**Note to Reader: This a reasonably stable draft of another probably too long to publish paper.**

**Networks, Dynamics and Explanation[[1]](#footnote-1)\***

**James Woodward**

**10.24.23**

**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, interact with other proteins. But in addition networks are also used to explain. The main focus of this paper is one particular form that such explanations can take. Put informally these are cases in which what is of interest is the spread of some process[[2]](#footnote-2) 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). 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 some mental disorder is present), 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. As noted below, they are not the only the only explanatory use of networks but they have distinctive features, particularly having to do the role played by independent dynamical assumptions governing the processes in question. For this reason I will call them independent dynamics explanations (IDEs).

Some of the cases in which networks have been claimed to explain in the philosphical literature differ from IDEs and one of my messages will be that discussions of network explanations need to be sensitive to these differences. I will also suggest that many standard network explanations are not (as sometimes claimed) "autonomous" or "stand alone" in the sense that they are independent of causal information. In addition, I will provide an account of directionality in network explanations. I have tried to indicate when my views agree with and depart from those elsewhere in the philosophical literature but this will not be my main focus. Finally, the networks I will discuss are in most cases undirected and edges (or the corresponding adjacency matrix) are unweighted. Directed networks, particularly when used to represent causal relationships are important in many areas of science but these function very differently in explanations from undirected graphs, as discussed in Section 19.

The rest of this paper is organized as follows. Section 2 introduces the idea that in many characteristic applications networks represent *constraints* concerning which components of a system can interact with others. Section 3 argues that such constraints are supported by physical structures. (A later Section --11-- elaborates on this idea. ) Section 4 introduces 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 Section 15. Sections 5- 8 distinguish such IDE cases from other uses of networks in explanation such as the explanation of impossibilities or of system robustness, in which dynamical considerations of the sort described in Section 4 play a less important role. Section 9 returns to networks with an independent dynamics and introduces some further nuances concerning the issue of whether network structure by itself can be explanatory. After some brief remarks on network discovery (Section 10), Section 12 draws together the implications of previous sections to spells out a general schema to which I believe network explanations conform. Sections 13 and 14 develop the implications of this schema in greater detail. Section 16 discusses the issue of whether network explanations are "autonomous", "stand alone" or independent of causal considerations, answering this question negatively for many network explanations. Section 17 uses the ideas developed in previous sections to provide an account of the directionality of network explanations and Section 19 compares network explanations with the explanations provided by causal graphs. Section 20 concludes by taking up the issue of when network explanations are most likely to be fruitful.

**2. Networks as Representing Constraints[[3]](#footnote-3)**

I begin with some general features of explanatory uses of networks, both IDEs and other varieties. 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[[4]](#footnote-4)-- differences in their internal structure is 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 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 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 below). 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 for 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[[5]](#footnote-5).

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 influence or interaction between these areas. (The correlation might arise because both areas are influenced by some third area.) However, as argued below, in many (perhaps all) cases of this sort the network does not have an independent dynamics and cannot figure in explanations.

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

When an edge represents such constraints, often but not always (always in the case of IDEs-- see below) 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. In the case of a brain, this 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. 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 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[[6]](#footnote-6).

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[[7]](#footnote-7). But, as explained in more detail in Section 11, 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. 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. As I suggest below, the absence of such additional commitments is part of the power of network representations.

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

In many cases of interest (those called IDEs above) 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. (I also distinguish (ii) from (iii) factors that affect the growth of networks -- how and why they change over time-- see Section 5.) Following a very common usage, I will call these rules or equations (and the behavior they describe) 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. It is for this reason I have called these Independent Dynamics cases. In such cases, the overall behavior of a system represented by a network will depend on both the network structure *and* the assumed dynamics.

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 small world network 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[[8]](#footnote-8). Thus Watts and Strogatz's model is *not* one in which disease spread depends on the network structure alone, independently of dynamical considerations[[9]](#footnote-9).

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

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

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

*(*3*) dR/dt= gI*

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

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 we shall see below (Section 15) 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 which will be 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[[10]](#footnote-10).

As yet another illustration (cf. Newman, 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 in a small interval of time the flow from *i* to *j* is

(4) *C (Sj-Si)dt.*

(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

(5) *dSi/dt= C ΣAij (Sj-Si)*

where *Aij* is the adjacency matrix for the network. This matrix encodes the connectivity or topology of the network and we see from (4) and (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) 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 another example of what I mean when I talk about the dynamics governing the processes that occur along a network being 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 (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 time scale than the time scale 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 time scale 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 wear out and 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 occuring on it.

Finally, 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. (*Functional* connectivty which describes temporal correlations among patterns of brain activity is one illustration-- see Section 12 below.) In such cases, in contrast to the examples considered above, the edges do not represent constraints on possible interactions. Thus, it will not make sense to specify an independent dynamics describing a process occuring along the network[[11]](#footnote-11). 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.

**5. Processes Governing Network Growth and Change**

In addition to (i) the dynamics governing the processes that may occur along networks, there are other sorts of relevant dynamics which may somewhat overlap with (i) or may be largely distinct. These have to do with whatever rules or principles of change describe how networks grow or respond to external perturbations. For example, some networks are thought to grow by a process of preferential attachment--that is, nodes with lots of edges are more likely to acquire additional edges than nodes with fewer edges, with several different mechanisms proposed for how this might happen in particular cases (e.g., Barabasi, 2016). This growth process in turn leads to a power law distribution for the edges and is thought to characterize a number of networks that have been investigated empirically, such as citation networks (papers that are already cited a lot tend to be cited more). Growth processes of this sort are also thought to often occur more slowly than particular instances of processes occurring along the network, although, as we saw with the example of synaptic growth, processes occurring along a network can cause network changes. Obviously explaining why and how, if at all, a network grows or changes or responds to perturbations is different from using that network to explain other explananda-- in the former case, the network structure is the explanandum, rather than part of an explanans

**6. Explaining Change Requires Dynamics**

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. (See Section 12.) 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, relatively static and unchanging and 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

However, there are putative cases of in which networks are used to explain that do not fit the pattern just described. The next two sections discuss two kinds of cases of this sort-- the first involves the use of networks to explain impossibilities. The second involves appeals to networks to explain facts about the robustness or non-robustness of a system to external perturbations.

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

One frequently discussed example is Euler's analysis of the bridges of Koningsberg (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 Koningsberg bridge configuration does not satisfy this condition and the result that it is impossible to cross each bridge exactly once follows. It might be argued that there is a sense in which even this example builds in dynamical assumptions 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 is at best very minimal (perhaps we can think of it as presupposed in the statement of the problem rather than a separate independent assumption) and this is reflected in several features that distinguish this example from many of the others considered above. First, the explanandum in the ngsberg bridge case is naturally regarded as an *impossibility*[[12]](#footnote-12)*,* 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 Koningsberg 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 use of the network in this case seems to just draw on its ability to represent possibilities and impossibilities. Suppose that (i) cars can only travel along roads and that (ii) there is no road leading from location A to location B-- 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. I take it there is nothing mysterious about this-- the impossibility result 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 is the same.

The moral that I draw from this is that Koningsberg bridge and similar examples are importantly different from the examples discussed above (IDEAs) in which it is important that there is the spread of some process occurring along the network. In particular, because of this difference, the Koningsberg bridge case seems more like a case in which the explanation appeals only to network structure, with no or a minimal role for dynamical considerations. If you like, it is an "edge" case which is not representative of many other network explanations. Many other explanations that appeal to networks do not have these "stand alone" features.

