More on Entropy

MSDM 5058

Prepared by S.P. Li

When Shannon was asked what he had thought about when he had finally confirmed his famous measure. Shannon replied:

"*My greatest concern was what to call it. I thought of calling it information, but the word was overly used, so I decided to call it uncertainty. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it* ***entropy****, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage.'*"

What is Entropy?

Entropy is a *scientific* concept, as well as a measurable physical property, that is most commonly associated with a state of *disorder, randomness, or uncertainty*. The term and the concept are used in diverse fields, from *classical thermodynamics*, where it was first recognized, to the microscopic description of nature in *statistical physics*, and to the principles of *information theory*.

(Wikipedia)

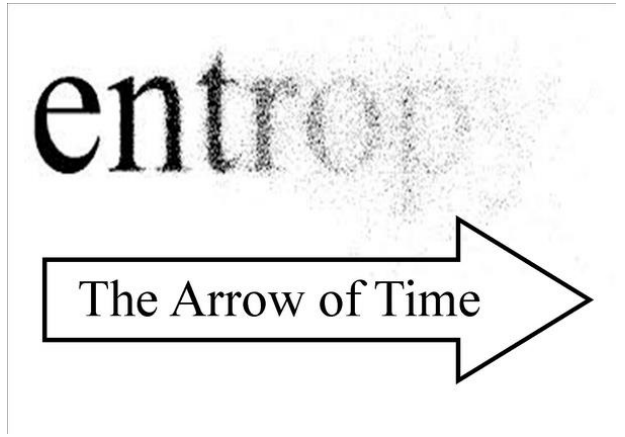
What is Entropy?

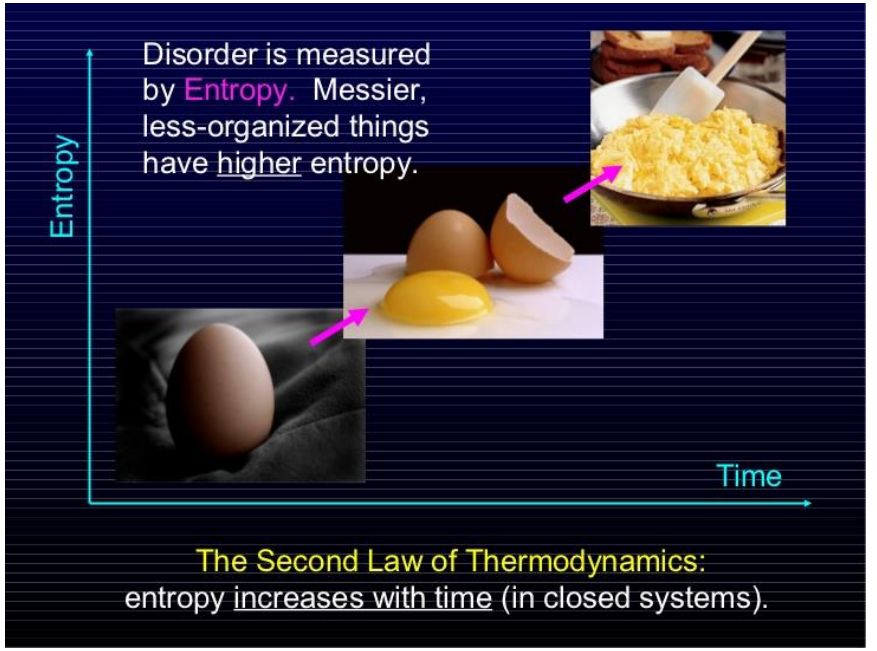
- Lack of order or predictability; gradual decline into disorder

- In ***Physics***: A thermodynamic quantity representing the unavailability of a system's thermal energy for conversion into mechanical work, often interpreted as the degree of disorder or randomness in the system.

- Second law of thermodynamics – Entropy always increases with time

- In ***Information Science***: A measure of uncertainty or surprise (***Shannon Entropy***)





What is Entropy?

Entropy increases; Complexity first increases, then decreases

Low entropy	Medium entropy	High entropy
Low complexity	High complexity	Low complexity

A Brief History of Entropy

- Coined by Clausius (1850) from Greek "en-" = in + "trope" = a turning. The word reveals analogy to "energy" and was designed to mean the form of energy that any energy eventually and inevitably "turns into" – a useless heat. The idea was inspired by an earlier formulation by Carnot (1824) of what is now known as the *Second Law of Thermodynamics*.

