A peer-reviewed version of this preprint was published in PeerJ on 7 June 2016.

View the peer-reviewed version (peerj.com/articles/2111), which is the preferred citable publication unless you specifically need to cite this preprint.

Zanin M. 2016. On causality of extreme events. PeerJ 4:e2111
https://doi.org/10.7717/peerj.2111

On causality of extreme events

Massimiliano Zanin

Multiple metrics have been developed to detect causality relations between data describing the elements constituting complex systems, all of them considering their evolution through time. Here we propose a metric able to detect causality within static data sets, by analysing how extreme events in one element correspond to the appearance of extreme events in a second one. The metric is able to detect both linear and non-linear causalities; to analyse both cross-sectional and longitudinal data sets; and to discriminate between real causalities and correlations caused by confounding factors. We validate the metric through synthetic data, dynamical and chaotic systems, and data representing the human brain activity in a cognitive task.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1827v1 | CC-BY 4.0 Open Access | rec: 7 Mar 2016, publ: 7 Mar 2016

On causality of extreme events

Massimiliano Zanin[1,2]

[1] *Innaxis Foundation & Research Institute, José Ortega y Gasset 20, 28006, Madrid, Spain*

[2] *Departamento de Engenharia Electrotécnica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisboa, Portugal*

Multiple metrics have been developed to detect causality relations between data describing the elements constituting complex systems, all of them considering their evolution through time. Here we propose a metric able to detect causality within static data sets, by analysing how extreme events in one element correspond to the appearance of extreme events in a second one. The metric is able to detect both linear and non-linear causalities; to analyse both cross-sectional and longitudinal data sets; and to discriminate between real causalities and correlations caused by confounding factors. We validate the metric through synthetic data, dynamical and chaotic systems, and data representing the human brain activity in a cognitive task.

PACS: 05.45.Tp, 05.45.-a, 87.10.Vg

I. INTRODUCTION

Detecting causality relationships between the elements composing a (complex) system is an old, though unsolved problem [1, 2]. The origin of the concept of *causality* goes back to the ancient Greek phylosophy, according to which causal investigation was the search for an answer to the question "why?" [3, 4]; but the debate was still hot in the late 18th century, in the work of David Hume [5]. Moving to the scientific research, in the last few decades there has been an increasing interest for detecting causality in real data, which has resulted in the creation of multiple metrics: Granger causality, cointegration, or transfer entropy [6–10], to name a few.

All proposed metrics share a common characteristic: causality is defined as a relation existing in the temporal domain, and the metrics thus involve a time series analysis. This probably originated in the way the human brain conceives causality, as sequences of related events close in time [11, 12]. This is an important limitation, especially when studying systems whose dynamics through time cannot easily be observed. Consider, for instance, genetic analysis; one single measurement is usually available per subject and gene, precluding the estimation of gene-gene interactions through a causal analysis solely based on expression levels.

Although correlation appears *prima facie* as an interesting solution, it presents the important drawback of not being able of discriminating between real and spurious causalities. Suppose one is studying a system composed of three interconnected elements, as the one depicted in Fig. 1 Left, with the aim of detecting if the dynamics of element $\mathcal{C}$ is *caused* by $\mathcal{B}$; additionally, no time series are available, and elements are described through vectors of cross-sectional observations. A statistically significant correlation between $\mathcal{B}$ and $\mathcal{C}$ may be found both when a true causality is present (Fig. 1 Right Bottom), and when both elements are driven by an unobserved (confounding) element $\mathcal{A}$ (Fig. 1 Right Top).

In order to tackle the scenario of Fig. 1, in this contribution we propose a novel metric for detecting causality from observational data. It entails three innovative points. First, it is defined on vectors of observation, which do not have to necessarily represent a time evolution. In other words, input vectors may correspond to gene expression levels measured in a population (a cross-sectional study), or (but not necessarily) to multiple observations of the same subject (a longitudinal study). Second, the method is based on the detection