**8. Interventions that Change the Structure of a Network**.

I emphasized above that it is common (and in many cases warranted) to assume that a network remains relatively unchanged as some dynamics governed process occurs on it. This may be a particularly plausible assumption if the network is just being used to predict the outcome of some process running on the network in the absence of any intervention that alters the network. An example is provided by the use of a contact network to help predict the course of a disease where it is assumed that quarantining or voluntary isolation does not occur that alters the structure of the network. Matters become more complicated if there is an intervention that alters the structure of the network-- e. g., by removing a node or removing some of the edges into it. Examples of such interventions are frequently discussed in the literature on networks. They include cases in which a node in an) airline transportation system is disabled or a certain species in a food network is removed from the network either via extinction or an external intervention (Huneman, 2010). A very common intuition about such cases is that the effects of deleting a highly connected node (or more precisely a node with a high centrality coefficient according to some measure[[13]](#footnote-13) ) will be very different from deleting a node which is less central -- shutting down O'Hare airport will be much more disruptive to the entire system than shutting down a small regional airport. Similarly certain network structures may be more vulnerable (in the sense of experiencing more extensive disruptions) to random deletions of nodes and edges than other network structures. If responsiveness to such deletions is taken to be a measure of a systems's robustness or stability, then (it is claimed) we can appeal to network structure to explain facts about a system's robustness or stability. For example, the robustness of ecological networks to species extinction or removal can be analyzed in this way. (Huneman, 2010).

Let me make two observations about such examples. First, these represent another kind of case which networks are purportedly used to explain and in which the dynamics governing the processes occurring along the network (at least insofar as these are different from whatever processes govern the network's response to perturbation-- see below) may not matter much to the explanation. But in contrast to Koningsberg bridge type examples, the explananda here have to do with robustness and response to perturbations rather than with what is possible or impossible. Second, while such explanations (or predictions) of robustness etc. may well be correct in particular cases, it should be recognized that they rest on an additional assumption that does not follow just from information about the original network. For this reason it is misleading to think of the explanations in question as appealing just to the structure of the network alone. The additional assumption is that the effects of the intervention are only local in the sense that the targeted node and edges are removed but the network is otherwise unchanged. (As discussed below, this is sometimes called a modularity assumption when it is made in connection with a directed graph used to represent causal relations-- see Woodward, 2003.) This assumption may be false as an empirical matter and the original network structure by itself does not tell us whether this is the case. It is certainly possible, for example, that if a species with high node centrality in a food network were to be eliminated, the entire structure of the network would reconfigure, with very different patterns of predation emerging. In such a case the removal of this species might have very different consequences from what is suggested just by considering the original network with the species removed. Of course in a case of shutting down an airline hub like O'Hare we do not expect such a reconfiguration to happen quickly but this is because of additional background information we have about how long it takes alter complicated scheduling logistics, expand the size of airports or build new ones and so on. In cases like this to predict what will happen we often need information both about the dynamics along the network (and how this may change in response to the intervention) as well as information about the processes that govern growth or change in the network (as, including the responses of these to interventions. Thus in these cases too the network by itself does not provide a stand-alone explanation (or even a prediction), independently of these additional considerations, which seem to be largely causal in character. On the other hand, in cases in which we have good "external" reasons to believe that a network will not change globally in response to a local intervention, knowledge of network structure can be very useful in predicting and explaning the results of such changes[[14]](#footnote-14).

**9. Some Consequences of the Independence of Dynamics and Network Structure Along with some Nuances and Complications**

The fact that in independent existence 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[[15]](#footnote-15) (in addition to the weak 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, 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. This is an issue that has led to considerable puzzlement and controversy in the recent literature. Having emphasized that, in independent existence cases, outcomes will in commonly reflect both network structure and dynamics let me now add some nuances. First, once one specifies or assumes a dynamics, one can of course ask what difference one network structure rather than another makes to an outcome. For example, given a 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 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 this or that general form. 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.

As an illustration, 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, although it will affect how quickly it occurs.

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)[[16]](#footnote-16). I suspect that it is possibilities like this that encourage some to construe explanations of disease spread etc. as explanations that appeal only to network structure and nothing else.

Of course the extent to which the details of the dynamics don't matter to some set of outcomes will also depend on the outcomes in question-- that is, what we are trying to explain. For example, whether or not a disease will spread widely in the long run on a network (that is, whether in the temporal limit (t--> ∞) a large number of people will be infected) may depend to a large extent on the structure of the network and not so much on the details of the disease dynamics. However, the *rate* of spreading will certainly depend on the dynamics-- see section 15.

It is also important that in many cases in which networks are employed, there may be little information about the details of the dynamics. In such cases researchers may in effect be assuming (or hoping) that whatever these details may be, they will matter relatively little (that is, over and above very generic features of the dynamics that are assumed) and thus can be safely abstracted away from, with the outcomes of interest depending largely just on various structural features of the network. These have the advantage that in many cases they are easier to discover than the details of the dynamics. For example, as discussed in Section 15 below, Bassett and colleagues, explore, 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 of course a broadly causal notion and reaching conclusions about controllability thus requires some dynamical assumptions. In their models Bassett 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. In such cases, it is not as though the dynamics don't matter at all; rather it is assumed that the conclusions are relatively insensitive to the details of the dynamics.

**10**. **Discovering Networks.**

I suggested in 5) that network structures may in some cases be easier to discover than the dynamics governing a system. This reflects the fact that in a number of cases we have means for detecting such structures that are independent or largely so of dynamical and other sorts of more "theoretical" assumptions[[17]](#footnote-17). For example, techniques like diffusion tensor imaging may be used to learn about structural connections in the brain without making assumptions about the dynamics of neural signaling (Of course, as noted above, it is a further question whether the structures so detected are actually used to transmit such signals-- in this case they often are, although the structures may not be very informative about the quantitative behavior of such signaling[[18]](#footnote-18)). In the case of traffic flow it is often straightforward to learn what the relevant road network is and one does not need a dynamic model of traffic flow to do this. Similarly a contact network can often be estimated from information about spatial proximity in a way that does not require modeling of how a disease or rumor spreads along the network, which is a different matter. Given that it may be easier to learn a network (or that claims about it may be more evidentially secure) than dynamical claims, it makes sense to explore the extent to which variations in outcomes can be explained largely by differences in network structure with only weak or generic dynamical assumptions. This provides an additional rationale for a focus on networks.

**11. More on the Role of Networks as Representing Constraints**

In this section I expand on the idea, introduced above, that an important role of networks, central to their explanatory use in many cases (particularly independent existence cases) is to describe constraints on possible interactions or influence. In a very illuminating paper, Ross (2023) characterizes what she calls constraint-based explanations in the following way:

In constraint-based explanations, the constraints that are cited in the explanation are factors that: (1) limit the values of the explanatory target of i nterest, (2) are often conceived of as separate from or external to the process they limit, (3) are consid ered relatively fixed compared to other explanatory factors, and (4) structure or guide the explanandum outcome, as opposed to triggering it.

This is very close to my own view about the way in which network structure functions in explanations. Readers will note that Ross' condition (2) closely resembles my claims above about the independence of the network structure (in independence existence cases) from the processes that occur along it and from the dynamics that characterizes these. Ross' condition (3) parallels my claim that networks typically change, if at all, on a time scale that is much slower than the time scale governing the dynamics. Condition (1) corresponds to my view that when networks are used in an explanatory way under discussion, they typically describe interactions or processes which are "possible" or "permitted" (or, alternatively, "impossible") with (in many cases) the dynamics and other information providing additional detail which is used to determine what actually happens.

