dynamics, one can ask, given that dynamics, what difference one network structure rather than another makes to an outcome. For example, given particular dynamics of disease spread like the SIR model, one can ask what difference it makes if the assumed network is small-world rather than a random graph. (This is what Watts and Strogatz, 1998 do.) Many of the examples of network or topological explanations discussed in the scientific and philosophical literature are plausibly viewed as cases in which a particular dynamics or a dynamics belonging to some general class is assumed, with the question addressed then being what difference it makes to some set of outcomes whether the network structure takes one or another general form. Note, though, that this does *not* mean that the dynamical assumptions are irrelevant--it still can be (and often is) the case that different dynamical assumptions will lead to a different outcome.

 Second, although some assumptions about dynamics are required, it may be that, given a particular network structure, the details of the dynamics (within some large range) may matter little or not at all to some set of outcomes. That is, it may be that, for certain networks, any one of a large number of assumptions about the dynamics (perhaps any "reasonable" dynamics) will lead to the same or largely the same outcomes among those of interest-- or at least researchers may claim or hope that this is the case. This consideration is particularly important in cases in which there is little information about the true dynamics. For example, Bassett and colleagues (Gu et al., 2015) explored, using network control theory, how the structure of brain networks affects their controllability -- roughly the extent to which an intervention in one neural structure (represented by an injection of "energy") can affect the state of other brain areas. Controllability is a broadly causal notion and reaching conclusions about controllability thus requires some dynamical assumptions. Gu et al. explicitly recognize this, remarking that to answer the questions that interest them we must move beyond "descriptive statistics" and that "we must have a framework that incorporates not just brain network structure but also models neural dynamics". In their models Gu et al. assume that neural signal transmission is governed by a very simple linear dynamics-- they recognize that the reality is far more complicated, but they think that the linear dynamics is a reasonable approximation from which some useful conclusions can be extracted. In other words, they assume that a linear dynamics is enough like the unknown real dynamics (at least in some relevant respects) that one may adopt it as a generic assumption and then use it to explore the implications of different network structures for controllability. Again, I emphasize that it is not that dynamics don't matter at all-- Gu et al. say explicitly that dynamical assumptions are required; rather it is assumed that the conclusions are relatively insensitive to departures from linearity of their dynamical assumptions. .

 Of course, it may also happen that some outcomes of interest may not depend on certain dynamical assumptions at all, although others do. As an illustration, it can be shown that given that the degree distribution in a network follows a power law, disease spread along the network will eventually result in an "epidemic" (disease spread through an entire compartment once one individual in the compartment is infected) regardless of the probability *p* of transmission. (Newman, 2010). Thus, this particular feature of the dynamics does not matter for whether an epidemic eventually occurs. However, it does affect how quickly spread occurs, illustrating the point that dynamics can be relevant to some outcomes even if it is not relevant to others.

 When, given some set of possible network structures, various details of the dynamics don't matter-- again for some specified set of outcomes -- and different outcomes follow for different network structures, there is an obvious sense in which much of the explanatory weight is being carried by facts about the network structure: it is the variations in network structure that explain the variations in outcomes (again given the generic assumptions about the dynamics)[[22]](#footnote-22). I suspect that it is possibilities like this that encourage some to construe explanations of disease spread and similar phenomena as explanations that appeal only to network structure and nothing else.

**7.** **A Schema for How Networks Figure in Explanations**

 With this as background I turn to a more explicit schema (**W** below) regarding how networks figure in explanations of the sort described above. This is intended to cover the generic case in which an independent dynamics, along with network structure plays a role in the explanation. As noted above, not all explanations in which networks figure directly fit this schema-- for example, additional assumptions may be required when networks are used to predict or explain what happens when a node is removed. Later (Section 10) I will consider the possible extension of **W** to cases like the Konigsberg bridges in which dynamics plays a minimal role.

 **W** has the following components:

**W1** A claim that some system *S* has a network structure *N* with various features where these may be either relatively local or more global (e.g., the network has a node with such and such a centrality measure, the network is small world, scale free, has a rich club structure etc.) That is, it is claimed that *N* represents, with some suitable degree of accuracy, relevant aspects of the connectivity in *S*.

**W2.** Claims about the dynamics *D* that operate on the network *N* in *S*. In addition to general rules governing *S*'s dynamics this may also include, when it matters, facts about the initial or starting conditions of the system[[23]](#footnote-23). That is, if it is the case that the system will only end up in some state of interest if it begins in a particular initial state, information about this state should be represented.

**W3**. The existence of a *dependency relation* between the facts in **W1**) and **W2)** and an explanandum *E* showing how and explanandum *E* depends on the facts in **W1 -W2**.

**W4**. *E* (the outcome to be explained). Usually this is some feature of the overall behavior of system *S*, rather, than, e.g., the value taken by some particular node-- the overall spread of a disease or the presence of a mental disorder. See below for more discussion.

**W1- W3** constitute an explanans for *E* in **W4**.

 By a dependency relation (**W3**) between **W1**, **W2** and *E*, I mean that there are both possible changes in *N* such that for the specified dynamics *D*, the outcome *E* will be different and possible changes in the dynamics *D* such that given the network structure *N*, *E* will be different. Often it will be appropriate to think of the possible changes in *N* as what would be produced by intervention-like processes that alter the structure of *N* or by some suitable generalization of these[[24]](#footnote-24). In these cases, the dependency between *E* and *N* can be understood in terms of counterfactuals that have the intervention-like properties described in footnote 22. Turning now to the dynamics, in some cases it may be possible to empirically alter these and see what difference this makes to outcomes (e.g., one might do something that changes the dynamics from linear to non-linear). In other cases, ascertaining the results for *E* of a change in the dynamics will require mathematical calculation. In my view, it is not clear how to apply the usual "surgical" notion of an intervention to manipulations that change the dynamics of a system. Still a more general notion of an unconfounded manipulation is often applicable, with "unconfounded" meaning that the change in the dynamics leaves the network structure intact. For example, one might alter the dynamics of a traffic system by changing the speed limit or by requiring that cars be a certain distance apart, but this will not change the network structure of the roads. If such a change alters patterns of traffic flow (the explanandum of interest) then this indicates the presence of counterfactual dependence between the flow and the dynamics.

