Climbing the ladder of causation

Casper van Elteren

Radboud University
caspervanelteren@gmail.com

October 4, 2018

Introduction

One of the principle aims of science is to provide causal explanation to (natural) phenomena[1, 2]. For example, we wish to understand why healthy cells can transform into cancerous cells, how populations of neurons produce cognition, what causes a financial market to crash, or how social interaction influences opinion consensus. These systems are typically characterized by many dynamical variables, non-equilibrium, non-linear, and time-dependent behavior, defying analysis by most of our current theories and tools. At the same time, exactly these systems are vitally important to the functioning of our bodies, society, and environment. Their resistance to reductionistic analysis has earned them the qualification of 'complex (adaptive) systems' and spawned the cutting edge field of complexity science aiming to understand the inner-mechanism of these systems[3, 4].

[5] Obtaining causal explanations is a multi-step process. Pearl identified three levels of causation[6]. The first step in understanding any problem is mapping regularities in the environment. This level is where science starts and where it is the state of the art for the many supervised machine learning methods specifically neural networks including deep learning. The second level entails predicting the effects of deliberate alterations of the environment. It focuses on 'what-if' scenarios, e.g. what will happen if we change the environment. It is important to emphasize that the first level will never be enough for causal explanations. No matter how large the dataset is, passive observation will only yield correlative associations where cause and effect cannot be disentangled. The final step involves constructing theories surrounding 'why' questions, e.g. questions such as 'Was it X that caused Y?' or 'What if X had not occured?'. That is, the final level op causation looks to the past and is able to integrate information above-and-beyond the second level. It allows for answering of counterfactuals through control and manipulation and is crucial for understanding the problems outlined above.

Reaching the final level of causation is not trivial however. Current available theories and algorithms for inferring the causal influence of dynamic variables are not readily applicable to complex systems. Four major issues can be identified from literature. First, many theories and algorithms are developed using at least one of the following assumptions regarding system behavior, e.g. stationarity of system state, (local) linearity of dynamics[7, 8], time-independence of the interaction, and/or having reached equilibrium. Second, techniques differ in strength of intervention. Many techniques are based on the 'overwhelming' control intervention [7], e.g. completely knocking out a gene from a cell's regulatory process, or replacing its signal altogether. Especially in complex systems such dramatic interventions may completely change the emergent system behavior or lead to undesirable side-effects and should therefore be avoided. Third, identification of causally important parts of a system, i.e. dynamically important for the systemic behavior, has long been studied from the viewpoint of topology[9–14]. This viewpoint is often based on concepts of 'flow'; the more outward connections a node in a system has, the more opportunities it has for spreading its information. For example, (structural) control theory identifies sets of driver-nodes based on their coverage of the edges in a graph. However, it is well-know that topology interacts with dynamics[9, 11, 15]. A recent study showed that the same static topoloy dictated by different dynamics yield different node influence [9]. This means that as the dynamic on a network structure changes, so does the dynamic importance of that node. Therefore, identifying possible causal driver-nodes cannot reliably be determined based on topology measures alone. Additionally, obtaining control using a topology based method may not be feasible on real-world network, e.g. it may require extraneous long time [14] or require abnormally large interventions[11, 16]. Lastly, the asymmetry of causal influence is not always accounted for. The rules of algebra allow to write relations between variables in a variety of syntatic forms. To illustrate consider Newton's second law of motion $f = ma$. This equation can be written in different ways, however it is common to think that force *causes* acceleration - not that acceleration causes force. The causal asymmetry of cause and effect cannot be inferred from a symmetrical equation. This lack of asymmetry is a problem for 'unification' accounts in mathematics and physics where the goal is to find one formula that could explain or account for different natural phenomena such as quantum particle motion, gravity. Providing causal explanations inherently requires scientists to embed the analysis in an asymmetrical environment using control and manipulation.

In this thesis, we aim to further develop and validate a novel algorithm for inferring the primary causal driver nodes of a complex system[17]. Previous research showed that it was not the high-degree nodes in scale-free artificial networks that were dynamically most important, rather the intermediately-connected nodes were proofed to be dynamically more salient. Here, we specifically focus on determining if topological features are a good predictor for the primary driver-node in a real-world signed network. Additionally, the aim is to expand the proposed methods through a framework of control and manipulation(I).

