

From the complexity of probability distributions to measures of graph complexity

Eckehard Olbrich, Nihat Ay, Nils Bertschinger, Jürgen Jost

MPI MIS Leipzig

Netsci'08 Norwich 25.06.2008

Algorithmic and Statistical complexity

- **Algorithmic** complexity: length of the shortest program which can generate the object
- **Statistical** complexity: Not all details are important. The measure should characterize the complexity of the structure.
- Not one object is considered, but a set of objects with a probability distribution. **Random** means now statistically independent.
- Measures of statistical complexity measure the amount of statistical dependencies in a distribution. They should be zero in both cases of a totally ordered and a total random system
- In the context of time series this idea led to the equivalent notions of *effective measure complexity* (Grassberger), *excess entropy* (Crutchfield) and *predictive information* (Bialek et al.), which measure the amount of information that is necessary for an optimal prediction.

Measures of statistical complexity and graph complexity

- Most measures of graph complexity rely on graph theoretical properties, which can be measured in a given graph.
- Measures of statistical complexity require some probability distribution.
- Two problems for a measure of statistical graph complexity:
 - 1 We need a measure of statistical complexity for finite systems without a linear order between the elements.
 - 2 We have to assign to a given graph a probability distribution.

- 1 Statistical complexity measures:
 - 1 Modified excess entropy for finite systems
 - 2 Tononi-Sporns-Edelman (TSE)-complexity
- 2 Assigning a probability distribution
 - 1 Graphical models: The graph represents the conditional independence structure of a distribution
 - 2 Random graphs: The graph is a typical representative from an ensemble of graphs.

- “World”: a set V of $1 \leq N < \infty$ elements with state sets \mathcal{X}_v , $v \in V$.
- Different possibilities for the elements in the case of a random graph:
 - the links — binary random variable $\mathcal{X}_v = \{0, 1\}$
 - the nodes — vector of links, row of the adjacency matrix
- Given a finite subset $A \subseteq V$ we write \mathcal{X}_A instead of $\times_{v \in A} \mathcal{X}_v$
- Given a probability vector p on \mathcal{X}_V we get random variables X_A .
- **Entropy** measures variety or uncertainty

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

- **Conditional entropy:** remaining uncertainty of X_C given X_B

$$H(X_C | X_B) := H(X_B, X_C) - H(X_B)$$

- **Mutual information** measures how much information the state of X_B contains about X_C and vice versa

$$MI(X_C : X_B) := H(X_C) - H(X_C | X_B)$$

- **Conditional mutual information** measures how much additional information about X_C one gets from observing X_B if one already knows X_A

$$MI(X_C : X_B | X_A) := H(X_C | X_A) - H(X_C | X_A, X_B) \geq 0$$

- **Mutual information** measures how much information the state of X_B contains about X_C and vice versa

$$MI(X_C : X_B) := H(X_C) - H(X_C | X_B)$$

- The **multi-information** or **integration** is a generalization of the mutual information for more than two random variables. The multi-information of the system X_V with respect to its nodes is defined as

$$I(X_V) := \sum_{v \in V} H(X_{\{v\}}) - H(X_V) = D \left(p(x_V) \parallel \prod_{v \in V} p_v(x_{\{v\}}) \right)$$

It is the difference between the sum of the variety of the elements and the variety of the system as a whole.

Excess entropy for a time series

- Elements X_i are the values at a certain time i
- Entropy rate:

$$\begin{aligned}h_\infty &= \lim_{N \rightarrow \infty} \frac{H(X_1, \dots, X_N)}{N} \\ &= \lim_{N \rightarrow \infty} H(X_N | X_{N-1}, \dots, X_1)\end{aligned}$$

quantifies how “random” the string is.

- The non-extensive part of the entropy, the **excess entropy**, can then be considered then as quantifying the structure in the sequence:

$$\begin{aligned}E &= \lim_{N \rightarrow \infty} E(N) \quad \text{with} \\ E(N) &= H(X_1, \dots, X_N) - NH(X_N | X_{N-1}, \dots, X_1)\end{aligned}$$

Excess entropy for a time series

- Elements X_i are the values at a certain time i
- The non-extensive part of the entropy, the **excess entropy**, can then be considered then as quantifying the structure in the sequence:

$$E = \lim_{N \rightarrow \infty} E(N) \quad \text{with}$$
$$E(N) = H(X_1, \dots, X_N) - NH(X_N | X_{N-1}, \dots, X_1)$$