As Ross also notes another way of describing the constraint idea is that we can often think of the network structure in terms of notions like "enabling" or "allowing" --- the circuit structure is such as to enable the current to reach the light, the blood vessels allow the blood to reach its cellular targets and so on. Put yet another way, networks encode non-specific information in the sense that in many cases they don't by themselves determine or "force" a particular outcome but instead are compatible with a range of these, with what actually happens also depending on the dynamics. As a simple illustration, suppose that the only way to get from A to B is via a bridge. Jones crosses the bridge from A in order to rob a bank in B. The bridge enables or allows Jones to carry out the robbery but the existence of the bridge is compatible with many other outcomes as well-- Jones might have crossed the bridge to go shopping instead. In this sense the role of the bridge is non-specific, in contrast Jones' intentions which have a more specific relation to the outcome.

Ordinary language often distinguishes between causes and enabling conditions, but nonetheless (as emphasized by Ross) there is an obvious sense in which the constraints described by a network are the upshot of the operation of various causal factors . When blood flow is constrained by the walls of a blood vessel there are complicated causal interactions occurring between the blood flow itself and the vessel walls. If the spread of a disease is constrained in such a way that it can only be transmitted by relatively close contact, there is a complicated causal story about the mechanism (e.g., virus containing particles expelled during respiration and then inhaled by others) by which the disease is transmitted.

A natural way to think about the network representation in such cases is that it "black boxes" this more detailed causal information. That is to say, it is assumed that for the explanatory purposes at hand these causal details do not matter except to the extent that they are reflected in the constraints described by the network structure. Insofar as we are interested in how a disease spreads in a population, the network portion of this explanation just needs to encode the assumption that the only possible way for the disease to spread is by contact (assuming this is true) -- additional causal details about the exact mechanism of transmission do not matter, independently of this. Similarly a network representation of a system of blood vessels black boxes the causal details of the interaction between the vessel walls and the movement of the blood-- what matters and what is encoded in the network is simply that these causal details operate in such a way as to constrain the blood to remain inside the vessels.

This picture of constraints is very similar to the use of this notion in other scientific contexts that do not involve networks. Consider the elementary treatment of a block, subject to gravity, sliding down an inclined plane. It is typically stipulated that the block is "constrained" to remain on the surface of the plane-- it cannot fly off the plane or fall through it. There is a very complicated causal story about why this constraint holds involving, among other considerations, complex contact forces between the block and the plane. But in stipulating that the constraint holds, we ignore or black box this complexity, recording only that it operates in such a way that the block remains on the plane. This enormously simplifies the analysis of the example-- we don't need to model the details of the contact forces and instead can focus only on the role of gravity and employ a very coarse-grained representation of the net frictional force opposing the motion of the block. Similar remarks apply to other textbook examples, such as a particle under the influence of a force but constrained to move on the surface of a sphere. Here this constraint can be represented as the requirement that the coordinates of the particle must conform to the relation x2 + y2 +z2- a2= 0, where a is the radius of the sphere[[19]](#footnote-19). Again this constraint represents that certain forces or causes are operative that restrict the motion of the particle but without explicitly modeling those forces.

From this point of view, the use of the network representation involves a strategy similar to what Mark Wilson calls *physics avoidance* (although in this case it might more accurately be described as science avoidance): avoiding representing or including in our analysis various causal details, by providing a kind of coarse-grained summary of what is relevant about them-- where, in the case of networks, this has to do with which interactions are possible[[20]](#footnote-20). This allows us to focus on other aspects of a system's behavior. That networks can be used in this way is another reason why they are so fruitful.

This picture allows us to make sense of other features of many network representations that may intially seem puzzling. For example, if we wanted to provide a detailed representation of *causal* relations between nodes in a network we would need, at a minimum, to employ a directed graph, since causal relations are certainly directed. However, as noted above, many of the networks employed in science, including those which are used for explanatory purposes, employ undirected graphs. Moreover, even ignoring the fact that they are undirected, there are often other grounds for thinking that a straightforward causal interpretation of the edges is inappropriate. For one thing, nodes in networks are often most naturally interpreted as objects, rather than variables or values of variables, as is usually required for full fledged causal interpretation in a causal graph. A road from Chicago to New York can be represented as an edge in a network but this is not to be interpreted as a claim about Chicago "causing" New York or vice -versa. Similarly for an edge representing that person A is in close physical contact with person B. Again what the edge represents in these cases is something more like the possibility of a general kind of causal influence being transmitted or the possibility of a process occurring along the edge that has causal import, but without any much more detailed characterization of that influence being provided or why it is constrained in this particular way. For this purpose, an undirected edge may suffice and as we have noted it is often much easier to provide evidence for the existence of such an edge/constraint than to provide further causal details.

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

With this as background I turn to a more explicit discussion of how networks figure in explanations. This is primarily intended to cover independent dynamics cases, but it arguably also covers cases like the koningsberg bridge case. I begin with a general schema **W** which will be refined and complicated below. **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[[21]](#footnote-21). 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 intitial state, information about this state should be represented.

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

**W1- W3** constitute an explanans for the explanandum E in **W4**:

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

By a dependency relation (**W3**) between **W1**, **W2** and E I mean a relationship such that it is possible to intervene on the network N so as to change its structure, holding fixed the dynamics D, with the result that E changes systematically under this intervention on N. ("Intervention" is understood in broadly the sense of Woodward, 2003.) In other words, it must be possible to change E by manipulating the structure of the network via an intervention. In many cases the overall dependency relation between E and **W1** and **W2** will be a contingent relation that, as a matter of empirical fact, obtains under some range of interventions rather than a relationship that corresponds to a mathematical truth. In such cases, it is often appropriate to regard the explanation of E in terms of **W1-W3** as broadly causal. (It qualifies as causal within an interventionist framework since the relationship is intervention-supporting and empirical.) For example, the relationships that appear to 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 alone 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). Thus if it is true that for some hypothesized dynamics or for whatever dynamics in fact occurs in the brain, there is a dependency relation between the structural network structure (or for that matter the effective connectivity network) and whether or not a disorder is present, that relation is plausibly regarded as causal. However, in some 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 will not count as causal-- the Koningsberg bridge case 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 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 (or even any 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 anatomical 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[[22]](#footnote-22).

Next consider the bridges of Koningsberg[[23]](#footnote-23). Assuming a "walking dynamics" (which in effect is what one assumes when one asks whether an Eulerian path exists) , the impossibility of crossing each bridge exactly once follows as a matter of mathematical necessity from the fact F that the bridge structure is not a connected one in which there exists 0 or 2 nodes with odd degree . Again assuming the same dynamics, if we were to alter the configuration of bridges so that F is not present (the network is fully connected and there are 0 or 2 nodes of odd degree) , the explanandum would change-- it would then be possible to cross each bridge exactly once. Note also that although (assuming the walking dynamics) the relation between F and whether it is possible to cross each bridge exactly once is "mathematical" there is nothing problematic about the notion of intervening to alter the connectivity structure of the bridges-- one might destroy bridges or build new ones. Thus, according to schema **W** the relation between bridge structure and whether it is possible to cross each bridge exactly once is explanatory but not causal.

