

CT 1.3.4

From the complexity of probability distributions to measures of graph complexity

E. Olbrich¹, N. Ay^{1,2}, N. Bertschinger¹, and J. Jost^{1,2}

¹Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, D 04103 Leipzig, Germany

²Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

Introduction

Instead of approaching graph complexity directly via graph theoretic properties such as the degree distribution or the community structure, we here propose a statistical approach. Our approach is based on measures of the complexity of a joint probability distribution of N random variables, such as the TSE-complexity [1] or the excess entropy [2-4]. Thus, we propose to quantify the complexity of a graph by assigning a probability distribution to it to which we can apply such statistical complexity measures.

Measures of statistical complexity

A *graph* \mathcal{G} is a pair $\mathbf{G} = \{V, E\}$, where V is a set of $1 \leq N < \infty$ nodes and E , a subset of the set $V \times V$ of ordered pairs of vertices, called the *edges* or *links* of \mathcal{G} , tells us which of the nodes are connected to each other. This describes a given graph. Our approach, however, consists in considering the graph as a member of some given ensemble.

In order to make this precise and to employ statistical complexity measures we need to introduce state sets X_v , $v \in V$ on the nodes. For a subset $A \subseteq V$, we write X_A instead of $\times_{v \in A} X_v$; the total configuration set then is X_V . The complexities are defined for probability measures on X_V . The *entropy* of X_C is defined as

$$H(X_C) := - \sum_{z \in X_C} \mathbf{Pr}(X_C = z) \log_2(\mathbf{Pr}(X_C = z))$$

We now list several complexity measures that can be applied in the framework just sketched:

The ***Multi-information*** quantifies the total amount of statistical dependencies in the system:

$$I(X_V) := \sum_{v \in V} H(X_{\{v\}}) - H(X_V)$$

In the case of time series the ***Excess entropy*** (also known as effective measure complexity [2] or predictive information [3]) is the best understood measure of statistical complexity. In this context it provides a lower bound for the amount of memory needed for an optimal prediction. A similar quantity can be defined for arbitrary joint probability distributions as the difference between the uncertainty of the state of the whole system and the sum of the remaining uncertainties of the state of the elements using all information available in the system:

$$E(X_V) := H(X_V) - \sum_{v \in V} H(X_{\{v\}} | X_{V \setminus \{v\}}).$$

It quantifies the “explainable” part of the variety of the system. The *TSE-complexity* was introduced in [1], motivated by the attempt to measure the potential ability of a neural system to produce consciousness. It is defined as

$$C_{TSE}(X_V) := \sum_{k=1}^N \left(H(k, N) - \frac{k}{N} H(X_V) \right),$$

with

$$H(k, N) = \binom{N}{k}^{-1} \sum_{\substack{Y \subseteq V \\ |Y|=k}} H(X_Y)$$

denoting the average entropy of subsystems of size k .

The complexity of random graphs

In order to use these measures for the characterization of graphs we have to assign a probability distribution to the graph. One possibility is to consider the graph as a typical realization of a graph drawn from an ensemble and to estimate the complexity of the probability distribution characterizing that ensemble of random graphs. In fact, this step is needed in order to distinguish the “structure” of the graph from the accidental “random” properties. This approach will be illustrated for Erdős-Renyi random graphs and random graphs with a given degree distribution.

Graphical Models

A second possibility to relate the graph with a probability distribution comes from the theory of graphical models [6]. There the conditional independencies are expressed by a graph, implying a product structure of the distribution. There are, however, many probability distributions corresponding to a given graph \mathbf{G} . We propose to resolve this ambiguity by choosing the distribution which maximizes the information flow along the edges of the graph. We shall discuss the corresponding graph complexities for simple examples.

Acknowledgement

We thank the Volkswagen Foundation for support.

References

- [1] G. Tononi, O. Sporns, and G. M. Edelman, PNAS 91, 5033 (1994).
- [2] P. Grassberger, Int. J. Theor. Phys. 25, 907 (1986).
- [3] W. Bialek, I. Nemenman, and N. Tishby, Neural Computation 13, 2409 (2001)
- [4] J. P. Crutchfield and K. Young, Phys. Rev. Lett. 63, 105 (1989).
- [5] E. Olbrich, N. Bertschinger, N. Ay, and J. Jost, European Journal of Physics B (2008), <http://dx.doi.org/10.1140/epjb/e2008-00134-9>
- [6] S. L. Lauritzen, Graphical Models, Oxford University Press, 1996