At its core the method consists of information-theoretic measures, specifically a time-delayed version of Shannon mutual information (MI)[18] which are computed based on cross-section of time-series of each dynamic variable in the system with the entire system. Information theoretical methods have the advantage of being suitable for non-equilibrium systems, and non-linear, time-dependent interactions. Additionally, they can be readily applied to different data-types such as categorical, discrete and continuous data. Furthermore, the technique will be a good predictor for specifically small interventions and not overwhelming interventions, making it suitable for sensitive control of complex systems while staying as close as possible to their natural functioning.

Similar approaches have been developed using other information theoretical based approaches such as transfer entropy [19], conditional MI [20], relative entropy[21] or information bottleneck[22]. However, these methods are not suitable for two reasons. Firstly, the question that these techniques attempt tot answer is qualitatively different from our proposed measure. Their main aim is to reconstruct the complete network of causal interactions by evaluating the (short-term) causal influence between every pair of nodes. In contrast, the emphasis here is on finding out whether a give-node has a long-term causal effect on the entire system - *regardless of where in the system this effect resides*. Second, they are ultimately based on the conditional MI measure which is known to either overestimate or underestimate the influence of so-called synergistic and redundant information[23]. The used methods here circumvents this issue by using only the non-conditional MI measure which does not have this problem. This has the disadvantage that only a handful (independent) driver-nodes can be identified instead of the full causal network. Regardless, this body of work will already provide essential innovation towards the management of complex systems and possible research avenue for climbing the ladder of causation.

I. Material and Methods

i. Information decay time

The goal is to identify driver-nodes in the system. A driver-node is defined as a node that has the most dynamic impact on the system over time, i.e. it will 'drive' the system. Here we identify the most impactful node not based on topological features, but rather on dynamic properties by using concepts from information theory [17].

A system $G = (V, E, D)$ is defined as set of nodes or vertices $v_1, v_2, \cdots, v_k \in V$ and edges $(v_i, v_j) \in E$ between which interactions can take place. Each node in the system is dictated by dynamics D. The state of node v_i at time t is defined as $v_i(t)$ and the state of the system (or 'snapshot') is the collection of node states $V^t = \{v_1(t), v_2(t), \cdots v_k(t)\}$. Each node in the network chooses its next state based on the current state and the state of each of its nearest-neighbors, i.e. node $v_i(t)$ chooses its next state x with the conditional probability distribution $p(v_i(t+1) = x|h_i(t))$ where $h_i(t) = \{x : (v_i(t), x) \in E\}$ are the nearest-neighbors of node i. This is also known as a Markov-network or memory-less dynamics.

The dynamic importance of a node v_i is quantified by the information decay time (IDT) [17]. The IDT is the time δ it takes for the information about the state of node s to disappear from the network state. Conversely, it is the time it takes for the network as a whole to forget a particular state of a single node. We quantify 'forgetting' in terms of the decay of *bits of information* using a time-delayed Shannon mutual information. A bit of information is conceptualized as the minimum required yes-no questions to determine the outcome of a stochastic variable X and is quantified using entropy[24]:

$$H(X) = -\sum_x p(x) \log p(x) \tag{1}$$

Please note that all log are base 2 in this paper. Entropy is often conceptualized as the amount of 'surprise' of a stochastic variable. To illustrate, consider a far coin. It requires exactly one yes-no question to determine the state of the coin after a flip, e.g. was the facing side heads or conversely tails? Hence, the fair coin can deliver 1 full bit of information. In contrast, if the coin was not fair, the conveyed information would be less because of the reduction in surprise[18].