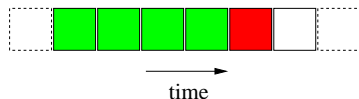
- If all limits exist, it can be rewritten as the mutual information between the future and the past — **predictive information**

$$E = \lim_{N_1, N_2 \rightarrow \infty} MI(X_{-N_1}, \dots, X_{-1} : X_0, \dots, X_{N_2})$$

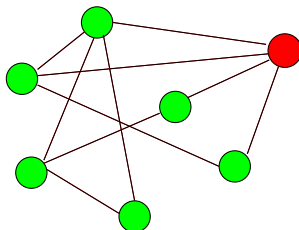
- or as the weighted sum over distance k information flows

$$E(N) = \sum_{k=2}^N (k-1) MI(X_k : X_1 | X_{k-1}, \dots, X_2)$$

Excess entropy for a finite system



time series



finite system

- Time series: Excess entropy

$$E_N = H(X_1, \dots, X_N) - NH(X_N | X_{N-1}, \dots, X_1)$$

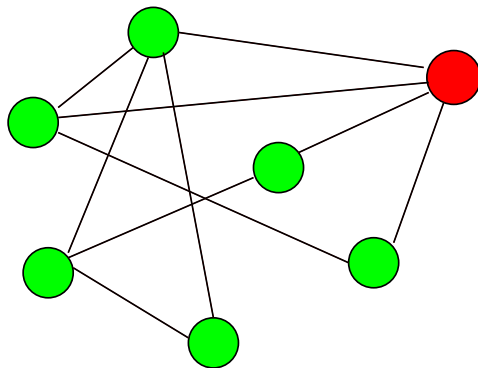
depends on the order of the elements.

- Order independent possibility

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

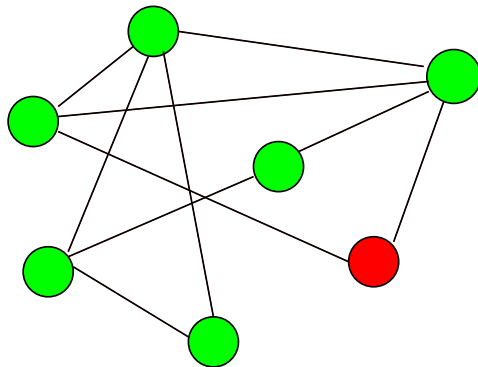
Excess entropy

- $H(X_{\{v\}}|X_{V\setminus\{v\}})$ quantifies the amount to which the state of a single element cannot be explained by dependencies in the system and is therefore considered as random.



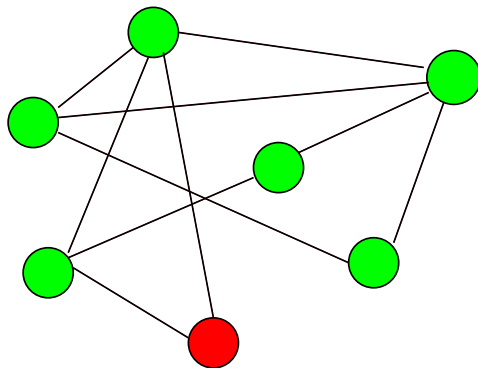
Excess entropy

- $H(X_{\{v\}}|X_{V\setminus\{v\}})$ quantifies the amount to which the state of a single element cannot be explained by dependencies in the system and is therefore considered as random.



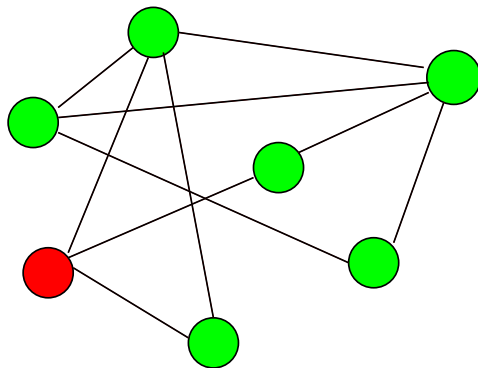
Excess entropy

- $H(X_{\{v\}}|X_{V\setminus\{v\}})$ quantifies the amount to which the state of a single element cannot be explained by dependencies in the system and is therefore considered as random.