 In a number of cases the dependency relation between *E* and the network structure *N* will be a contingent relation that holds as a matter of empirical fact, rather than a relationship that corresponds to a mathematical truth. For example, the relationships that hold between structural connectivity networks in brains and mental disorders is empirical in the sense that no one supposes, at least at present, that one can deduce from the network structure, even given information about the brain dynamics, whether a disorder will be present just by mathematical reasoning. This is reflected in the general recognition that claims about the relationship between connectivity and disorders need to be investigated empirically. However, in other cases in which we assume a fixed dynamics, the dependency relationship between network structure and *E* may hold as a matter of mathematics and for this reason the associated explanation (or at least this part of it) will not count as causal, assuming that causal relations hold contingently. The Konigsberg bridge case (discussed below) is a candidate for such a description. Note, however, that in such cases it may still be possible to intervene to alter the network structure (e.g., by adding or removing a bridge) with a corresponding change in *E*, so that we can still view the explanation as explaining in virtue of citing an intervention/ supporting (but non-causal) dependency relation. (cf. Woodward, 2018.)

 Schema (**W**) seems to fit a number of the cases of explanations involving networks that have been discussed above. Suppose the more rapid spread of a disease among one set of people rather than another is attributed to the fact that, the contact relations in the first group have a small world structure while the relations in the second group do not, and this difference in network structure leads to a difference in the rate of spread, given a plausible disease dynamics. This claim will be correct to the extent that intervening to change the network structure of contacts (something that might be achieved by, for example, quarantining) will alter the spread of the disease. As another illustration, suppose that a mental disorder is attributed to the fact that the subject's structural neural network is different in some way from healthy controls (too many or too few long-range connections, abnormal patterns of node clustering etc.). This suggests that if we could somehow intervene to change that structural connectivity pattern to something more like that in normal subjects this would change whether the subject has a mental disorder[[25]](#footnote-25). Similarly consider a subject thought to have abnormal brain dynamics (perhaps because of epilepsy) with various behavioral consequences-- the explananda *E* of interest. If the abnormal dynamics are part of the explanation of *E,* this implies that doing something that alters the dynamics (e.g. by administering a drug) can alter *E*.

 What about cases in which the explananda depend on network structure alone, with the dynamics playing no role at all? As suggested earlier, these can be treated as a special case of **W,** in which there is no need for the information in W2 and it is dropped from the explanation, since there is no dependence between this information and the outcome. For example, in the case mentioned in Section 6 in which, given a certain network structure, spread of disease throughout a compartment is independent of the probability of transmission *p*, the explanation proceeds (we are supposing) just by appeal to this network structure, so there is no need for a W2 type premise. In the case of the Konigsberg bridges, it is not clear that the explanation is entirely independent of dynamical considerations (see Sections 5 and 10) but to the extent this is the case, this explanation will also proceed just via W1 and W3, without any contribution from W2.

 Schema **W** is similar in some respects to the schema T advanced by Kostic (e.g., 2022, 2023) and Kostic and Khalifa (2021) to characterize topological explanations. In particular, **W** agrees with Kostic and Khalifa in requiring that there be a pattern of counterfactual dependence between the network structure and the explanandum. However, **W** departs from schema T in several respects. First, it assigns a role to dynamics in many successful network explanations-- this feature is absent in T. Second, schema **W** does *not* take counterfactual dependence in general to be sufficient for explanation. Instead, what matters is that there be a specific kind of counterfactual dependence, described by **W3**, between W**1**, **W2** and the explanandum E. This variety of counterfactual dependence is captured either by *interventionist counterfactuals* (e.g., if one were to intervene to change the network structure, *E* would change) or, as explained above, by some generalization of these. In both cases, this excludes so-called backtracking counterfactuals. Kostic and Khalifa (2021) also attempt to exclude backtracking counterfactuals (or at least some of them) but do so by means of a "pragmatic" relevance condition which they label T4 and which requires that the explanans be "an answer to the question" of why the explanandum obtains. While this condition is unexceptionable, it seems to me desirable to have more guidance into when and why it obtains. The use of interventionist counterfactuals and generalizations of these in **W** attempts to do this.

 Consider in this light Kostic and Khalifa’s (2021, 2022) claims that one can appeal to functional connectivity in the brain to explain, where (recall) functional connectivity has to do with patterns of correlation between different brain regions:

 ... if the brain dynamics counterfactually depends on variability in functional connectivity, while the direct anatomical connections remain fixed, then such a model satisfies [the conditions T for topological explanation] . This shows that functional connectivity models can be explanatory and are not always merely evidential.

By contrast, a common view among researchers is that patterns of functional connectivity are *effects* of brain dynamics in conjunction with network structure. Abnormal patterns of functional connectivity can be evidence for abnormal dynamics and thus can be used to *classify* patients as suffering from some brain disorder (or mental illness) but abnormal functional connectivity is not a cause of these.

 This view is reflected in the following remarks of Friston (2011):

 Functional connectivity has a distinct role from effective connectivity [Recall this has to do with causal relations between different brain regions.] Functional connectivity is being used as a (second-order) data feature to classify subjects or predict some experimental factor. It is important to realize, however, that the resulting classification does not test any hypothesis about differences in brain coupling. The reason for this is subtle but simple: in classification problems, one is trying to establish a mapping from imaging data (physiological consequences) to a diagnostic class (categorical cause). This means that the model comparison pertains to a mapping from consequences to causes and not a generative model mapping from causes to consequences (through hidden neurophysiological states). (p 15)

 The fact that differences in functional connectivity can be evidence for brain coupling (i.e., brain dynamics) implies that there is a kind of counterfactual dependence between whether differences in the dynamics and differences in functional connectivity, just as Kostic and Khalifa claim. However, if Friston and others are right, this is not the right kind of counterfactual dependence to establish that functional connectivity can explain brain dynamics. Instead, functional connectivity (*E*)is an effect of (and explained by) the combination of *C1* neural dynamics and *C2*structural connectivity. Claiming that *E* can be used to explain *C1* amounts to attempting to explain a cause by appeal to one of its effects. Thus, the condition T4 misclassifies this case-- we need a condition that is more discriminating. By contrast, intervening on or manipulating *E* (e.g. by altering *C2*) will not change *C1*, assuming as, we have been, an independent dynamics. Thus, schema **W** classifies the counterfactual dependence of *C1* on *E* as not indicative of an explanatory connection.