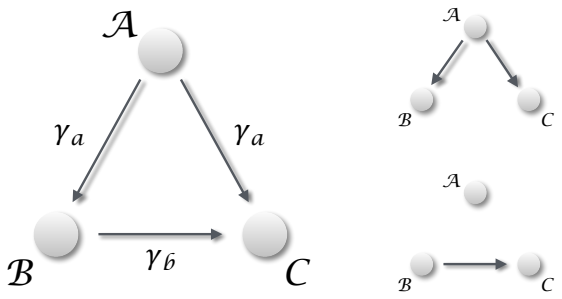

FIG. 1. Distinguishing causality from correlation. (Left) General situation, in which three elements ($\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$) interact in a simple triangular configuration. If one is interested in the relation between $\mathcal{B}$ and $\mathcal{C}$, two different scenarios may arise. (Right top) When $\mathcal{A}$ is dominating the dynamics, any common dynamics between $\mathcal{B}$ and $\mathcal{C}$ will be a correlation, generated by the external confounding factor. (Right bottom) The situation corresponding to a real causality between $\mathcal{B}$ and $\mathcal{C}$.

of extreme events, and on their appearance statistics. This is not dissimilar to Granger causality, as the latter measures how shocks in one time series are explained by a second one; but without the need of a time evolution. Third, it is optimised for the detection of non-linear causal relations, which are common in many real-world complex systems [13], but that may create problems in standard causality metrics [14].

II. METRIC DEFINITION

Suppose two vectors of elements $\mathcal{B} = \{b_i\}$ and $\mathcal{C} = \{c_i\}$ of equal size. The two elements of each pair (b_i, c_i) must be related, *e.g.* they may correspond to the measurement of two biomarkers in a same subject. In the case of $\mathcal{B}$ and $\mathcal{C}$ being time series, clearly (b_i, c_i) would correspond to measurements at time i; yet, as already introduced, such dynamical approach is not required.

Starting from these vectors, some of their elements are labelled as *extreme* when they exceed a threshold, *i.e.* $b_i > \tau_b$ and $c_i > \tau_c$. If a causality relation is present between them, such that $\mathcal{B} \to \mathcal{C}$, this should affect the way extreme events appear. First, under non-extreme dynamics, the two systems $\mathcal{B}$ and $\mathcal{C}$ are loosely coupled. Especially when the relation is of a non-linear nature, small values in the former system are dampened during the transmission. Second, most (ideally, all) of the extreme values of $\mathcal{B}$ should correspond

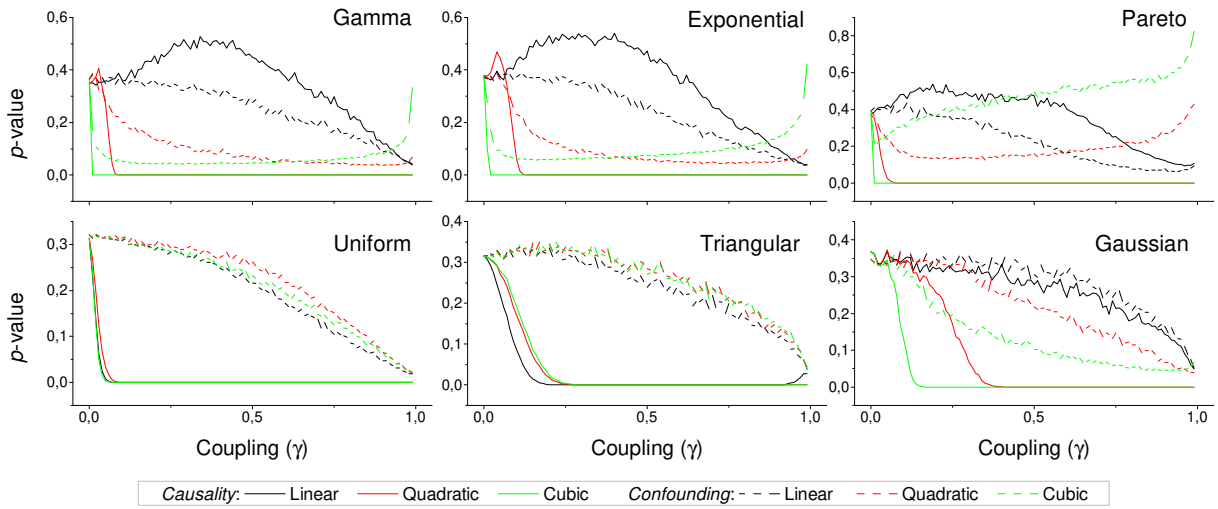

FIG. 2. p-value obtained by the proposed causality metric, for vectors of synthetic data drawn from six different distributions, as a function of the coupling constant γ - see main text for details. Black, red and green lines respectively correspond to linear, quadratic and cubic couplings; solid lines depict true causalities (as in Fig. 1 Right Bottom), dashed lines spurious ones (Fig. 1 Right Top). Each point corresponds to 10.000 realisations.

to extreme values of $\mathcal{C}$, as extreme signals will be amplified from the former to the latter by the non-linear coupling. Third, extreme values of $\mathcal{C}$ only partially correspond to extreme values of $\mathcal{B}$; due to its internal dynamics, $\mathcal{C}$ can display extreme events not triggered by the other element.