Note also that in all of these cases, schema **W** requires more than merely that the conditions cited in the explanans be sufficient (even nomologically or mathematically) for the explanandum E. It is also required that changes in those conditions under interventions are associated with changes in E-- the what-if -things-had-been-different condition described in Woodward, 2003. According to **W**, the explanations under discussion work by exhibiting how their explananda depend on the factors cited in their explanans-- a feature also shared by many non-network explanations. This is one of several reasons why it seems inadequate to locate their explanatory power merely in the fact that their explanans necessitate their explananda, perhaps with some particularly strong flavor of necessitation. As a number of familiar examples involving other sorts of causal or nomological show, S can be sufficient for E without explaining E. The same seems true of network explanations.

Schema **W** is similar in some respects to the schema T advanced by Kostic (e.g., 2022 ) 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 two crucial 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 **W1**, **W2** and the explanandum E. This variety of counterfactual dependence is captured by *interventionist counterfactuals* --- it must be the case that if one were to intervene to change the network structure in **W1**, E would change. Kostic and Khalifa do not impose a restriction of this sort in their schema and, as is well-known, there are many counterfactuals that are true under some interpretation but not under an interventionist interpretation and do not seem to be explanation-supporting. For example, there are natural ways of interpreting the counterfactuals

If it were impossible to cross each bridge connecting a set of land masses exactly once, then that bridge structure would lack the Eulerian feature F

If it were possible to cross each bridge connecting a set of land masses exactly once, then that bridge structure would possess the Eulerian feature F

according to which these counterfactuals are true. But although whether the bridge structure possesses feature F counterfactually depends on whether it is possible to traverse each bridge exactly once, presumably no one thinks that the possibility or not of traversing in this way explains why the bridges do or do not have feature F. An interventionist account of counterfactual dependence captures this assessment: intervening to change whether or not the bridges can be traversed exactly once will not change the connectivity structure of the bridges, thus capturing our sense that is is the network structure that helps explain whether or not each bridge can be crossed once but not vice-versa. (I say more about this issue in Section 17 below-- changing whether the bridges are traversable without changing the connectivity of the bridges can be accompished by, for example, expanding the abilities of those traversing the bridges.) By contrast, without restrictions on what counts as the right kind of counterfactual dependence in schema T, it is unclear why this "backwards" dependence does not count as explanatory.

A similar point holds for relations of counterfactual dependence that track evidential relationships but are not intervention supporting. For example, there are good reasons to think that abnormal patterns of functional connectivity can be evidence for mental disorders and thus that there is a kind of counterfactual dependence between whether or not such disorders are present and patterns of functional connectivity. However, it is a further question whether intervening to change a patients functional activity will change whether that patient has a mental disorder. According to schema **W** it is only if this is the case that the pattern of functional activity explains the presence of the disorder. A common view among researchers is that abnormal patterns of functional connectivity are *effects* of mental disorders (or of whatever brain pathologies underlie the disorder) rather than causes (or explanations) of such disorders and that when patients are classified by different patterns of functional activity we are classifying on the basis on effects of such disorders. Friston (2011) distinguishes effective and functional connectivity as follows:

Functional connectivity has a distinct role from effective connectivity. 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 di- agnostic 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)

..., one can associate the generative models of effective connectivity with hypotheses about how the brain works, while analyses of functional connectivity address the more pragmatic issue of how to classify or distinguish subjects given some measurement of distributed brain activity. In the latter setting, functional connectivity simply serves as a useful summary of distributed activity, usually reduced to covariances or correlations among different brain regions (p 15)

Consider in this light Kostic and Khalifa's (2021, 2022 ) claims that one can appeal to functional connectivity to explain:

This is a *prima facie* methodological reason to countenance functional connectivity explanations that are autonomous of any mechanistic explanation. Recall that functional connectivity edges are synchronization likelihoods. Thus, many of the functional connectivity models’ topological properties are different ways of describing how well synchronized a brain is.

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

Again, an obvious thought is that functional connectivity E is an effect of (and explained by) the combination of C1 neural dynamics and C2 structural connectivity so that in considering (as Kostic does) a counterfactual in which structural connectivity is fixed and the dynamics is claimed to depend on functional connectivity one is considering a counterfactual in which a cause C2 is fixed and the other cause C1 is claimed to depend on E. This kind of dependence is not explanatory dependence.

In the second quotation, Kostic seems to agree that evidential dependence is distinct from explanatory dependence but also claims that we can appeal to functional connectivity to explain. Of course I agree with the first part of this claim but it seems to me to make good on the second, Kostic needs some positive account of the kind of counterfactual dependence that is sufficient for explanation while also distinct from evidential dependence. It also needs to be shown that relations involving functional connectivity satisfy this counterfactual condition on explanation. An interventionist account provides a candidate for this explanatory condition.

**13. The Importance of Distinguishing Whether Edges within a Network Represent Causal or Explanatory Relations from Whether the Network Structure Can be Used to Explain**

In thinking about the explanatory role of networks it is important to distinguish two questions which are sometimes conflated: (i) whether a network is such that the state of one node can be used to explain the state of another node (which, as I see it, is the question of whether the edges in the network represent causal or some other explanatory relationship ) from (ii) the question of whether we can appeal to various structural features of the network itself (or more precisely the features of the system it represents) to explain other explananda having to do with the behavior of that system, taken as whole. (ii) has to do with whether e.g. the fact that a contact network is small world can figure in an explanation of disease spread while (i) has to do with whether the state of one node in that network explains the state of some other node. It is possibility (ii) that the schema **W** attempts to capture. If we assume (in my view plausibly) that explanatory relations, whether causal or not, must have a direction, then in a network with undirected edges (e.g a contact network, we cannot appeal to the state of one node to explain the state of another, at least absent additional information[[24]](#footnote-24). However it does *not* follow from this that we cannot appeal to a graph with undirected edges to provide explanations, including causal explanations, of various *other* explananda E. Here (to repeat) E will typically have to do with the behavior of the whole system characterized by the network--e.g., disease spread on the whole network. This last sort of explanation is legitimate as long as the conditions described in **W** are satisfied-- as long as it is true that intervening to change N the structure represented by the network will change E. Moreover, as I will argue below (section 17), such an explanation of E in terms of N (and, as we have seen, often other information involving the dynamics D) can be and typically is "directional" (that it is has a direction from N and D to E) even if the edges in the network are not directed and do not have a straightforward causal interpretation. (To anticipate and speaking very roughly, this directionality has to do with the independence of N and D.) For example, a network of roads is standardly represented by an undirected graph (perhaps particularly if the roads are two way) but it can still be true that we can appeal to features of the road network to explain patterns of traffic congestion, where this explanation has a definite direction and is (at least sometimes) causal-- that is, features of the road network (in conjunction with other assumptions) causally explain the congestion and not vice -versa. Similarly when the configuration of bridges and land masses in Koningsberg is represented by an undirected graph it would make little sense to interpret the individual edges as representations of causal or explanatory relations-- the undirected edge connecting land mass A and B does not mean that land mass A (or its state) "explains" land mass B. Despite this, facts about the configuration of the bridges do figure in a (directional) explanation of the impossibility of crossing each bridge exactly once.