Mutual information can be thought of as the reduction of uncertainty about one random variable given the knowledge of another random variable, i.e. $I(X;Y) = H(X) - H(X|Y)$, where $H(X)$ and $H(X|Y)$ indicates the entropy and conditional entropy respectively. The conditional entropy is similar to the entropy except that it takes into account the effect of a second variable, specifically how much uncertainty is left after taking into account another variable. High mutual information indicates a large reduction in uncertainty, and zero mutual information between two random variables means that the variables are independent.

ii. Defining information dissipation of a node

Up until now, we haven't addressed how we can ensure that the information of a node will be lost as a function of time. This can be ensured using the data processing inequality. This concept from information theory entails that the information content of a signal cannot be increased through physical operation in a Markov chain; information is generally lost, not gained when sent through a noisy channel, see **??** for proof.

At every time step every node can partially transmit its information to its nearest neighbors. In order to compute the IDT we will need a time-delayed mutual information which is defined as: The state $S(0) = X$ indicates a random state from the state space of the system, i.e. a specific 'snapshot' or configuration of the system, and $s_i(t)$ indicates the state of node a node at some time t from snapshot $S(0)$, please see vi.

Here we consider Markovian networks in which each node determines its next state according to the Gibbs measure. The Gibbs measure describes how a unit changes its state subject to the combines potential of its interacting neighbors (see v). We define the IDT for node s_i as : where α can be chosen between $[0, 1)$. In Markov networks the data processing inequality ensures that the mutual information will decay as $t \to \infty$. Data processing inequality is an information theoretical concept which states that the information through a noisy channel cannot be increased through local operation, see appendix **??** for proof.

iii. Network characteristics

In graph theory and network analysis centrality measures are often used to describe properties of network topology. Formally, a network topology is depicted as graph G; an order structure $G = (V, E)$ consisting of a set of vertices or nodes V and a set of edges E that connects these nodes. Centrality measures are used to answer the question 'what characterizes an important node'. The answer is provided by a real-valued function on the nodes in a graph where the the values provide a ranking which identifies the most important nodes [25, 26]. What quantifies 'important' depends on the question being asked. For example one might be interested in answering 'What nodes is closest to any other node?' or 'Does a node with high degree connect often to other high degree nodes?'. Here we will use four of the most commonly used centrality measures to will answer a diverse set of questions.

Closeness centrality

In connected graphs, closeness centrality of a node is a measure of centrality of a node. It is calculated as the reciprocal sum of the shortest path of a node between all other nodes in a network; the more central a node is, the closer it is to other nodes. Closeness of node v is defined as [**Bavalas1950**]:

Betweenness centrality

Eigencentrality

Weighted degree centrality

Prior research noted that nearly all centrality measures in fact decompose the walk structure of a graph. A walk is a sequence of alternating node and edge such as $v_0, e_1, v_1, \cdots e_k, v_k$[25, 26]. Bogatti and colleagues identified four dimensions all centrality measure decompose the structure of a walk[26]; type of summary measure (sums or averages), the type of walks considered (shortest path, paths, trail or walk), the dimension of the walk (frequency of occurrence or length of walk) and the involvement of walk dimension (radial or medial).

iv. Data

The data originates from the Changing Lives of Older Couples (CLOC) and compared depressive symptomology assessed via 11-item Center for Epidemologic Studies Depression Scale (CES-D) among those who lost their partner (N=241) with still-married control group (N=274) [27]. Each of the CES-D items were binarized with the aid of a causal search algorithm using Ising spin model developed by [28] and represented as a node with weighted connections (1). For more info on the procedure please see [27–29]. The 11 CES-D items are (abbreviated names used in the remainder of this text in brackets): 'I felt depressed' (depr), 'I felt that everything I did was an effort' (effort), 'My sleep was restless' (sleep), 'I was happy' (happy)', 'I felt lonely' (lonely), 'People were unfriendly' (unfr), 'I enjoyed life' (enjoy), 'My appetite was poor' (appet), 'I felt sad' (sad), 'I felt that people disliked me' (dislike), and 'I could not get going' (getgo).