Let us denote by p_1 the probability that an extreme event in $\mathcal{C}$ also corresponds to an extreme event in $\mathcal{B}$, *i.e* $p_1 = P(b_i > \tau_b | c_i > \tau_c)$. Conversely, p_2 will denote the probability that an extreme event in $\mathcal{B}$ corresponds to an extreme event in $\mathcal{C}$, *i.e.* $p_1 = P(c_i > \tau_c | b_i > \tau_b)$. In the case of a real causality, the second condition implies that $p_1 \approx 1$, the third one that $p_2 \ll 1$. On the other hand, in the case of an external confounding effect, and if the two thresholds are chosen such that the probability of finding extreme events is the same for both elements, it is easy to see that $p_1 \approx p_2$. Notice that the same is true if $\mathcal{B}$ and $\mathcal{C}$ are bidirectionally interacting.

The previous analysis suggests that the necessary condition for having a $\mathcal{B} \to \mathcal{C}$ causality is $p_1 > p_2$. The statistical significance can be quantified through a binomial two-proportion z-test:

$$z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \tag{1}$$

with n_1 and n_2 the number of events associated to p_1 and p_2, and $\hat{p} = (n_1 p_1 + n_2 p_2)/(n_1 + n_2)$. The corresponding p-value can be obtained through a Gaussian cumulative distribution function.

Before demonstrating the effectiveness of the proposed causality metric, it is worth discussing several aspects of the same.

First of all, the attentive reader will notice the similarity of this method with some metrics for assessing synchronisation in time series. For instance, local maxima and their statistics were considered in Ref. [15], and event coincidences in Ref. [16]. In both cases, an essential ingredient is the time evolution: extreme events in one time series are identified and related to those appearing in a second time series, and the delay required for their transmission assessed through a time shift optimisation. While this yields an estimation of the direction of the information flow between two time series, it cannot be applied to systems whose time evolutions are not accessible. The metric here proposed has the advantage that can be applied to static data sets, in principle paving the way to the construction of data mining algorithms based on causality.

Second, the metric definition requires setting two thresholds, *i.e.* τ_b and τ_c. This can be done using *a priori* information, *e.g.* when a level is accepted as abnormal for a given biomarker; or by simply explore all the parameters space, in order to assess the values of (τ_b, τ_c) corresponding to the lowest p-value. This may result especially useful in those situations for which the input elements are not well characterised: beyond the identification of causality relations, this method may also be used to define what an abnormal value is.

Third, we have previously stated that the presence of a confounding effect can be correctly detected, and that in such situations the metric would not detect a statistically significant causality. According to the Common Cause Principle [1], two variables are unconfounded *iff* they have no common ancestor in the causal diagram; and ensuring this requires including the confounding effects in the analysis, *i.e.* detect if there are causalities $\mathcal{A} \to \mathcal{B}$ and $\mathcal{A} \to \mathcal{C}$ in the diagram of Fig. 1. In the context here analysed, a confounding effect would be detected as the presence of co-occurring extreme events (generated by the confounding element) in both vectors of data. This requires the confounding element to influence in the

same way both analysed elements, or, in other words, to have the same coupling strength between $\mathcal{A} \rightarrow \mathcal{B}$ and $\mathcal{A} \rightarrow \mathcal{C}$. Additionally, if the causality $\mathcal{B} \rightarrow \mathcal{C}$ is mixed with an external influence, the latter cannot be detected if the strength of the former is greater - that is, a strong causality can mask a confounding effect. For all this, the proposed method does not always allow to discriminate true causalities from spurious relationships, although it provides important clues about which one of these two effects is having the strongest impact.