As another illustration of the importance of distinguishing the (i) the possible explanatory role played by relations between nodes within a network from (ii) whether the network structure itself can be used to explain some other (non-node) explanandum, consider Huneman 's (2010) claim that the fact that causality must have a temporal dimension and the fact that a network topology lacks this dimension ("topology is not something that takes place in time"[[25]](#footnote-25)) implies that network explanations are non-causal. It is true of course that a system of roads or an anatomical neural network do not encode temporal information. One cannot use such networks by themselves to explain why some node is in a certain state given the state of other nodes that are connected to it. But again this does not preclude our appealing to facts about the structure of such networks to help explain other more general features of the systems they characterize such as the occurence of traffic congestion or abnormal brain functioning, especially if there are also dynamical assumptions in play in such an explanation.

A related point is that even if it is true that the edges in a network represent *causal* relations, this claim needs to be distinguished from the issue of whether the relation between the structural features of the network and some other explanandum playing the role of E in schema **W** is also a causal relation. As we have noted, there is considerable evidence that there are correlations between measures of neural connectivity and other features of behavior/life outcomes-- income, IQ, impulse control, drug use, various measure of health-- and so on. But even if the connectivity measures are interpreted causally (e.g., as measures of effective connectivity) it does not follow that that the correlations between such connectivity and the life outcomes are also causal-- that is a completely different issue. Effective connectivity has to do with whether activity in one neural area causally influences activity in another. It is a separate question whether the presence of some particular pattern of causal influences (rather than some alternative pattern of causal influences) itself causally influences life outcomes rather than merely being correlated with them (or whether such a causal claim even makes sense[[26]](#footnote-26)). In any case whether the possible causation here is to be attributed to effective connectivity or instead to whatever is responsible for it, the right way to test such claims would be to do something equivalent to intervening to alter effective connectivity or what is responsible for it and seeing how if at all life outcomes change. Again note that this is very different from the way in which one would test a claim about effective connectivity *within* a brain.

**14.** **Topological Properties**?

On the picture I have presented so far many networks (including those with an independent dynamics as well as others) 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 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 "material", so that when one appeals to network structure to explain some some explanandum, one is appealing to non- physical/non- material/non-empirical 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, Kostic's (2022) characterization of topological explanations requires that the explanans makes reference to "topological" properties which he thinks of as "non-causal" (see below) and also requires that the explanandum make reference to "empirical" properties with the apparent suggestion that these are two kinds of properties that should be contrasted with each other.

I think views of this sort introduce unnecessary complexities and threaten to make network explanations 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. 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 (that is the actual path ) 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. When we intervene on a structure represented by a network to change it to a structure represented by a different network, we intervene on physical features represented by the network, not some non-physical mathematical property, although when we do so, what we change is the system's connectivity, which is characterized via abstract concepts drawn from graph theory. 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.

**15. More on Networks and Dynamics**

I observed above that while the network structure can influence outcomes so can dynamical assumptions about the processes occuring 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, with a different dynamics leading to a different outcome. 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 this claim, focusing first on different models of disease dynamics from epidemiology and then on issues about controllability in brain networks.

I briefly mentioned the SIR model of disease dynamics 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 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.

The exploration of network controllability in Gu et al. (2015) provides another illustration of the interplay between network structure and dynamics. These authors are interested in the influence of brain circuitry on the "controllability" of different brain areas. They distinguish several different varieties of controllability but all have have to do with the extent to which interventions on one neural area can influence others. They employ network control theory to answer this question and they emphasize that doing so requires not just information about neural network structure but also assumptions about neural dynamics along that structure:

Graph theory specifically and network science more generally have provided a toolbox of diagnostics to describe the organization of graphs or networks. Yet, the relationships between this organization and the system’s function remain speculative at worst and correlative at best. Groundbreaking new discoveries will necessitate a fundamental turn from descriptive statistics towards mechanistic predictions. What are the mechanisms by which network structure affects functional dynamics? Moreover, how could one intervene in a network to push the system dynamics towards a specific, targeted goal? To address these questions, we must have a framework that i ncorporates not just brain network structure but also models neural dynamics. Network control theory offers exactly such a framework...

The authors acknowledge that actual neural dynamics are highly complex and non-linear but they nonetheless employ a simplified linear model of the dynamics, instantiated in an adjacency matrix which is weighted by assumed connection strengths. Thus what they investigate is controllability in the network defined by this matrix. They provide a number of different justifications for their choice of a linear dynamics:

Decades of research demonstrate that neural dynamics are nonlinear. Yet, our approach is built on a linear model of these dynamics, and it is therefore imperative to delineate its strengths and weaknesses. First, we note that nonlinear behaviour may be accurately approximated by linear behaviour in certain scenarios .... Second, we note that the controllability of a linearized model has implications for the controllability of a nonlinear model: if the linearized system is controllable, then the nonlinear system is locally controllable.

...controllers based on linearized dynamics can be effectively used for the control of nonlinear dynamics. Thus, while neural dynamics are inherently nonlinear, the study of linear models of neural dynamics can offer fundamental insights into system function.

Assuming this linear dynamics, the authors are then able to investigate how different patterns of neural circuitry in different brain areas provide different possibilities for control.

I take this discussion to illustrate the general pattern described in sections 12-14. Researchers assume a particular dynamics and investigate how, given this dynamics, different network structures affect outcomes, in this case outcomes having to do with controllability. In carrying out this exploration it may be acceptable to employ a simplified or "unrealistic" dynamics if there is reason to think that a more realistic dynamics would not yield fundamentally different conclusions about the consequences of network structure. In other words, it is assumed that the effects of network structure would be similar for a range of different dynamics, including an overly simplified choice. It is important, though, as the quotations from Gu et al. illustrate, that *some* assumptions about dynamics are necessary for this sort of investigation-- it is not the case that dynamics play no role at all.

**16. 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 alone provides free-standing explanations. 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 often do need to have enough information about these to identify the relevant network. [[27]](#footnote-27) A network of roads is the relevant network for understanding traffic flow because or to the extent that cars have 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[[28]](#footnote-28).

Consider in this light claims like the following[[29]](#footnote-29):

*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)

Topological explanations explain the dynamics of complex systems by making use of topological properties, i.e., properties of a complex system that are mathematically quantified using graph theory (Kostić, 2019).

In their seminal paper, Watts and Strogatz (1998) applied these concepts [small world vs random network structure etc.] to show how a network’s topological structure determines its dynamics. (Kostic and Khalifa, 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 processes. Although I have acknowledged that some network explanations (like Euler's treatment of the Koningsberg bridges) are non-causal (or perhaps "edge cases" in terms of their very minimal causal commitments), what is non-causal or mathematical in such cases is the *dependency relations* on which they rely rather than the fact they appeal to non-causal or mathematical properties (whatever exactly that means) in their explanans. In other words, the properties that are related, even in the Koningsberg bridge case, are ordinary physical properties-- it is the dependency relations between these properties that involve mathematical rather than causal connections.

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) these connectivity changes, 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 typically, 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[[30]](#footnote-30). 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 -- neurons and not blood vessels -- are the relevant networks to consider. To the extent that strong claims about autonomy deny any of this they seem overstated[[31]](#footnote-31).

For similar reasons, the claim that a network's topological structure "determines" its dynamics seems misleading or at least in need of qualification. If "dynamics" means what happens overall on a network then very often this is influenced but not determined by the network structure. And if "dynamics" has its more standard meaning having to do with the rules or equations describing the processes occurring along the network-- e.g., the equations describing how neural signals propagate or assumptions like those made in models of disease spread like the SI and SIR models-- these are, as we have seen, typically independent of assumptions about network structure.