**8.** **Topological Properties?**

 On the picture I have presented so far many networks that figure in explanations represent patterns of connectivity in physical structures (like roads etc.) that constrain possible interactions or patterns of influence. In such cases, a network representation of course will be an abstract or coarse-grained description of such structures since the network representation just represents facts about connectivity and does not represent other more specific material details of the system. For example, a network representation of a system of roads will not represent the material out of which the roads are constructed or, at least usually, their state of repair. Similarly, a network representing structural connectivity within a brain will not provide detailed information about the material composition of neurons. Characterizations of more global features of a network -- e.g., whether it is small world or modular -- as well as as such features as measures of node centrality are even more abstract characterizations of features of connectivity patterns.

 I stress this because some writers describe networks as encoding "topological" or "mathematical" properties. This is unexceptionable if it simply means that that topological, mathematical or graph-theoretical concepts are used to describe patterns of connectivity in physical systems. Sometimes, however, something stronger seems to be suggested -- that the properties and relations represented in a network should be thought of as "mathematical" or "topological" in a sense that contrasts with "physical" or "empirical", or even "causal", so that when one appeals to network structure to explain some some explanandum, one is appealing to non- physical/ non-empirical/non-causal properties to explain this explanandum. This in turn carries the apparent implication that such explanations differ from those in which the explanans describes ordinary physical facts and relationships. It is also sometimes taken to imply that explanations appealing to topology or network structure are, for this reason alone, non-causal (since causal explanations appeal to ordinary physical properties). For example, Darrason (2018) writes:

 To be perfectly clear, let’s specify that material properties and topological properties are not merely distinct but completely different kinds of properties. Material properties are directly related to the physical and concrete properties of an object.

In a somewhat similar vein, the characterization of topological explanations in Kostic's and K's schema **T** requires that the explanans makes reference to "topological" properties while the explanandum makes reference to "empirical" properties, with the apparent suggestion that these are two different kinds of properties that should be contrasted with each other.

 I think views of this sort introduce unnecessary complexities and make network explanations seem more mysterious than they actually are. In my view, the use of an edge in a graph to represent, e.g., the fact that two brain reasons are structurally connected or that two people are in close spatial proximity is no different in principle from the use of, say, a non-negative real number to describe an object's mass or the use of a continuous function to describe an object's trajectory (that is, the spatial path it follows). These characterizations use mathematics to describe, and continuity is certainly a "topological" property, but this does not mean that the mass of an object or its trajectory is "mathematical" in a sense that contrasts with "physical" or "empirical". Like the network representation, these characterizations also leave out detail-- for example, they don't tell us what the object is made of or just what the object's trajectory is (other than the fact that is is continuous). However, this absence of detail does not mean that they involve the ascription of non-physical properties. Indeed, if the use of abstract, mathematical characterizations from graph theory (such as connectivity) is enough to establish that we are dealing with "non-physical" etc. properties, it would follow that all of mathematical physics is in the business of ascribing non-physical, non-causal properties. This cannot be right.

 None of this is to deny that topological or network explanations have distinctive features--- I've been discussing a number of these-- but what is distinctive about such explanations is not that they relate non- physical or non-material properties to physical explananda[[26]](#footnote-26).

**9. Are Network Explanations "Autonomous"?**

 According to the view I have been defending, in many cases both network structure and assumptions about dynamics of processes along the network are relevant to explaining outcomes-- they work together. Indeed, in several important respects this way of putting matters *understates* the extent to which network structure relies on other sorts of information. As we have noted, when we ask about the network structure that is relevant to the behavior of some system, while we don't need to have detailed information about the interactions and processes occurring along the network or the dynamics governing these, we do need to have enough information about these to identify the relevant network. This typically includes broadly causal information in the sense that it has to do with possibilities of causal influence involving what is transmitted along the network. A network of roads is the relevant network for understanding traffic flow because or to the extent that cars have causal properties that require them to move only along roads-- if cars could move arbitrarily off roads or could jump over buildings, the relevant network would be different (or perhaps there would not be any interesting network constraints). Similarly, it is part of the nature of certain neural signals that they require specific anatomical connections for their propagation. This in turn leads us to a network representation of those connections rather than others. Signaling or influence that works differently-- e. g., by general diffusion -- might require a different or no network representation[[27]](#footnote-27).

 Consider in this light claim like the following:

*autonomists* about topological explanation, such as myself, maintain that topological explanations are a new and distinct kind of explanation. (Kostic, 2021)

 ..there are cases in which topological properties explain independently of any causal or mechanistic considerations. (Kostic, 2022)

 In non-causal explanations [where this is understood to include topological explanations-- JW], most broadly speaking, some non-causal facts (such as mathematical, modal or metaphysical) are used to explain some physical facts (Kostic, 2023)

 I argued above against the claim that the connectivity properties of a physical system, represented in a network, should be understood as 'mathematical" in a sense that contrasts with ordinary physical properties. Putting this consideration aside, if talk of the "autonomy" of topological explanations means that one can fix or assume a dynamics for a system and then, under this assumption, consider how the behavior of the system changes (and is explained by) changes in its network connectivity relations, then as suggested above, this is a perfectly legitimate undertaking-- one is tracking how, given that dynamics, network structure by itself makes a difference. I have also agreed that in some cases, given a network structure, some behavior will follow for many different or perhaps even *any* reasonable dynamics, with different behavior following from other network structures so that in that sense, it is variations in network structure that play the most important role in explaining variations in system behavior.