Although inferring the exact network of causal relations is difficult, the main point of this thesis is to show that the topological features are not predictive for the node with the highest IDT. The results from this study may provide insights in what symptoms are more dynamically important than others. This could lead to novel treatment methods for psychia-

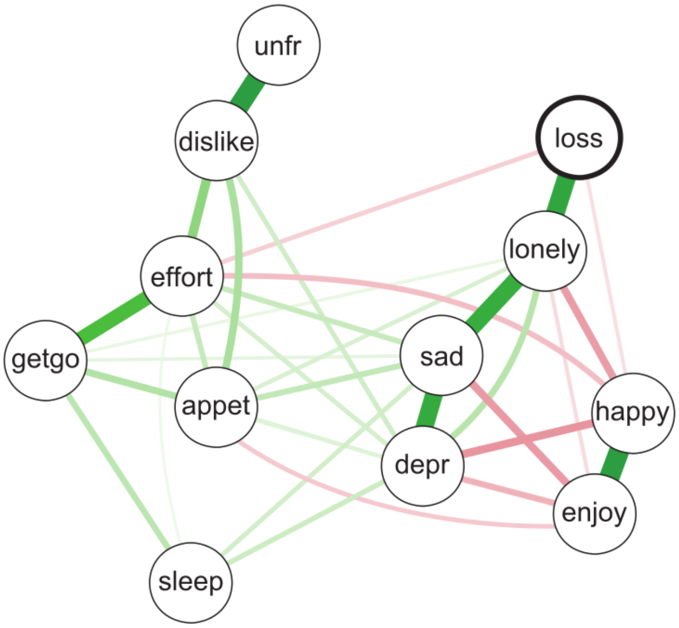

Figure 1: *Psycho-symptom network obtained from [29]. The thickness of the line indicates the strength of the connection. Green lines indicate positive weights, red lines indicate negative weights.*

trist and psychologist alike to target symptoms that have the highest impact. Since the methods presented here do not indicate whether the impact is positive or negative for the patient, future studies will need to validate whether these simulations contributed meaningful insights for improving the patient's quality of life.

v. Model

Here we consider a real-signed network in which each unit is dictated by the Gibbs measure:

$$p(s_i = x) = \frac{1}{Z(\beta)} \exp(-H(s_i)) \qquad (2)$$

where $H(s_i)$ is the Hamiltonian function;

$$H(s_i) = -\sum_{<ij>} J_{ij} s_i s_j - \sum_j h_j s_j$$

The sum j is over the nearest neighors of node i The Gibbs measure is a concept from statistical mechanics providing a probablistic description of ensemble of particles [**Gibbs1996**].

vi. Numerical methods

In order to compute **??**, we will estimate $P(x)$ and $P(x|X)$ using Markov chain Monte-Carlo sampling. The core procedure for a pre-defined temperature involved obtaining $N = 10000$ snapshots based on a single Markov-Chain. Starting from a random state to eliminate, the system was equilibrated using $k = 100000$. Next, using the state distribution, $k = 1000$ independent simulations were repeated for $\delta = 20$ time steps each to construct $p(x|X)$. Each simulation step consisted of asynchronous updating. Each simulation step used asynchronous updating rule; every simulation step every node in the system was considered for a flip in random order once.

Through regression with a double exponential $y = a + b\exp(-cx) + d\exp(-ex)$, we were able to estimate the decay times for each node for an α.

Previous research indicated that there is an interaction with temperature and IDT[17]. Hence, in order to see the effect of noise on the symptoms we performed the procedure above for 6 temperatures. The temperatures were based on the 'willingness' of the system to stay magnetized (2). Namely, the system was set to full magnetization, i.e. all nodes had state -1 or $+1$, and the system was ran for $N = 10000$ steps. Temperatures were chosen between $20 - 80$ percent of the max magnetization (2) and as such provide a nice cross-section of the effect of noise on the system.

The code used in this paper is build from the ground up in python 3.6 and can be readily adapted for use with other models, please see the appendix B for more information.