**17. Directionality and Network Explanations**

Issues about the directionality of network explanations have elicited considerable discussion recently (e.g., Craver and Povich, 2019, Lange, 208, Kostic and Khalifa, 202 ). It is worth recalling that directionality issues originally arose in dicussions of causal explanation and the DN model. In a standard illustration, one can deduce the period T of a simple pendulum from the relation

(P) T= 2Π√l/g

and the values of l (length of the pendulum ) and g (gravitational acceleration) . (Call this the forward derivation) This counts as explanation according the DN model and this seems to accord with intuitive judgment. However, the "reverse" derivation of the value of l from (P) and g is also counted as an explanation by the DN model, contrary to the judgment of most. A common assessment (in my view correct but not very illuminating-- see below) goes like this: the forward derivation is acceptable because it conforms to the direction of causation-- l and g cause T. By contrast the reverse derivation runs contrary to the direction of causation: T is not a cause of l but rather one of its effects. The difference between the two derivations arises from the fact that we explain effects in terms of their causes rather than vice-versa.

A similar issue can be posed for network explanations. We think that the structure of a contact network helps to explain facts about disease spread but the latter (even in conjunction with dynamical assumptions) does not explain the structure of the contact network. The fact that the Koningsberg bridges lack the Eulerian feature explains (or explains in part) why they cannot be crossed exactly once-- call this NC. However, from NC one can also deduce that the bridges lack the Eulerian feature but NC does not explain facts about the connectivity structure of the bridges. 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 or derivation is explanatory when it runs in one direction but not in the other. Thus, in this case, one needs to find some other account of the directionality of such explanations-- this is the route taken by Kostic and Khalifa. 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.

I will not attempt to discuss all of the ideas in Kostic and Khalifa (2021) but will focus instead on just two of their suggestions about the directionality of network explanations[[32]](#footnote-32). First, recall their claim that a "topological explanation" is one that appeals to a "topological property" in its explanans. A derivation that appeals to network structure to explain some other property (NC above, the presence of a mental disorder) can thus count as a topological explanation. By contrast since NC is not a topological property, a putative explanation that appeals to TR will not count as a topological explanation. While this conclusion is of course unexceptionable given the way that "topological explanation" has been defined, in my view this argument doe not really address the underlying issue about directionality: even if we don't count the reverse derivation as a topological explanation, we still face the question of why we should not count it as an explanation of some other kind, and it seems clear that it is not. That is, we face the question of why we cannot appeal to facts about disease spread to explain facts about a contact network, even if we don't regard this as a topological explanation.

Another suggestion from Kostic and Khalifa is this:

Topological explanations must be answers to the relevant explanation- seeking question. An explanation exhibits *perspectival directionality* whenever its original satisfies this condition but its reversal does not. We submit that only the [ forward candidate explanation with a topological property in its explanans-- JW] *typically* satisfies the perspectival requirement.

By contrast a backwards candiate explanation in which the topologocal property is the explanandum *typically* does not satisfy the above condition:

in typical contexts in which [ the why question associated with the backwards candidate explanation] is asked, the relevance relation will likely *prohibit* any topological answers. If someone asked why exactly two of Kaliningrad’s [Koningsberg's] landmasses have an odd number of bridges, they would expect some retelling of the city’s history of urban planning, but would be sorely disappointed by a math lesson. So, insofar as we are concerned with typical contexts, perspectival directionality is achieved.

They add, though, that there may be very special contexts in which the reverse derivation *is* explanatory:

Against this backdrop, we think that defenders of topological explanation can safely bite the bullet: the reversal in this *very specific* interpretation of a *toy example* in a *highly idiosyncratic* context is explanatory. After all, nobody should be remotely surprised that atypical contexts call for atypical explanations.

Other things being equal, I think that it desirable not to appeal to contextual, or perspectival or other "pragmatic" considerations in providing an account of explanatory direction when an alternative that is not contextual is available. Moreover, for what it is worth, I doubt that there are any contexts in which, say, NC explains facts about the connectivity of Koningsberg's bridges. It counts against an account of explanatory directionality that it permits this possibility.

My own proposal about the source of directionality in network explanations (or at least many of them) 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. I claim they can be extended to capture the directional features of many network explanations, including both those that are causal and those that are arguably non-causal. I emphasize this to underscore the point that the suggestions that follow do not appeal to ad hoc considerations developed specifically to account for directionality in network explanations but rather to a set of ideas about directionality that have been successfully applied in other contexts, both philosophical and non-philosophical. In my opinion, these ideas deserve more serious attention from philosophers interested in network explanations than they have hitherto received.

The basic idea appeals to interventionist considerations: 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. There is also a third variable Z which is not correlated with X. 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 inteventions on Y do not change X, then of course if folows 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. 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) that such interventions on Z would also be associated with changes in X and, *ex hypothesi*, they are not. 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 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 equation (P) above but g (the third variable playing the role of Z above) is unaffected (assume we remain in a gravitational field of the same strength). 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 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 without changing the dynamics in a way that also influences E. 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 or from E to N to D or from E to D to N etc. 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.

I claimed earlier that it can be true that an intervention on N changes 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 non-interventionist counterfactuals of this latter sort (If E were such and such, N would be so and so) 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.

Arguably, the same account can apply even if we think of the explanation provided by a network as non-causal as long as there is some additional factor, independent of the network, which also affects the outcome of interest and long as both can be objects of intervention. We think that the network structure of the Koningsberg bridges and in particular the fact that this network does not have the Eulerian feature F explains the fact that the bridges cannot be traversed exactly once (OC) rather than OC explaining F. One way of capturing this asymmetry appeals to the fact that the example involves an implicit dynamics (a "walking dynamics") that is independent of F. In particular, there are ways of changing OC that do not involve changing the bridge configuration. As a somewhat fanciful possibility we might imagine that all the pedestrians are equipped with jet packs that allow them to leap over canals without walking over bridges. This alters whether it is possible to cross each bridge exactly once but will not alter the bridge structure[[33]](#footnote-33). On the other hand if we were to intervene to alter the network structure of the bridges so that there were zero or two nodes of odd degree while keeping the dynamics the same (it still involves walking), this would alter OC, in accord with our expectation that the network structure explains OC rather than vice-versa[[34]](#footnote-34).

In a discussion of previous comments of mine about this example, Lange and Elliott (2022) claim that interventionism lacks the resources to distinguish cases in which the explanatory direction runs from F (again, whether the bridge configuration has the Eulerian feature) to OC from cases in which the direction runs from OC to F. As I understand it, their argument goes as follows: Suppose that one carries out what one supposes is an intervention I (e.g via a construction project that alters or removes bridges) on the bridges that changes F and observes a change in OC, so that one is tempted to think that the explanatory direction is I--> F--> OC with F explaining OC. (Think of I as representing the intentions of the builders to realize one structure rather than another). Elliott and Lange claim that since this intervention also changes OC, one is equally justified in thinking of the situation as one in which the intervention is, so to speak, directly on OC which in turn results in a change in F: I--> OC-->F. This argument overlooks the role of the factor D (the dynamics) which is independent F but also relevant to OC. It is this that allows us to break the supposed symmetry between F--> OC and OC--> F. If, in the scenario above, I--> OC--> F is the correct direction, it should be possible to intervene on OC by changing D in a way that is independent of I (builder's intentions) and under some such intervention F should change (since it is assumed that OC causes F) even if I remains unchanged. (That is, bridges are not changed while D is changed in a way that affects OC). Obviously F will not change under this intervention. And if we keep D unchanged but appropriately change F (via the builders' intentions), OC will change. This shows that I--> F--> OC is the correct account of the explantory direction.