 On the other hand, I have also emphasized that often, when network structure has an explanatory role, some non-network assumptions about dynamics, even if rather generic and perhaps other assumptions as well are also required. Indeed, in many cases of interest it seems difficult to make sense of the idea that the network by itself, independently of any conception of processes or interactions occurring along the network, can explain system behavior. A network characterizing the structural connections in a brain does not by itself explain why the brain behaves as it does-- we also need to consider the role of signals propagating along these connections[[28]](#footnote-28). Moreover, to the extent that different dynamical assumptions can lead to different behavior on the same network, this is another respect in which the network by itself does not explain the behavior. And some behaviors of interest, such as rates of disease spread (as opposed to how much spread will occur in the limit), will inevitably depend on dynamics as well as network structure. In addition, as argued above, the very construction of a network typically requires assumptions about processes or interactions or flows along the network -- assumptions about neural signals tell us that certain connections such as neurons but not blood vessels are the relevant networks to consider. Much of this information is causal information. To the extent that claims about the autonomy of topological explanations or their non-causal character deny these observations they seem overstated.

 To avoid misunderstanding, let me add that there are other notions of autonomy which are legitimately applicable to network explanations. It is obviously true that ascriptions of network structure abstract away from the material details of the systems represented. The same network (or networks with the same global properties but differing in connectivity details) can be used to represent connectivity relations in systems that differ radically in more concrete detail-- a network of roads, a contact network associated with disease transmission, structural connectivity in a brain and so on. The use of a common network structure applicable across these different systems amounts to a commitment to the claim (or working hypothesis) that for the purposes at hand (the explananda we are trying to explain) these differences in material detail don't matter or matter very little-- it is the overall patterns of connectivity among components (in conjunction with more or less precise dynamical systems) that drives what happens. This involves a kind of autonomy (of behavior from material details and so on) that is familiar from philosophical discussions of multiple realizability. However, one can acknowledge this point without concluding that explanations that appeal to networks make no use of causal information or involve appeal to non-physical properties.

**10. Directionality and Network Explanations**

Issues about the directionality of network explanations have elicited considerable discussion recently (e.g., Craver and Povich, 2019, Lange, 2018, Kostic and Khalifa, 2022 ). We think that the structure of a contact network *N* helps to explain facts about disease spread *E.* However, although *E*, in conjunction with dynamical assumptions, may allow the deduction of facts about *N*, and although, again given the dynamics, there may be a sense in which *N* counterfactually depends on *E*, *E* does not explain *N*. Information about patterns of traffic flow can, on the assumption that traffic can only move on roads, tell us about the structure of a road network but this does not explain why the road network has the structure it does. Similarly, as noted previously, there is a sense in which brain dynamics “depends" on functional connectivity, but the dynamics is not explained by functional connectivity-- rather the explanation runs from dynamics to functional connectivity.

 Note that if one thinks that some or all network explanations are non-causal, one cannot appeal to causal considerations to explain why in such cases the dependency is explanatory when it runs in one direction but not in the other. Thus, if we assume (as I shall) that all explanations have a direction and that at least some network explanations are non-causal, one needs to find another account of their directionality. Alternatively, one might conclude that to the extent that we lack a clear account of the directionality of alleged network explanations, this suggests that they are non-explanatory. Of course, to the extent that network explanations are causal, one might take their directionality to be due to the directionality of causation. However, in my opinion, the directionality of causation is also not well understood. It would be desirable to have a more general framework that provides insight into the sources of directionality in both causal and non-causal explanations.

 My proposal about this draws on (my versions) of work by Dan Hausman (1998) and more recent developments in machine learning (Janzing et al. 2012) some of which is described in Woodward (2021a). These ideas were originally developed for thinking about directionality in non-network causal explanations involving structural equations. I claim they can be extended to capture the directional features of many network explanations, including both those that are causal, those that may be non-causal and any that contain a mix of causal and non-causal elements. I emphasize this to underscore the point that the suggestions that follow do not appeal to considerations developed specifically to account for directionality in network explanations (and that apply only to these) but rather to a set of ideas about directionality that have been successfully applied in other contexts, both philosophical and non-philosophical.

 The basic idea appeals to interventionist/manipulationist considerations and, more generally, to the role played by an independence condition: Suppose one has two variables, *X* and *Y* and one is trying to determine whether the explanatory direction runs from *X* to *Y* or from *Y* to *X*. Assume there is also a third variable *Z* which is not correlated with *X*. (This is the independence condition.) It is known that *X* and *Y* are correlated but we do not have direct information about whether interventions on *X* change *Y* or conversely. (If we do know that interventions on *X* change *Y* and interventions on *Y* do not change *X*, then of course if follows immediately that *X* causes *Y*.) However, we observe that interventions on Z change *Y* but interventions on *X* do not change Z and interventions on Z do not change *X*. (Also an independence condition on *X* and *Z*.) Alternatively we infer these interventionist claims from from the correlations between *X* and *Y*, *Z* and *Y* and the absence of a correlation between *X* and *Z*. Then, absent special circumstances, the correct direction of explanation is that *X* explains *Y* rather than *Y* explaining *X*. To motivate this claim, suppose instead that *Y* explains *X*. Then, since it is assumed that interventions on *Z* change *Y*, one would expect (under the above supposition and barring some special circumstances[[29]](#footnote-29)) that such interventions on Z would also be associated with changes in *X* and, *ex hypothesi*, they are not[[30]](#footnote-30). On the other hand, if the correct direction is from *X* to *Y,* it is completely understandable why an intervention on Z which changes *Y* does not change *X*. The key feature that is exploited here is the independence of *X* and *Z* and the fact that *Y* changes under interventions on *Z*-- this gives us a way of changing *Y* via Z that does not require changing *X* and when *Y* changes under this intervention, we conclude that the correct explanatory direction runs from *X* to *Y*.