II. Results

III. Discussion

References

1. Woodward, J. *Making things happen: A theory of causal explanation* ISBN: 2002192596 (2005).
2. Pearl, J. Causality Second Edition, 1–386 (2000).
3. Mitchell, M. An Introduction to Genetic Algorithms (Complex Adaptive Systems). *The MIT Press,* 221. ISSN: 08981221 (1998).
4. Sayama, H. *Introduction to the Modeling and Analysis of Complex Systems* ISBN: 9788578110796. doi:10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3 (2015).
5. Aaronson, S. Why Philosophers Should Care About Computational Complexity. *In Computability: Gödel, Turing, Church, and beyond (. . .* 58. ISSN: 1433-8092 (2011).
6. Pearl, J. *et al. The Book of Why* ISBN: 9780465097616 ().
7. Chen, G. Pinning control and controllability of complex dynamical networks. *International Journal of Automation and Computing* **14.** ISSN: 17518520. doi:10.1007/s11633-016-1052-9 (2017).
8. Liu, Y. Y. & Barabási, A. L. Control principles of complex systems. *Reviews of Modern Physics* **88,** 1–61. ISSN: 15390756 (2016).
9. Harush, U. & Barzel, B. Dynamic patterns of information flow in complex networks. *Nature Communications* **8,** 1–11. ISSN: 20411723 (2017).
10. Šikić, M., Lančić, A., Antulov-Fantulin, N. & Štefančić, H. Epidemic centrality - Is there an underestimated epidemic impact of network peripheral nodes? *European Physical Journal B* **86,** 1–23. ISSN: 14346028 (2013).
11. Cowan, N. J., Chastain, E. J., Vilhena, D. A., Freudenberg, J. S. & Bergstrom, C. T. Nodal dynamics, not degree distributions, determine the structural controllability of complex networks. *PLoS ONE* **7.** ISSN: 19326203. doi:10.1371/journal.pone.0038398. arXiv: 1106.2573 (2012).
12. Muldoon, S. F. *et al.* Stimulation-Based Control of Dynamic Brain Networks. *PLoS Computational Biology* **12.** ISSN: 15537358. doi:10.1371/journal.pcbi.1005076. arXiv: 1601.00987 (2016).
13. Yan, G. *et al.* Network control principles predict neuron function in the Caenorhabditis elegans connectome. *Nature* **550,** 519–523. ISSN: 14764687 (2017).
14. Pequito, S., Preciado, V. M., Barabási, A.-L. & Pappas, G. J. Trade-offs between driving nodes and time-to-control in complex networks. *Scientific Reports* **7,** 39978. ISSN: 2045-2322 (Jan. 2017).
15. Gates, A. J. & Rocha, L. M. Control of complex networks requires both structure and dynamics. *Scientific Reports* **6,** 1–11. ISSN: 20452322 (2016).
16. Yan, G., Ren, J., Lai, Y. C., Lai, C. H. & Li, B. Controlling complex networks: How much energy is needed? *Physical Review Letters* **108.** ISSN: 00319007. doi:10.1103/PhysRevLett.108.218703. arXiv: 1204.2401 (2012).