**18. Networks and Causal Graphs Compared.**

The networks considered in this paper have been, for the most part, undirected graphs. These contrast in a number of respects with causal graphs which, as their name implies, are used to represent causal relationships and are always directed, even if they are cyclic. Some of the differences between these two sorts of representations have been mentioned above in passing. Here I want to make the contrast more explicit-- doing so will do more to bring out what is distinctive about undirected networks. When a direct graph is used to represent causal relationships, the nodes/vertices in the graph represent *variables* (income, years of schooling etc., the concentration of a protein) which can take different values. An edge from variable X to variable Y represents that X is a direct cause of Y. This implies, in a standard interventionist interpretation, that there is a possible intervention that changes the value of X and which is such that if all other variables in the graph are held fixed at some value, also by independent interventions, the value of Y would change. It is also assumed that each variable Y is a specific deterministic function of the full set of its direct causes (its "parents) and possibly an additional "error term" which is assumed to be uncorrelated with the other parents of Y[[35]](#footnote-35). Moreover in many applications, the causal graph is assumed to be accompanied by a probability distribution over the the vertices in the graph which is related to the graph structure in a specific way, via the Causal Markov Condition. Needless to say, none of these asssumptions are required or even usually made when a graph with undirected edges is employed. As noted above, in an undirected graph, the nodes often represent things or objects--- individual people or brain areas (which don't have "values" in any literal sense) -- rather than variables. Moreover although an edge from X to Y represents the possibility of influence there need be no suggestion that this is characterized by a specific deterministic function. In addition causal graphs are assumed to behave in a characteristic way under interventions: an intervention on a variable X is understood to "break" or "delete" all of the edges directed into X while preserving all other edges in the graph. Put differently, the causal structure represented by the graph is assumed to be "modular" in the sense that in principle one can break edges into one variable without altering what holds elsewhere in the graph -- this seems to be required if the graph is to have a full interventionist interpretation. As noted above, while similar modularity assumptions are sometimes made in connected with undirected graphs (e.g., in explanations of robustness) they are not required and by no means always warranted. Finally, there is another difference. As we have seen, in a causal graph, the notion of intervention on a *node* is well-defined and has a characteristic effect in terms of breaking *all* edges directed into that node. However, in this context, the notion of intervening to break an individual edge (among several that may be directed into a variable while leaving the other edges intact) is not defined. By contrast, in at least some cases-- particularly in what I have called independent existence cases-- the notion of intervening to remove or change a particular edge as well as a node seems well defined: if there are several structural neural connections linking neural area A to neural area B, it may be possible to remove one of these connections while leaving the others intact. Moreover, when undirected graphs are employed, we seem to have a distinct notion of removing a node that has no counterpart in a standard causal graph. Given an undirected graph in which brain areas A and B are connected by an edge, it will typically be possible to remove or ablate A itself. It seems this operation will necessarily remove *all* edges between A and other brain areas, in contrast to just removing edges directed *into* the node intervened on, as is the case with a causal graph.

Although I have been critical of the idea that explanations involving networks with undirected edges are entirely free from causal committments, the considerations just described show that such networks are in important respects very different from ordinary causal graphs. This is so even though undirected networks can provide explanations-- even causal explanations. Those who emphasize the distinctivenes of such network explanations are not wrong, even though, in my view, the nature of this distinctiveness is not always adequately characterized.

I conclude this section with another observation. In many typical uses of causal graphs what is of interest is the discovery and representation of particular causal connections-- e.g., whether variable X causes Y. But it is certainly possible to focus instead on overall features of the connectivity of a causal graph and their consequences, rather than individual links. It makes sense to do this when some outcome of interest depends more on the overall structure of a causal graph than on any individual link . For example, in a bowtie structure (Alon, 2006), many directed edges converge on a central node from which other edges diverge. The edges often have a straightforward causal interpretation but for some purposes what matters is not so much the causal details associated with the individual edges but the role the the central node has in virtue of its overall position in the graph. In particular, a consequence of this position is that disruption of the central node will disrupt all of the causal processes and interactions that are funneled through it, so that this structural feature explains why the system is not robust to such interventions. Note that this explanation can be correct even if one has mischaracterized (or lacks information about) some of the particular edges directed into or out of the central node as long as the overall architecture has a bow tie structure. Here it certainly matters that the graph is directed and at least in many cases it also matters that the edges have a causal interpretation, so it is not as though the explanation provided is lacks causal committments[[36]](#footnote-36). But at the same time, the explanation draws on facts about the overall architecture or connectivity of the graph in the way in which other network explanations do.

**19. When are Network Explanations Likely to be Most Fruitful?**

I began this essay by noting that networks are most useful when dealing with systems with a number of components which are restricted in how they can influence each other, as opposed to systems with components each of which can interact with every other. In this section I draw together some strands in my earlier discussion to describe other system features that are relevant to the successful use of networks.

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, 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. More ambitiously we may also ask whether, given a network structure, the same outcomes will follow for some range of different dynamics.

2. Given such a networks/dynamics separation it may happen that we have ways of discovering network structure that require only minimal assumptions about dynamics -- we 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. Systems for which such independent identification of network structure is possible are likely to be particularly suitable for network analysis.

3. The use of a network to represent constraints on possible interactions 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. 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 area A and area 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.

5) 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 etc. 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.

6) To the extent that the abstraction from detail described in 3) - 5) 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**

# Alon, U. (2006) *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Boca- Raton, FL: Taylor and Francis.

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". *Philosopy of Science*; 573–593.

Borsboom. D. (2017) "A Network Theory of Mental Disorders" *World Psychiatry* 16:5-13.

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.

Lange, M. (2018) "A reply to Craver and Povich on the directionality of distinctively mathematical explanations" *Studies in the History and Philosophy of Science* 67: 85-88.

Elliott, K. and Lange, M. (2022) "Running it up the flagpole to see if anyone salutes" *Theoria* 37: 53-62.

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.