 As an illustration involving an ordinary causal explanation, consider the question of whether the correct direction of explanation is from the length *l* of a simple pendulum to its period *T* or from *T* to *l*. We find that if we intervene to change *l* (e.g., by substituting a longer string or stretching the original string) *T* changes in accord with the equation

(P) *T= 2Π√l/g*

Here *g,* the gravitational acceleration, plays the role of Z above -- it is unaffected by this intervention, assuming that we remain in the same gravitational field. We can also intervene to change *g* by moving the pendulum to a stronger or weaker gravitational field. This will also influence the period in accord with equation (P) but will not change *l.* If the correct explanatory direction is from *T* to *l*, one would expect that an intervention that makes use of a change in *g* to change *T* would also change *l* and this is not found. One the other hand, if the correct direction is from *l* to *T* and from *g* to *T* with *l* and *g* being causally independent, the observed pattern is just what one would expect.

 My proposal is that a similar analysis holds for directionality in many network explanations even when they are (in at least some respects) non-causal. Here again the fact that there are two independent factors-- the network structure *N* and the dynamics *D*  -- that influence the explanandum of interest *E* is crucial. Because of this independence, we can influence *E* by changing the dynamics without changing the network structure and we can change the network structure in a way that also influences *E* but without changing the dynamics. This information tells us that the correct direction of explanation runs from *N* and *D* to *E* rather than from *E* to *N* or from *E* to *D*. For example, as we have seen, for a fixed dynamics, changes in the structure of a contact network often influence the spread of disease along the network but it is also true that for a fixed network, the variations in the dynamics along the network can lead to variation in disease spread. Moreover, the dynamics and network structure can change independently. This tells us that the correct direction of explanation runs from the network structure and the dynamics to the spread of the disease. Notice that this argument does not depend on the assumption that the explanation is causal, although in some cases or respects it may be. It is the independence of *N* and *D* that drives the conclusion about direction.

 I claimed earlier that it can be true that an intervention/manipulation on network structure *N* changes an explanandum *E* while not true that an intervention on *E* changes *N*, even if we can derive information about *N* from *E* (and perhaps other assumptions). That is, an interventionist counterfactual of the form:

 If *N* were to change as a result on an intervention, *E* would change

 can be true, even though counterfactuals of the form:

 If *E* were to change as a result on an intervention, *N* would change

are false and even though there are interpretations of some counterfactuals of this latter sort (e.g., interpretations that allow backtracking) according to which they are true. My discussion immediately above should make it clearer how this is possible. The key is the existence of an alternative route (involving the dynamics *D*), in addition to and independent of *N* that allows for manipulation of *E*. When *N* does not change under a manipulation of *E* by means of *D*, we can conclude that *E* does not cause *D*.

 Return to the fact (*F*) that the network structure of Konigsberg bridges does not have an Eulerian path and that this explains the fact that the bridges cannot be traversed exactly once (*OC*) rather than *OC* explaining *F*. Even if one thinks of the relationship between *F* and *OC* as non-causal, it is still the case that one might change the network (e.g., by building additional bridges) so that an Eulerian path is present. This would change *OC*, so that there is counterfactual dependence of *OC* on *F*. Note also that whatever one thinks about the status of the walking dynamics *D* assumed in the example (that people traverse paths by ordinary walking which requires something like continuous spatial contact with the path), changing *F* will not change this dynamics. In fact this is assumed when we claim that the right sort of change in *F* will change *OC--* we are assuming that *D* is unaffected. So at least to this extent, we are assuming a dynamics, however implicit, that is independent of *F.* This independence assumption can be verified by noting that changing *F* does not change *D*, again understood as above. And although contemplating changes in *D* may seem fanciful and contrary to our implicit understanding of the case, it is nonetheless true that if, say, people were able to leap over canals without crossing bridges, this would change OC without changing the original bridge configuration. We thus have the independence pattern that is present in the other examples in which the explanatory direction is well defined: *F* and *D* are independent, *OC* depends on *F* and *OC* depends on *D*. This is my proposal for why we think in this case the explanatory direction runs from *F* and *D* to *OC*.

**11. Conclusion: When are Network Explanations Likely to be Most Fruitful?**

 I conclude by summarizing some of the main strands of my discussion and by applying these to the general question of when networks are likely to be most fruitful for explanatory purposes. Interestingly, this important question has hitherto received little attention in the philosophical literature.

1. First, as we have noted, the explanatory use of networks works well in contexts in which we can separate the network structure from assumptions about the dynamics governing whatever processes occur along the networks. Given such a separation, we may ask, among other questions, given some assumed dynamics, how differences in network structure make a difference to outcomes. Here a focus on network structure is reasonable, not because the dynamics is irrelevant but because variation in network structure explains variation in outcomes, for a given dynamics. More generally (and ambitiously) we may also ask whether, given a network structure, the same or different outcomes will follow from different dynamical assumptions. As described in Section 7, this naturally fits with an account of network explanation, captured by schema **W**, according to which this involves tracing patterns of counterfactual dependence between (a) network structure, (b) dynamics and (c) outcomes-- seeing what difference variations in (a) and (b) make for (c).

2. There certainly are cases (other than those described in 1) in which network structure alone, independently of dynamical assumptions or in which dynamical assumptions are minimal or implicit, can explain outcomes. These cases can be unified with those in which dynamical assumptions do play a role within the framework of schema **W**.

3. When networks are used for explanatory purposes, network structure often represents constraints on possible interactions. This interpretation amounts to the assumption that, for the purposes at hand, we can blackbox or avoid committing to claims about the details of those interactions-- that is, what matters, roughly, is just the generic fact that one node can, in some way or other, influence another node (or transmit something to it that influences its state etc.) where this is relevant to what we are trying to explain but without further details needing to be included. Network explanations will be most fruitful when this empirical assumption (about details of what happens along the edge not mattering) is correct.

4. Related to 3) and returning to a theme mentioned only in passing above, given such a networks/dynamics separation, it may happen that we have ways of discovering network structure that require only minimal assumptions about dynamics. For example, one can identify structural connections in the brain through techniques like diffusion tensor imaging without assuming much about the dynamics of neural signal processing other than that such processes, however they may work in detail, propagate along these connections. Identifying a contact network for purposes of explaining disease spread can be accomplished without knowing much about the details of how the disease spreads, as long as it is transmitted by contact. This is important because in such cases identifying relevant network structure may be much easier than achieving a detailed understanding of dynamics and causal details of transmission, thus allowing for at least partial understanding in the absence of the latter information. Systems for which such independent identification of network structure is possible are likely to be particularly suitable for network analysis.