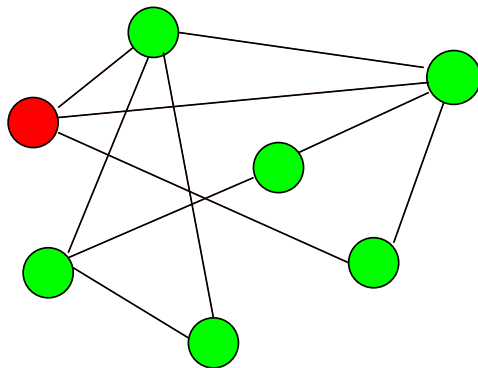
Excess entropy

- $H(X_{\{v\}}|X_{V\setminus\{v\}})$ quantifies the amount to which the state of a single element cannot be explained by dependencies in the system and is therefore considered as random.



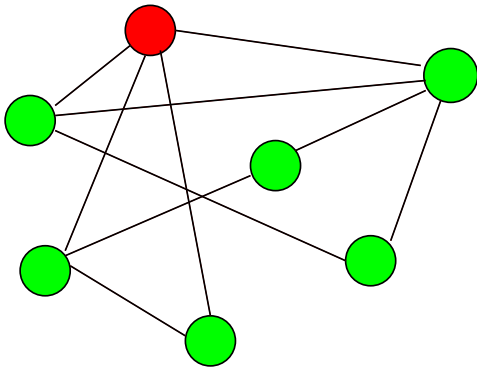
Excess entropy

- $H(X_{\{v\}}|X_{V\setminus\{v\}})$ quantifies the amount to which the state of a single element cannot be explained by dependencies in the system and is therefore considered as random.



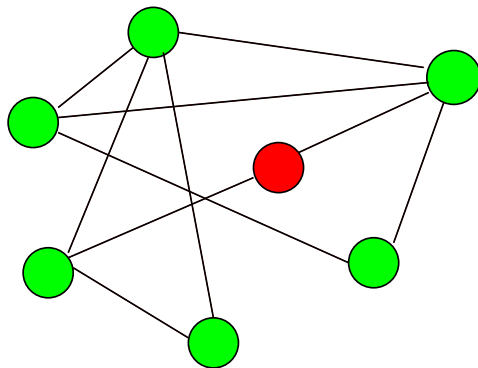
Excess entropy

- $H(X_{\{v\}}|X_{V\setminus\{v\}})$ quantifies the amount to which the state of a single element cannot be explained by dependencies in the system and is therefore considered as random.



Excess entropy

- $H(X_{\{v\}}|X_{V\setminus\{v\}})$ quantifies the amount to which the state of a single element cannot be explained by dependencies in the system and is therefore considered as random.



- $H(X_{\{v\}}|X_{V\setminus\{v\}})$ quantifies the amount to which the state of a single element cannot be explained by dependencies in the system and is therefore considered as random.
- The excess entropy is then the difference between the uncertainty of the state of the whole system and the sum of the irreducible 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.

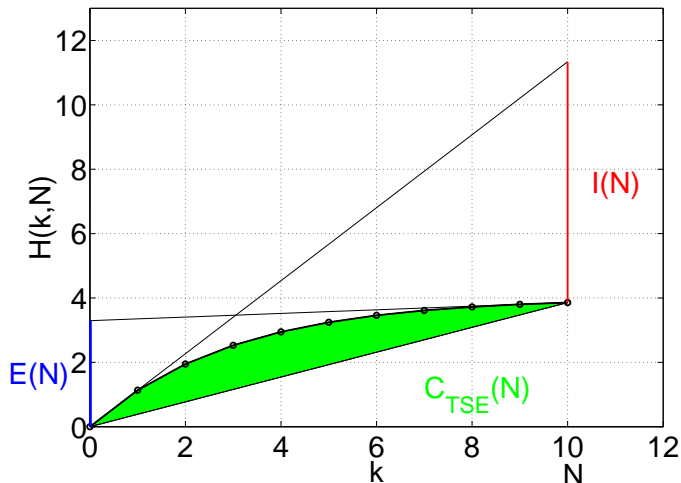
$$C_{TSE}(X_V) = \sum_{k=1}^N \left(H(k, N) - \frac{k}{N} H(N) \right) \quad H(k, N) = \binom{N}{k}^{-1} \sum_{\substack{Y \subseteq V \\ |Y|=k}} H(X_Y)$$

- High TSE-complexity requires low integration for small subsystems and high integration at the system level:

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

$$I(k, N) = \binom{N}{k}^{-1} \sum_{\substack{Y \subseteq V \\ |Y|=k}} \left(\sum_{v \in Y} H(X_{\{v\}}) - H(X_Y) \right)$$

Excess entropy and TSE-complexity



Excess entropy and TSE-complexity

- Excess entropy relates the system level to the level of elements, the TSE-complexity considers also the level in between. Excess entropy depends on the system level entropy and its growth at the last step from $N - 1$ to N , the TSE is governed by the growing behavior at all levels.
- We have shown that

$$C_{TSE}(X_V) = \frac{1}{2} \sum_{k=1}^N E(k, N) = \frac{1}{2} \sum_{Y \subseteq V} \frac{1}{\binom{N}{|Y|}} E(X_Y),$$

i.e. the TSE-complexity is proportional to the sum over the mean excess entropies averaged over all subsets of the same size.