III. METRIC EVALUATION

We first test the proposed metric with synthetic data. Fig. 2 presents the evolution of the p-value for two vectors $\mathcal{B}$ and $\mathcal{C}$, whose values are drawn from different distributions. Two situations are compared. First, a real $\mathcal{B} \rightarrow \mathcal{C}$ causality, such that $c_i = c_i + \gamma b_i^n$ (n being the order of the coupling) - solid lines in Fig. 2. Second, a confounding effect in which $b_i = b_i + \gamma a_i^n$ and $c_i = c_i + \gamma a_i^n$ - dashed lines in Fig. 2. It can be appreciated that the p-values of real causalities drop to zero with small values of coupling constants; and that non-linear couplings perform better than linear ones. In some cases, a confounding effect (especially when highly non-linear) can foul the metric and yield a low p-value - see, for instance, the cubic confounding coupling for a gamma distribution in Fig. 2. Such situations can easily be identified by comparing the p-values for $\mathcal{B} \rightarrow \mathcal{C}$ and $\mathcal{C} \rightarrow \mathcal{B}$: in the case of a true causality, which is by definition directed, the p-value should be small only for one of them. An example of this is depicted in Fig. 3, which shows the evolution of the p-values for a confounding effect (top panel) and a causality (bottom panel), for vectors of Gamma distributed values. Once the limitations and requirements about confounding effects, as defined in the previous section, are taken into account, discriminating between true and spurious causalities only requires calculating the two opposite p-values, and checking whether they are both small.

The necessity of detecting extreme events introduces a drawback in the method, *i.e.* the need of having a large set of input values to reach a stable statistics. This problem is explored in Fig. 4, which depicts the p-value obtained as a function of the number of input values. Depending on the kind of relation to be detected, between 2 and 4 thousand values are required.

One of the advantages of the proposed metric is that it can be applied both to cross-sectional and longitudinal (*i.e.* time evolving) data. Here we show such flexibility in the

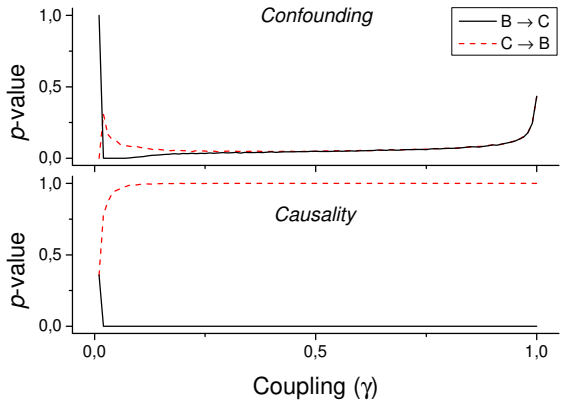

FIG. 3. Evolution of the p-value of the causality, when considering both $\mathcal{B} \to \mathcal{C}$ and $\mathcal{C} \to \mathcal{B}$ tests for a cubic coupling and for data drawn from a Gamma distribution (as in green lines of the first panel of Fig. 2. The top panel reports the results for a confounding effect, the bottom one for a true causality between $\mathcal{B}$ and $\mathcal{C}$.

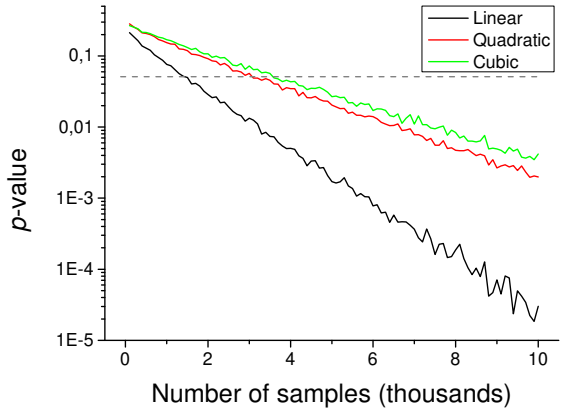

FIG. 4. Evolution of the p-value of the causality, for a triangular distribution, as a function of the number of values included in the input vectors. Black, red and green lines respectively correspond to linear, quadratic and cubic couplings.

detection of the causality between two noisy Kuramoto oscillators [17, 18]. Suppose two oscillators whose phases are defined as:

$$\dot{\phi_{\mathcal{B}}} = \kappa_{\mathcal{B}} + \xi \tag{2}$$

$$\dot{\phi_{\mathcal{C}}} = \kappa_{\mathcal{C}} + \gamma sin(\phi_{\mathcal{B}} - \phi_{\mathcal{C}}) + \xi. \tag{3}$$

κ is the natural frequency of each oscillator ($\kappa_{\mathcal{B}} \neq \kappa_{\mathcal{C}}$), and ξ an external uniform noise