5. Just as the edges in a network can be thought of as encoding the assumption that for certain explananda it is generic facts about what can influence what that matter, it is also the case that network explanations will be most successful when the details of what goes on *within* individual nodes does not matter much for whatever it is we are trying to explain. Thus, a network representation of structural connections between brain areas *A* and *B* will not tell us much about what goes on within *A* and *B* -- these will just be represented as nodes with labels. Again, this reflects the empirical assumption that for certain explananda such within node information is not required.

6. Network analyses are also most likely to be fruitful when the outcomes explained depend (given a dynamics) on relatively large scale or global features of the network-- whether it is small world, random, whether edges are distributed in accord with a power law, whether or not nodes are highly interconnected, general relations connecting network structure to controllability and so on. Even when a particular node is of interest-- e.g., because it is central according to some measure, its centrality will reflect more global features of the network-- not just whether it is extensively connected to other nodes, but whether those other nodes are themselves extensively connected. Again, the use of networks amounts to a kind of bet that, as an empirical matter, more local details of the connectivity will not matter much to the outcome of interest.

7. To the extent that the abstraction from detail described in 4) - 6) is empirically supported, this allows networks with similar properties to be used to describe a wide variety of systems which differ in their material properties or causal details-- ecological networks with the same structure can represent relations of predation or of parasitism, small world networks can represent relations in the brain or social relations and so on.

**References**

Barabasi, A.-L. (2016) *Network Science.* Cambridge: Cambridge University Press.

Bechtel, W. (2018). "The Importance of Constraints and Control in Biological Mechanisms: Insights from Cancer Research". *Philosophy of Science*; 573–593.

Craver, C. F. (2016). The explanatory power of network models. *Philosophy of Science, 83*(5), 698–709. https://doi.org/10.1086/

Craver, C. F., & Povich, M. (2019). The directionality of distinctively mathematical explanations. *Studies in History and Philosophy of Science Part A, 63*, 31–38. https://doi.org/10.1016/j.shpsa.2017.04.005

Darrason, M. 2018 "Mechanistic and topological explanations in medicine: the case of medical genetics and network medicine" . *Synthese* 195, 147–173. (doi:10.1007/s11229-015-0983-y)

Friston, K. (2011) "Functional and Effective Connectivity: A Review" *Brain*

Hausman, D. (1998) *Causal Asymmetries*. Cambridge: Cambridge University Press.

Gu, S., Pasqualetti, F., Clesaks, M., Telesford, Q., Yu, A., Kahn, A., Medaglia, J., Vettel, J., Miller, M., Grafton, S. and Bassett, D. (2015)" Controllability of Structural Brain Networks" *Nature communications 6.*

Huneman, Philippe. 2010. “Topological Explanations and Robustness in Biological Sciences.” *Synthese* 177 (2): 213–45. https://doi.org/10.1007/s11229-010-9842-z.

Janzing, D., Mooji. J., Zhang, K., Lemeire, J., Zscheischler, J., Daniusis, P. Steudel, B. and Scholkopf, B. (2012) “Information-geometric approach to inferring causal directions” *Artificial Intelligence* 182-183: 1-31.

Jones, N. (2014). "Bowtie structures, pathway diagrams, and topological explana- tion". *Erkenntnis* , 79(5), 1135-1155.

Kostić, D. (2018). Mechanistic and topological explanations: an introduction. *Synthese*, 195 (1):1–10.

Kostić, D. (2020) "General Theory of Topological Explanations and Explanatory Asymmetry" *Philosophical Transactions of the the Royal Society B: Biological Sciences* 375: 1-8.

Kostić, D. (2022) "Topological Explanations: An Opionated Appraisal" In I. Lawler, E. Shech & K. Khalifa (eds.), *Scientific Understanding and Representation: Modeling in the Physical Sciences* 96-115.

Kostic, D. (2023) " On the Role of Erotetic Constraints in Non-causal Explanations" Forthcoming in *Philosophy of Science*.

Kostić, D. and Khalifa, K. (2021). “The Directionality of Topological Explanations.” *Synthese*, November. https://doi.org/10.1007/s11229-021-03414-y.

Kostić, D. and Khalifa, K (2023) "Decoupling Topological Explanations from Mechanisms" *Philosophy of Science* 90: 245–268.

Lange, M. (2018). *Because Without Cause: Non-Causal Explanations in Science and Mathematics.* Oxford: Oxford University Press.

Newman, M. (2010). *Networks: An Introduction*. Oxford: Oxford University Press.

Pincock, C. (2012) Mathematics and Scientific Representation. Oxford University Press.

Rathkopf, C. (2015). "Network representation and complex systems". *Synthese* 195: 55–78 DOI 10.1007/s11229-015-0726-0

Ross, L. (2021) "Distinguishing Topological and Causal Explanation**.**
*Synthese* 98:  9803–9820.

Ross, L (2023) "The Explanatory Nature of Constraints: Law-Based, Mathematical, and Causal". *Synthese* 202:56. https://doi.org/10.1007/s11229-023-04281-5.

Schiff, S. (2012) *Neural Control Engineering*. Cambridge: MIT Press.

Silberstein, M. and Chemero, A. (2013) "Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences". *Philosophy of Science*, 80(5):958–970.

Watts, D.J. and Strogatz, S.H. (1998) "Collective dynamics of ‘small world’ networks". *Nature* 393: 440-442.

Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation.* New York: Oxford University Press.

Woodward, J. (2018) “Some Varieties of Non-Causal Explanation”. In *Explanation Beyond Causation,* ed. Juha Saatsi and Alex Reutlinger. Oxford University Press, pp 117-137.

Woodward, J. (2021a) "Flagpoles Anyone? Causal and Explanatory Asymmetries” *Theoria” pp* 1-52.

# Woodward, J. (2021b) “Explanatory autonomy: the role of proportionality, stability, and conditional irrelevance" *Synthese* 198: 237–265.