- Even if the Excess entropy remains constant for growing N , the TSE-complexity will grow extensively \Rightarrow renormalized TSE-complexity

$$\tilde{C}_{TSE}(N) = \frac{1}{N+1} C_{TSE}(N)$$

- 1 Graphical models: The graph represents the conditional independence structure of the probability distribution.
Problem: There is more than one distribution with the same independence structure. Which one should we consider?
- 2 Random graphs: The graph represents a typical example drawn from a distribution of Graphs.

- Graphical models represent the conditional independence structure of a probability distribution.
- Nodes correspond to the random variables.
- If a undirected graph \mathcal{G} is used as a graphical model the probability distribution factorizes according to

$$p(x_V) = \prod_{C \in \mathcal{C}} a_C(x_C)$$

with \mathcal{C} denoting the set of cliques (complete subgraphs) of \mathcal{G} .

- Pairwise Markov property: Two nodes are conditionally independent, given the rest of the graph, if they are not neighbors.
- A whole family of distributions corresponds to one graph.

Selecting one distribution

- Maximum entropy? **No**, because the maximum entropy distribution would be the distribution of independent variables.
- Look at the information flow between nodes:

$$\begin{aligned} IF(i,j) &= MI(X_{\{i\}} : X_{\{j\}} | X_{V \setminus \{i,j\}}) \\ &= H(X_{\{i\}} | X_{V \setminus \{i,j\}}) - H(X_{\{i\}} | X_{V \setminus \{i\}}) \end{aligned}$$

Because of the **pairwise Markov property** we can restrict ourselves to neighbours.

- Maximal information flow $\sum_{i \sim j} IF(i,j)$? **No**, because this can lead to distributions, with no information flow on certain links.
- Our actual (preliminary) answer: Choose the distribution with the maximal minimal information flow on the links.

Problems with this approach

- Finding the corresponding distribution is difficult already for very small simple graphs.
- At the moment there are no general results for the complexity measures available, e.g. we do not know how they depend on the concrete choice of the random variable (binary or more states).
- In the case of binary variables, the excess entropy is maximized by a distribution, where $N - 1$ observables are independent, with $p(0) = p(1) = 0.5$ and the last one is the parity function. The graphical model of this distribution corresponds is the fully connected graph, The information flow on each link is maximal, i.e. 1 bit. Thus the **fully** connected graph would have **maximal complexity** in this setting.
- Alternative: Use TSE-complexity.

- Concrete graph is considered as a representant of an ensemble of random graphs described by a probability distribution.
- Parameters of the distribution have to be estimated from the single graph.
- It would be nice to have a hierarchy of correlations related to increasing complexity such as Markov-models of increasing order in the case of time series.
- Zeroth order - no correlation \Rightarrow Erdős-Renyi random graphs

- Only one parameter: Probability p of links between two nodes
- If the elements are the **links**, the distribution factorizes and all complexities are zero.
- If the elements are the **nodes**, i.e. the rows of the adjacency matrix. Only in the case of undirected graphs we get non-vanishing complexities. The complexity measures are monotonic functions of $H(p) = -p \log p - (1 - p) \log(1 - p)$ and maximal for $p = 1/2$.
- Degree distribution: constructing an ensemble of undirected random graphs with a given degree distribution already introduces degree correlations as shown by Park and Newman (2003) — thus leading to non-trivial complexities.
- Next steps: Taking into account higher order correlations.

Summary

- The excess entropy and the TSE-complexity are useful measures of the statistical complexity of finite systems described by a probability distribution.
- In order to apply these to graphs one has to related a given graph to a probability distribution.
- Interpreting the graph as a graphical model of a distribution using an information flow criterion we found that the fully connected graph has maximal complexity if measured by the excess entropy.
- At the moment more promising: Random graph approach.
- Results depend on what is considered as the elementary unit: link or node.
- Next steps: Constructing a sequence of random graph ensembles taking more and more correlations into account, similar to Markov models for temporal sequences.