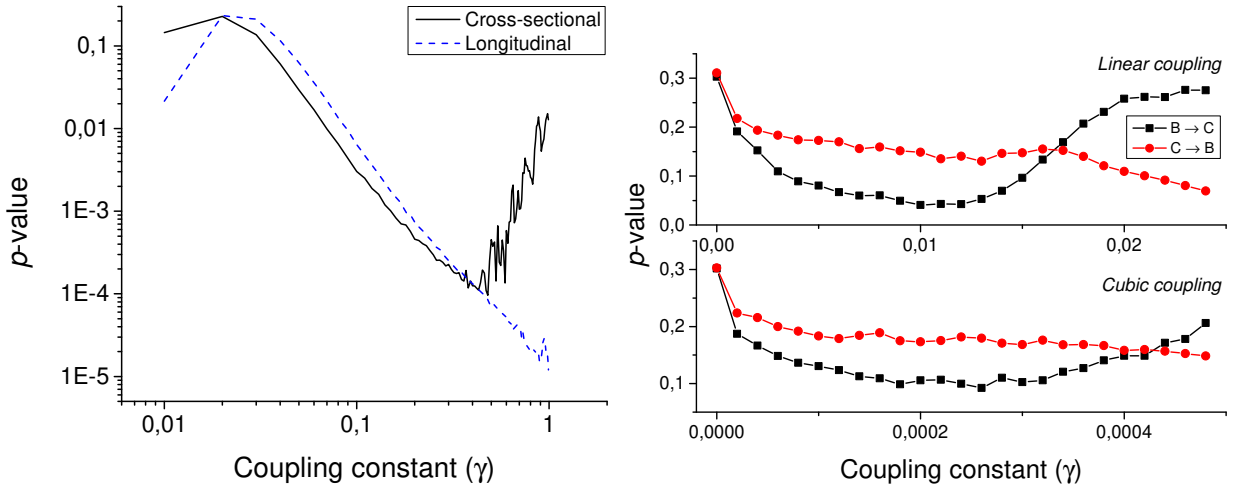

FIG. 5. (Left) Evolution of the p-value of the causality test between two Kuramoto oscillators, for different values of the coupling constant γ. The solid black line and the dashed blue one respectively correspond to a cross-sectional and longitudinal study - see main text for details. (Right) p-value for two coupled Rössler oscillators as a function of the coupling constant γ, for a linear (top graph) and cubic (bottom graph) coupling.

source. The coupling constant γ defines the way the two oscillators interact, with independent dynamics for $\gamma \approx 0$, and a causality $\phi_{\mathcal{B}} \rightarrow \phi_{\mathcal{C}}$ for $\gamma > 0$. The longitudinal causality can be detected by considering the time series created by $\dot{\phi}_{\mathcal{B}}$ and $\dot{\phi}_{\mathcal{C}}$, thus focusing on how abnormal *jumps* in the phase of the oscillators is transmitted from the former to the latter. The p-value of the metric is represented in Fig. 5 Left by the blue dashed line. The equivalent cross-sectional analysis requires multiple realisations of the previous dynamics; for each one of them, one single pair of values $(\dot{\phi}_{\mathcal{B}}, \dot{\phi}_{\mathcal{C}})$ is extracted, corresponding to the largest variation of $\phi_{\mathcal{B}}$ (and thus, to the most extreme jump in the phase of the first oscillator). The evolution of the corresponding p-value is shown in Fig. 5 Left by the black solid line. Both the longitudinal and cross-sectional analyses yield similar results, suggesting that dynamical and static causalities are equivalent under the proposed metric.

An important characteristic of complex systems is that their constituting elements usually have a chaotic dynamics [13], making more complicated the task of detecting causality between them. We here test the proposed metrics by considering two unidirectionally coupled Rössler oscillators ($\mathcal{B} \rightarrow \mathcal{C}$) in their chaotic regime - see [19] for details. We consider both linear and cubic couplings; following the notation in [19], this means:

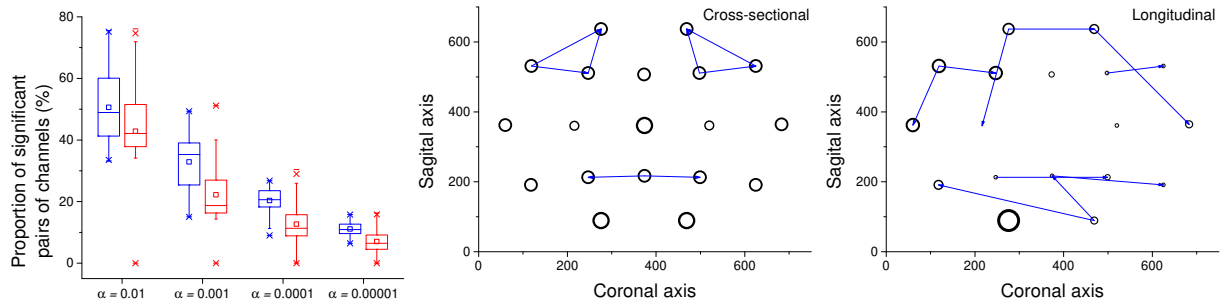

FIG. 6. Analysis of causality in EEG data. (Left) Proportion of pairs of channels in which causality has been detected, for cross-sectional (blue) and longitudinal (red) analyses, as a function of the significance level α. (Center) Top-10 causality links in the cross-sectional analysis. (Right) Top-10 causality links in the longitudinal analysis. In the central and right panel, the size of the node is proportional to its weight.

$$\dot{y}_1 = -(y_2 + y_3) - \gamma(y_1 - x_1), \text{and} \tag{4}$$

$$\dot{y}_1 = -(y_2 + y_3) - \gamma(y_1 - x_1)^3. \tag{5}$$

Time series are created by sampling the second dimension of each oscillator (*i.e.* x_2 and y_2) with a resolution lower than the intrinsic frequency. Fig. 5 Right depicts the evolution of the p-value for low coupling strengths γ, thus ensuring that the system is *generalised synchronised*. For $\gamma \approx 0.01$ ($\gamma \approx 2 \cdot 10^{-4}$ for cubic coupling), a true causality is detected, while for $\gamma > 0.015$ ($\gamma > 4 \cdot 10^{-4}$) the two oscillators start to synchronise.

The possibility of combining a cross-sectional analysis of extreme values with a longitudinal analysis opens new doors towards the understanding of systems for which both aspects can be studied at the same time. Here we show how this can be achieved in the analysis of functional networks representing the structure of brain activity in healthy subjects [20, 21]. The data set corresponds to electroencephalographic (EEG) recordings of 40 subjects during 50 trials of an object recognition task (details can be found in [22] and references within), obtained through the UCI KDD archive [23]. For each trial and subject, 19 time series (corresponding to 19 EEG channels in the $10-20$ configuration) of 256 samples were available. The longitudinal analysis was performed by calculating the causality using the raw time series. On the other hand, the cross-sectional analysis relies on identifying the propagation of extreme events, as in the case of the Kuramoto oscillators. Extreme events are defined

as those for which the energy of the signal is maximum in a given time series; the energy is defined, at each time point, as the deviation with respect to the mean, normalised by the standard deviation of the signal - *i.e.* as the absolute value of the Z-Score.

Fig. 6 (Left) depicts a box plot of the proportion of significant pairs of channels (*i.e.* pairs of channels for which a causality was detected), in both the cross-sectional (blue) and longitudinal (red) analyses, for different significance levels α. In the case of the cross-sectional analysis, each value corresponds to the results for a single subject. Results are qualitatively equivalent, with the longitudinal analysis detecting slightly less links than the cross-sectional one for small values of α. Fig. 6 Center and Right depict the 10 most significant links, as detected by both analyses. While not completely equivalent, both graphs suggest that some areas are identified as active by both methods, *e.g.* the frontal lobe on the top and the visual and somatosensory integration area in the bottom. Remarkably, these two regions are expected to be relevant for the task studied, *i.e.* object identification: the former for higher function planning (react to the image shown), the latter in the processing of visual inputs.

IV. CONCLUSIONS

In conclusion, we presented a novel metric able to detect causality relationships both in static and time-evolving data sets, thus overcoming the limitation of existing metrics that rely on time series analysis. The proposed metric is designed to detect the propagation of extreme events, or shocks, and as such is more efficient when non-linear relations are present; it is further able to discriminate real from spurious causalities, thus enabling the detection of confounding effects. The effectiveness of the metric has been tested through synthetic data, data obtained from simple and chaotic dynamical systems (Kuramoto and Rössler oscillators), and on EEG data representing the activity of the human brain during an object recognition task.