**Appendix: Additional Illustrations of the Distinction between Networks and Dynamics**

I observed above that while network structure can influence outcomes so also for dynamical assumptions about the processes occurring along the network. Indeed, different dynamics can produce different outcomes on the same network and in some cases, a specified dynamics will produce similar outcomes on networks that differ considerably in their structure. In other cases it may be reasonable to claim that, given a network structure, a range of different dynamics will lead to similar outcomes. In this section I provide more illustrations and detail in support of these claims, focusing first on different models of disease dynamics from epidemiology and then on issues about controllability in brain networks.

 The SIR model of disease dynamics was briefly discussed in Section 4. The even simpler SI model assumes that that there are only two possible states, susceptible (*S*) and infected (*I*) with a specified transmission rate *b* when a susceptible person comes in contact with someone who is infected. Here it is assumed that those who are infected do not become infected again. A third possible model, the *SIS* model, differs from the *SI* and *SIR* model in allowing individuals who become infected and recover to be reinfected with rate *a*. Each model can be combined with different assumptions about the network over which the disease spreads. Different dynamical models can predict different long-term outcomes on the same network and, for some dynamics, some outcomes are the same regardless of network structure. For example, with an *SI* dynamics, every individual who can be infected (that is, every individual in a component to which an infected person belongs) will be infected in the limit as *t--> ∞* (Newman, 2010, 640). The further details of the structure of the network do not matter to this long run behavior. By contrast with a dynamics described by the *SIR* model, this is not the case (Newman, 2010, 642). Instead with this dynamics, and infected individuals within a component, the disease can die out without infecting everyone in that component, with the probability of this happening depending on the transmission rate and the recovery rate. In general, in the *SIR* model, and in contrast to the *SI* model, there will be transition thresholds depending on the value of the transmission rate *b*, in which the disease goes fairly abruptly from being isolated in small clusters to becoming endemic, in analogy to percolation phenomena of the sort studied in physics (Newman, 2010, 644) . In the *SIS* model with *b > ga*, the whole population is not affected but the disease settles down to a steady state. (Recall that *g* is the recovery rate.)

The temporal evolution of a disease (as opposed to its long run state) is even more obviously governed in part by its dynamics and by parameters like the transmission and recovery rate. For example, Newman (2010) derives the following equation for the spread *x(t)* of a disease over time under the assumptions of the SIR model:

*x(t) = ∑ ar (o)* ***v****r e (bkr-g) t.*

Here the first terms are a function of the adjacency matrix (and thus the network structure) but the constant in the exponential term is a function of the infection the infection and recovery rate, just as one would expect.