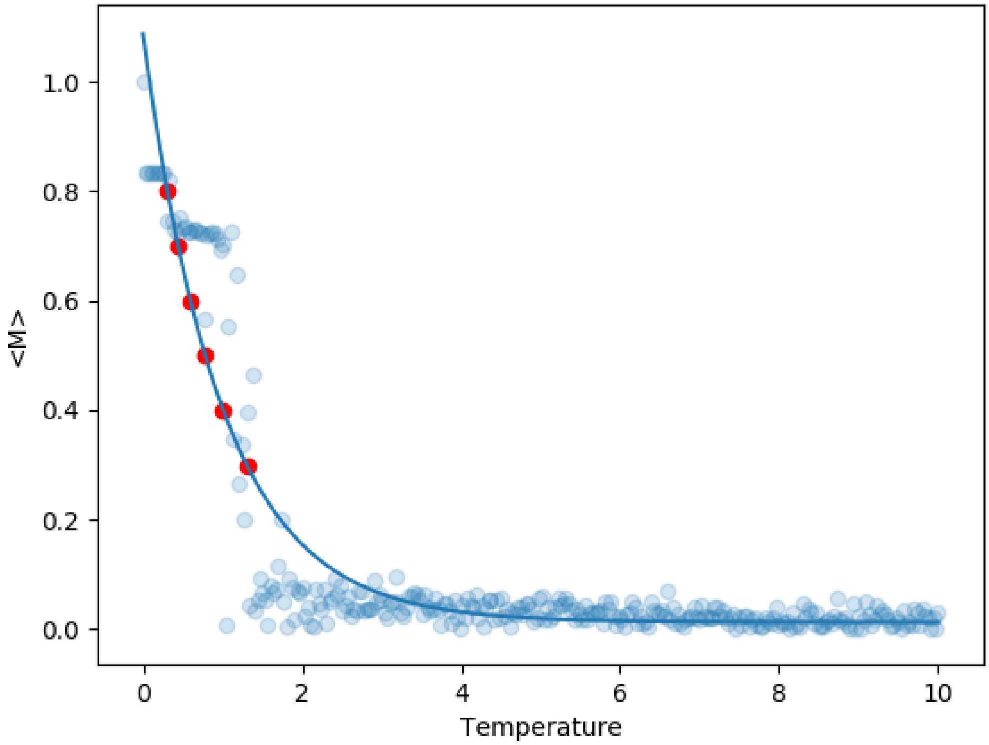

Figure 2: *Absolute magnetization as a function of temperature. The system starts from full magnetization and is simulated for 100 000 steps. The red dots indicate the magnetization between 80-20 percent of the maximum magnetization found and used for further simulations.*

17. Quax, R., Apolloni, A. & Sloot, P. M. a. The diminishing role of hubs in dynamical processes on complex networks. *Journal of the Royal Society, Interface / the Royal Society* **10,** 20130568. ISSN: 1742-5662 (2013).
18. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* 1–748. ISBN: 9780471241959. doi:`10.1002/047174882X`. arXiv: `ISBN0-471-06259-6` (2005).
19. Eichler, M. On the evaluation of information flow in multivariate systems by the directed transfer function. *Biological cybernetics* **94,** 469–82. ISSN: 0340-1200 (June 2006).
20. Wibral, M., Editors, J. T. L. & Kelso, S. *Directed Information Measures in Neuroscience* ISBN: 978-3-642-54473-6. doi:`10.1007/978-3-642-54474-3` (2014).
21. Ay, N. & Polani, D. Information Flows in Causal Networks. *Advances in Complex Systems* **11,** 17–41. ISSN: 0219-5259 (2008).
22. Tishby, N., Pereira, F. C. & Bialek, W. The information bottleneck method, 1–16. ISSN: 0305-5728 (Apr. 2000).
23. James, R. G., Barnett, N. & Crutchfield, J. P. Information Flows? A Critique of Transfer Entropies. *Physical Review Letters* **116,** 1–6. ISSN: 10797114 (2016).
24. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27,** 379–423. ISSN: 07246811 (1948).

25. Borgatti, S. P. Centrality and network flow. *Social Networks* **27,** 55–71. ISSN: 03788733 (2005).

26. Borgatti, S. P. & Everett, M. G. A Graph-theoretic perspective on centrality. *Social Networks* **28,** 466–484. ISSN: 03788733 (2006).

27. Fried, E. I. *et al.* From loss to loneliness: The relationship between bereavement and depressive symptoms. *Journal of Abnormal Psychology* **124,** 256–265. ISSN: 19391846 (2015).

28. Van Borkulo, C. D. *et al.* A new method for constructing networks from binary data. *Scientific Reports* **4,** 1–10. ISSN: 20452322 (2014).

29. Epskamp, S., Kruis, J. & Marsman, M. Estimating psychopathological networks: Be careful what you wish for. *PLoS ONE* **12,** 1–13. ISSN: 19326203 (2017).

A. Data processing inequality

The data processing inequality shows that no manipulation of the data through a Markov chain can increase the inferences made from that data.

Consider a Markov chain of random variables X, Y, Z:

$$X \to Y \to Z$$

where $X \perp Z$.

B. Code Manual

C. Design philosophy

The main goal was to provide a package that is readily accessible and easy to adopt for use with other types of models. It is based on separation, speed, and expandability. It consists of an engine; the code used to compute the IDT and related information-theoretical measures, and methods to perform monte-carlo simulations. Other modules will hook into the engine including a model framework, basic input-output operations, see **??** for overview.

D. Simple examples

Simple ipynb? Simple graphs