The possibility of detecting causality in static data sets is expected to be of increasing importance in those research fields in which time dynamics are not available, and that require ensuring that a causality is not just the result of the presence of a confounding factor. For instance, one may considering the raising field of biomedical data analysis [24–26]. The custom solution is to resort to data mining algorithms, which allow to detect and

make explicit patterns in the input data, with the final objective of using such patterns in diagnostic and prognostic models [27]. Nevertheless, data mining (and machine learning in general) is based on the Bayes theorem, a form of statistics of co-occurrences, and thus on a generalised concept of correlation. These methods are thus sensitive to the confounding effects that are frequently in place, as genes and metabolites create an intricate network of interactions. Resorting to classical causality metrics, like Granger's one, is not possible, as time series are seldom available - measuring gene expression or metabolite levels is an expensive and slow process. In spite of this, causality is an essential element to be detected: if one only focuses on correlations, there is a risk of detecting elements whose manipulation does not guarantee the expected results on the system [28–30]. We foresee that the proposed causality metric can be an initial solution to this problem, by providing a causality test that can be applied to static data, and that could be used as the foundation of a new class of data mining algorithms.

A Python implementation of the proposed causality metric is freely available at `www.mzanin.com/Causality`.

---

[1] J. Pearl, Econometric Theory **19**, 675 (2003).

[2] J. Pearl, *Causality* (Cambridge university press, 2009).

[3] M. G. Evans, Philosophy and Phenomenological Research , 466 (1959).

[4] R. J. Hankinson, *Cause and explanation in ancient Greek thought* (Clarendon Press Oxford, 1998).

[5] D. Hume *et al.*, *An enquiry concerning human understanding* (Alex Catalogue, 1965).

[6] C. W. Granger, Journal of econometrics **39**, 199 (1988).

[7] C. W. Granger, Journal of Economic Dynamics and Control **12**, 551 (1988).

[8] T. Schreiber, Physical review letters **85**, 461 (2000).

[9] M. Staniek and K. Lehnertz, Physical Review Letters **100**, 158101 (2008).

[10] P. Verdes, Physical Review E **72**, 026222 (2005).

[11] A. M. Leslie and S. Keeble, Cognition **25**, 265 (1987).

[12] S. C. Tanaka, B. W. Balleine, and J. P. O'Doherty, The Journal of Neuroscience **28**, 6750 (2008).

[13] S. H. Strogatz, *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering* (Westview press, 2014).

[14] C. W. Granger and T. Terasvirta, OUP Catalogue (1993).

[15] R. Q. Quiroga, T. Kreuz, and P. Grassberger, Physical review E **66**, 041904 (2002).

[16] J. F. Donges, C.-F. Schleussner, J. F. Siegmund, and R. V. Donner, arXiv preprint arXiv:1508.03534 (2015).

[17] Y. Kuramoto, "Chemical oscillations, waves, and turbulence," (2012).

[18] Rodrigues, Francisco A, Peron, Thomas K DM, Ji, Peng, and Kurths, Jürgen, Physics Reports (2015).

[19] N. F. Rulkov, M. M. Sushchik, L. S. Tsimring, and H. D. I. Abarbanel, Physical Review E **51**, 980 (1995).

[20] E. Bullmore and O. Sporns, Nature Reviews Neuroscience **10**, 186 (2009).

[21] M. Rubinov and O. Sporns, Neuroimage **52**, 1059 (2010).

[22] X. L. Zhang, H. Begleiter, B. Porjesz, W. Wang, and A. Litke, Brain Research Bulletin **38**, 531 (1995).

[23] S. D. Bay, D. Kibler, M. J. Pazzani, and P. Smyth, ACM SIGKDD Explorations Newsletter **2**, 81 (2000).

[24] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, and W. E. Hammond, in *Proceedings of the AMIA annual fall symposium* (American Medical Informatics Association, 1997) p. 101.

[25] K. J. Cios and G. W. Moore, Artificial intelligence in medicine **26**, 1 (2002).

[26] J. Han, in *BIOKDD* (2002) pp. 1–2.

[27] V. Vapnik, *The nature of statistical learning theory* (Springer Science & Business Media, 2013).

[28] E. Salmon, F. Collette, C. Degueldre, C. Lemaire, and G. Franck, Human brain mapping **10**, 39 (2000).

[29] L. R. Cardon and L. J. Palmer, The Lancet **361**, 598 (2003).

[30] V. A. Vakorin, O. A. Krakovska, and A. R. McIntosh, Journal of neuroscience methods **184**, 152 (2009).