1. "Process" is not meant to suggest anything metaphysical or something necessarily meeting the requirements of a "process theory" of causation. [↑](#footnote-ref-1)
2. The networks I will discuss are in most cases represented by undirected graphs and with unweighted edges. These contrast with directed networks, particularly when used to represent causal relationships, as with so-called causal Bayes nets. The latter *can* be used to explain why a node is in a certain state. I lack space for a detailed comparison, but these function very differently from explanations from undirected graphs. [↑](#footnote-ref-2)
3. Thus while it is part of the definition of an IDE that it includes dynamical information as well as information about network structure, my claim that many of the explanations that appeal to network structure are IDEs is an empirical claim about that literature. [↑](#footnote-ref-3)
4. In explanations of robustness and fragility, claims about the centrality or not of nodes in a network and other claims about network structure are used to support claims about what would happen if that node were removed: removal of central nodes will be more disruptive than removal of non-central nodes. For example, it may be claimed that structures in which most nodes are non-central will be robust to removal of those nodes, although not to the removal of central nodes. Such claims rest on an additional empirical assumption-- that the network will not reorganize on a global scale with the removal of a node but will instead retain most or all of its previous structure, just absent the affected node. [↑](#footnote-ref-4)
5. This understanding of a limiting or special case of a more generic one is common in science. For example, when a body falls through a medium that provides resistance this is a generic case; fall through a vacuum is a special case of this in which the resistance is zero. Ordinary language (as well as first order logic) can be misleading about such cases because it may suggest treating the case as one in which non-zero resistance is present is "special" in the sense that this adds an additional condition to the case in which there is zero resistance. Obviously on this understanding the case in which there is zero resistance and the case in which there is non-zero resistance are disjoint. The notion of special case that I am referencing in connection with IDEs behaves differently from this. [↑](#footnote-ref-5)
6. As in the well-known Konigsberg bridge example in which the explanandum can be construed as the impossibility of traversing each bridge exactly once. [↑](#footnote-ref-6)
7. My notion of constraint here is very similar to the notion discussed by Ross in her 2023. The examples of circuits, roads and other constraints on patterns of flow are also used in Ross, 2021, 2023. The notion of a constraint is used in somewhat different ways by other writers including Bechtel, 2018 and Chemero and Silberstein, 2013. I lack space to discuss the relation between my notion and these. [↑](#footnote-ref-7)
8. Again, assuming that we are dealing with graphs with unweighted edges. The use of edges that are not further distinguished is appropriate to the extent that is true that whatever differences may exist in interactions along individual units, these are less important to the outcome to be explained than the overall pattern of connectivity of the network. Similarly for differences among the nodes. Needless to say, it is empirical matter whether this is the case for various systems. [↑](#footnote-ref-8)
9. More generally, network representations are likely to be most fruitful when mean-field analyses (that is, analyses in which the interaction between a single component in a system and all of the other components can be understood in terms of the mean of the aggregate behavior of the other components) or other representations based on simple averaging (as is the case with the dilute gas) are inadequate and instead information about how some components and not others are connected up in such a way that they are able to influence one another is required. [↑](#footnote-ref-9)
10. By "support" I do not intend anything metaphysical like "grounding". I just mean that there are physical or empirical stucture that mediate or constrain the possibilities of causal interaction or transmission. [↑](#footnote-ref-10)
11. Again, how the traffic will in fact move depends not just on the road structure but on such dynamics involving facts as the traffic density, vehicle type and so on. [↑](#footnote-ref-11)
12. Roughly a network with high clustering and small average shortest path length. [↑](#footnote-ref-12)
13. Rathkopf (2018) is an important exception. [↑](#footnote-ref-13)
14. Watts and Strogatz do speak at one point of the network structure "determining" the dynamics. But I take it that what they mean by "dynamics" here is the overall behavior of the system-- that is, the combined upshot of the network structure and the dynamics of the SIR model. [↑](#footnote-ref-14)
15. The *causal* relations among different neural areas that reflect the transmission of neural signals reflect the brain's *effective* connectivity, which contrasts with the *structural* connectivity reflecting the presence of anatomical connections which make signal transmission possible and also with functional *connectivity* which reflects patterns of correlations between activity in neural areas that need not be causal. The presence of a structural connection does not by itself determine whether an effective connection is present; effective connectivity is the result of both structural connectivity and whatever dynamics governs neural transmissions. The very fact that there is a distinction between structural and effective connectivity shows that network structure in the sense of anatomical structure is not by itself sufficient to provide an explanation of how the brain works. [↑](#footnote-ref-15)
16. Of course, one may be able to formulate a relationship describing the changes in functional activity over time but this is not a dynamics in the sense under discussion. [↑](#footnote-ref-16)
17. It is important to note the difference between (i) the claim that, given some specified network structure, an outcome would be the same if one dynamical assumption was replaced with another or even replaced with any dynamical assumption within some large set, although there are some dynamical assumptions for which the outcome would be different and (ii) the claim that dynamical assumptions play no role in the outcome-- i.e., that the outcome would be the same for no matter what dynamical assumptions are adopted. It is only (ii) that justifies the claim that the dynamics is irrelevant. The examples discussed in Section 4 are cases of (i) but not (ii). [↑](#footnote-ref-17)
18. That the explanandum in this example is an impossibility is also emphasized in Lange, 2018 and Woodward, 2018. [↑](#footnote-ref-18)
19. Thus, it is not clear that the graph per se (or the structure represented by the graph) is doing much explaining here--it seems the graph just encodes facts like (i) and (ii) which do the explaining. Note also that the quasi-dynamical assumption (i) is a crucial part of the explanation-- the absence of a road from A to B only matters because it is assumed that cars can only travel along roads. [↑](#footnote-ref-19)
20. The example discussed in Section 4 in which whether a disease spreads through an entire compartment depends just on the network structure and not on the probability *p* of transmission is another case in which an outcome is entirely independent of a "dynamical" parameter. Again, one might think of this as an impossibility result -- it is impossible that the disease not spread. [↑](#footnote-ref-20)
21. Some may think that in at least some cases these dynamical assumptions should not be interpreted causally. Whether or not this is correct, the assumptions do at least concern propagation, or spread of something, etc. These are not assumptions about the network structure itself. [↑](#footnote-ref-21)
22. "Greater explanatory weight" here means simply that for a given network any one of a large range of dynamics will lead to the same result, while for different network structures, these generic dynamical structures lead to different results. [↑](#footnote-ref-22)
23. For example, a disease will only spread through a compartment if the starting state is one in which at least one individual in the compartment is infected. Notice that the network itself does not provide this explanation. [↑](#footnote-ref-23)
24. The notion of an intervention in Woodward, 2003 applies to manipulations that change the values of variables in an unconfounded way. In the case of networks, we may want to talk analogously about manipulations that remove a node or an edge, but we also want to be able to talk about manipulations that change more global features of networks such as whether they are small world. This is not the place to attempt a more precise definition but think in terms of a change in network structure that affects the explanandum of interest but does not affect the explanandum via some route that does not go through the network structure. Thus, a process that both alters the network structure and the dynamics with both in turn affecting the explanandum will not meet this condition. By contrast, a change in the brain that just removes a structural connection but does not alter the equations governing the dynamics will qualify as intervention-like. [↑](#footnote-ref-24)
25. If intervening to change this connectivity pattern seems far fetched as a practical matter, consider that this is likely what (or part of what) anti-depressant drugs do. [↑](#footnote-ref-25)
26. A referee interprets me as arguing that "topological properties are not "non-causal" or "non-physical" because they are "grounded "in physical facts. The referee objects that this is a non-sequitur, because (according to the referee) non-physical facts can be grounded in physical facts. I agree that this is a non-sequitur. However, this is not my argument which says nothing about "grounding" and does not rest on the claim that topological properties are grounded in physical facts. My argument is that when a physical system is described mathematically, one is not entitled to infer just on that basis that the system is being described in terms of non-physical, non- empirical or non-causal properties. I don't offer a positive account of what is going on when mathematics is used to describe a physical system-- I merely claim that the inference described above is invalid. I will add, for what it is worth, that even if one thinks that topological descriptions ascribe topological properties and these are different from the more specific physical properties that (to use what I hope is a neutral word) "underlie" these, it does not follow that the topological properties are non-physical or non-causal. I will also add that it is far from obvious to me that the right way to talk about the use of mathematics to describe the physical world is in terms of mathematics ascribing distinct "properties" that have to be related in some way to physical "properties", with attendant puzzles about how to understand this relation. [↑](#footnote-ref-26)
27. Consider in the light of these observations Kostic's claim that "topological properties of a given object [ in this case topological properties in the sense of the connectivity of a network- JW] are derived from its spatial relationships with the other parts of a system" and similar claims by other writers. Even when the network structure corresponds to well-defined spatial relations this seems misleading. Different neural areas stand in all sorts of different spatial relations but this by itself does not tell us what the relevant network connectivity is for tracing neural signaling. We need in addition to recognize which are the connections that transmit neural signals-- and this is not something that can be "derived" just from information about spatial relationships. Of course, spatial relations constrain what sorts of transmission are possible.

 [↑](#footnote-ref-27)
28. Here I agree with Craver, (2016). [↑](#footnote-ref-28)
29. Roughly these special circumstances correspond to a failure of faithfulness in the sense of Spirtes et al. 2000. [↑](#footnote-ref-29)
30. More weakly, suppose simply that *Z* and *Y* are correlated. Then if *Y* causes *X*, we would expect (absent special circumstances) that *X* and *Z* are correlated and, *ex hypothesi*, they are not. [↑](#footnote-ref-30)