1. \* Many thanks to Lauren Ross for very helpful comments on an earier draft [↑](#footnote-ref-1)
2. "Process" is not meant to suggest anything metaphysical or something necessarily meeting the requirements of a "process theory" of causation. [↑](#footnote-ref-2)
3. My notion of constraint here is very similar to the notion discussed by Ross in her 2023. See Section 11. 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-3)
4. 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-4)
5. 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 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 to is required. [↑](#footnote-ref-5)
6. Thus, although network connectivity relationships are commonly described as "topological" one should be careful about how this is understood-- "topological" should not be understood as meaning "purely spatial". In part for this reason I prefer the term "network explanation". [↑](#footnote-ref-6)
7. I'm aware of course that a number of philosophers have claimed that network explanations are sharply distinct from causal explanations and at least in some important respects entirely free of causal commitments -- they are "stand alone" non-causal explanations (see, e.g., Kostić, 2018, 2020, 2022.) This issue is discussed in detail below but for now I'll just say that I think such claims are often misleading. I suspect that the fact that the construction of a network requires only weak causal assumptions is part of what has suggested "stand alone" claims to some some. Another consideration is when a dynamics is assumed in a network explanation this is often rather generic and the focus is on what difference network structure makes to outcomes, assuming that dynamics. Let me also emphasize that while I see networks as describing possibilities of interaction, the kinds of networks under discussion (with undirected edges) do not by themselves provide causal explanations and an edge connecting X and Y should *not* be interpreted as the claim that X causes Y. An edge from Jones to Smith in a contact network does not tell us that Smith's infection was caused by contact with Jones, but merely that this is possible if Jones is infected. In these respects the networks under discussion are very different from the standard directed graphs or causal Bayes nets that are used to represent causal relations. [↑](#footnote-ref-7)
8. Rathkopf (2018) is an important exception. [↑](#footnote-ref-8)
9. 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-9)
10. 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. 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-10)
11. 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-11)
12. That the explanandum in this example is an impossibility is also emphasized in Lange, 2018 and Woodward, 2018. [↑](#footnote-ref-12)
13. There are a number of different measures of node centrality or importance. One commonly used measure is eigenvector centrality which measures (roughly) the extent to which a node is connected to other nodes that are also highly connected but there are other possible measures which focus on instead on such features as the extent to which deletion of a node results in a disconnected graph. In general the appropriateness of a centrality measure will depend both on what one is trying to measure and empirical assumptions about the nature of the flow or transfer along the network. This is an additional reason why one cannot read off the emprical results of node removal just from formal measures defined from the structure of a graph. [↑](#footnote-ref-13)
14. A number of researchers (e.g., Borsboom, 2017) have constructed models of mental illnesses as networks of symptoms and have then explored using these to infer what the results of intervening to alter symptoms might be. The general consensus is that, when understood as dynamical systems, the response of such systems to interventions can be difficult to predict just on the basis of network structure-- a more systematic analysis in terms of control theory is required. [↑](#footnote-ref-14)
15. 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, spread of something, interactions among nodes etc. These are not assumptions about the network structure itself. [↑](#footnote-ref-15)
16. "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-16)
17. Rathkopf (2018) emphasizes this. [↑](#footnote-ref-17)
18. See Friston, 2011. [↑](#footnote-ref-18)
19. Note that although this constraint is represented as a geometric-looking relation what it is really encoding is the presence of unspecified forces that restrict the particle to the surface of the sphere. This is not a case in which a purely geometrical or spatial relation explains. A similar point holds for many of the constraints represented by network structure. [↑](#footnote-ref-19)
20. Note that it is not the case that these details are causally irrelevant simpliciter -- rather they are irrelevant, *conditional* on the coarse-grained summary represented by the presence or absence of an edge, to the system behavor we are interested in explaining. For more on this conditional irrelevance idea, see Woodward, 2021b. [↑](#footnote-ref-20)
21. For example, a disease will only spread through a compartment if the starting state is one in which at least one individual in the apartment is infected. [↑](#footnote-ref-21)
22. If intervening to change the 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-22)
23. For additional discussion of this example and non-causal explanation more generally, see Woodward, 2018. [↑](#footnote-ref-23)
24. Suppose that in a contact network there is an undirected edge between the node representing individual 1 and the node representing individual 2. If both 1 and 2 are infected, we don't know whether the infection of 1 explains the infection of 2 or vice versa or whether the infection is caused by some third infected individual. Temporal information may help (e.g., we observe 1 infected before 2) but this is information in addition to that provided by the network. Moreover, it is unlikely to be helpful in every case. [↑](#footnote-ref-24)
25. Huneman 2010 , quoted in Ross, 2021. [↑](#footnote-ref-25)
26. One might think that the real causal actors here are the factors responsible for the pattern of effective connectivity-- the structural connections, the neural transmitters, properties of neural signals etc-- and the choices, behavior etc. to which they contribute that in turn influence life outcomes. This would avoid attributing causal efficacy to the overall pattern of effective connectivity itself. [↑](#footnote-ref-26)
27. Again, this is not to say that we can't distinguish between the network and the dynamics occuring on it. We can know that certain kinds of neural connections are necessary for the propagation of neural signals without knowing much about the dynamics governing those signals. [↑](#footnote-ref-27)
28. 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. In other cases, the point seems even clearer. The edges in a network representing friendship relations do not derive in any straightforward way from the spatial relations among people but rather reflect whether they interact in some appropriate way. [↑](#footnote-ref-28)
29. Similar claims are made in Huneman, 2010and Jones, 2014. Huneman, 2010 writes, "... there is a kind of explanation that relies upon “topological” properties of systems in order to derive the explanandum as a consequence, and which does not consider mechanisms or causal processes".

    [↑](#footnote-ref-29)
30. Here I agree with Craver, (2016). [↑](#footnote-ref-30)
31. These remarks may provoke a response along the following lines: It is true that in constructing a network we may need to make use of causal information of various sorts-- e.g., about what can influence what or about how a network may respond to perturbations. But, it may be said, we need to distinguish the role played by such information from what, properly speaking, is part of the network explanation itself. In particular when we explain using a network, the causal information just described should be treated as part of the background to the network explanation or what is presupposed or assumed rather than as a component of that explanation. I find this use of the "part of the explanation" vs. "background" opaque. Those who insist on this distinction are, I suspect, operating in accord with something like the following methodology. The philosopher proceeds by contemplating various candidate examples of explanation, which are assessed according to whether they produce a feeling of explanatoriness or not. If they do, the philospher then asks what in the example seems (introspectively) seems to be responsble for this feeling. This is then taken to be what belongs to the explanation per se-- everything else is background or presupposed or taken to be not a working part of the expalnation. One obvious problem is that this way of proceeding is that it can be subjective and arbitrary-- people won't always agree about what is making the explanitoriness light bulb go off in them and what is instead mere background. But a deeper problem is that it limits the scope of the project of developing models of explanation in an unhelpful way. A satisfactory philosophical treatment of some form of explanation should, among other desiderata, elucidate the information that is needed to construct and evaluate the explanation and the circumstances in which the explanation will be successful or not. For these purposes, identifying the relevant background information or what is presupposed in applying the explanation can be highly relevant and needs to be included in an account of how the explanation works. [↑](#footnote-ref-31)
32. I will also not discuss the ideas about the source of directionality in examples like the Koningsberg bridge problem in Craver and Povich, 2017 or Lange, 2018. My view is that the solution proposed here is clearer and more general. [↑](#footnote-ref-32)
33. This goes to the issue, raised above, of what the network structure in this case represents. If the network structure is just represents the bridges, then their once traversabiity OC can change independently of this structure by endowing the citizens of Koningsberg with new capabilities. On the other hand, the network might be understood to represent something like the possibility of traversing by ordinary walking via a continous path. Now the walking dynamics is in this respect built into assumptions about the network. [↑](#footnote-ref-33)
34. As another but related possibility, if one is willing to assume that the explanatory structure involving F, W = whether the walking dynamics holds and OC satisfies the faithfulness condition imposed in causal modeling (even if the relationship among these variables is not causal), then, assuming that F and W are statistically independent, and F and OC and W and OC are dependent but F and W are dependent conditonal on OC (all of which are plausible), it follows that the only possible structure is one in which F and W explain OC. [↑](#footnote-ref-34)
35. When a directed graph is used to represent causal relations, the graph does not tell you what this function is, but it is assumed that there *is* such a function. [↑](#footnote-ref-35)
36. See Ross, 2021 for a similar conclusion. We wouldn't be dealing with a bowtie structure if the edges were undirected or all the edges were directed out of the central node. [↑](#footnote-ref-36)