- Boltzmann (1877) put entropy into the probabilistic setup of statistical mechanics.

- John von Neumann (1932) generalized to quantum mechanics.

- Shannon (1948) introduced *Information Entropy* as a term in probability and information theory

- …….

Information and Thermodynamic Entropy

Thermodynamic entropy:

$$\Delta S = S_f - S_i = {}_R \int_i^f \frac{\delta Q_R}{T}$$

Boltzmann entropy:

$$S_B[\Gamma_{D_i}] = -k_B \log W$$

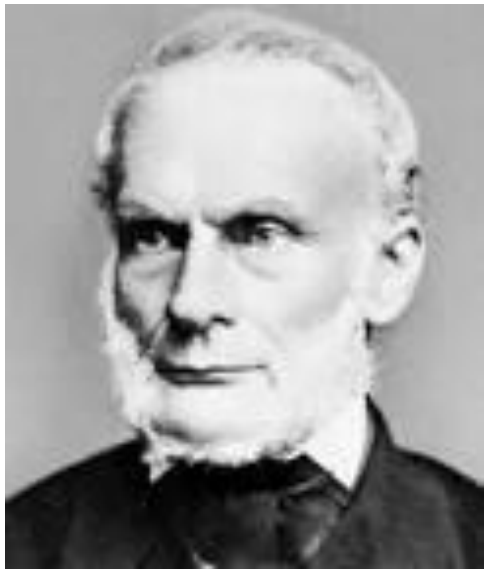
Gibbs entropy:

$$S_G[\rho] = -k_B \int_\Omega \rho(x,t) \ln\big(\rho(x,t)\big)\ dx$$

Shannon entropy:

$$H[X] = -\sum_{j=1}^{l} p_j \log_2 p_j$$

Rudolf Julius Emanuel Clausius



Jan 2, 1822 – Aug 24, 1888

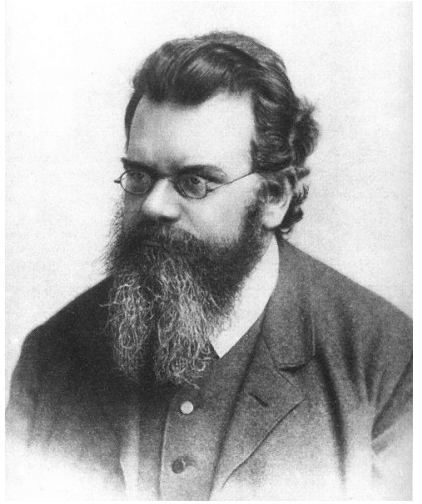
German theoretical physicist and mathematician who played an important role in establishing theoretical physics as a discipline. His most famous paper was read to the Berlin Academy on February 18, 1850 and published in Annalen der Physik in the same year, which marks the foundation of the modern thermodynamics. In his 1865 paper, Clausius introduced the concept of ***entropy***, and stated the First and Second laws of thermodynamics in the following form

1) The energy of the universe is constant.
2) The entropy of the universe tends to a maximum.

A quote from Clausius:

"I prefer going to the ancient languages for the names of important scientific quantities, so that they mean the same thing in all living tongues. I propose, accordingly, to call S the ***entropy*** *of a body, after the Greek word "****transformation****." I have designedly coined the word entropy to be similar to* ***energy****, for these two quantities are so analogous in their physical significance, that an analogy of denominations seems to be helpful."*

Ludwig Boltzmann



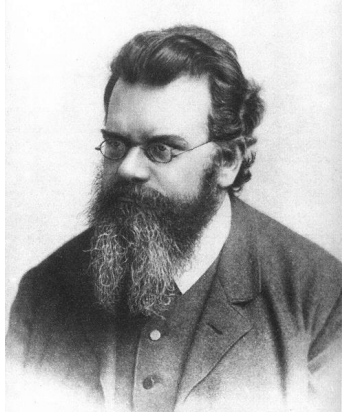
Feb 20, 1844 – Sept 5, 1906

Austrian physicist and philosopher. His greatest achievements were the development of statistical mechanics, and the statistical explanation of the second law of thermodynamics, which he did independently of Willard Gibbs. Their theories describe how macroscopic observations (such as temperature and pressure) are related to microscopic parameters that fluctuate around an average.

Entropy in Statistical Mechanics

- *Goal:* To explain the behavior of *macroscopic* systems in terms of the dynamical laws governing their *microscopic* constituents.

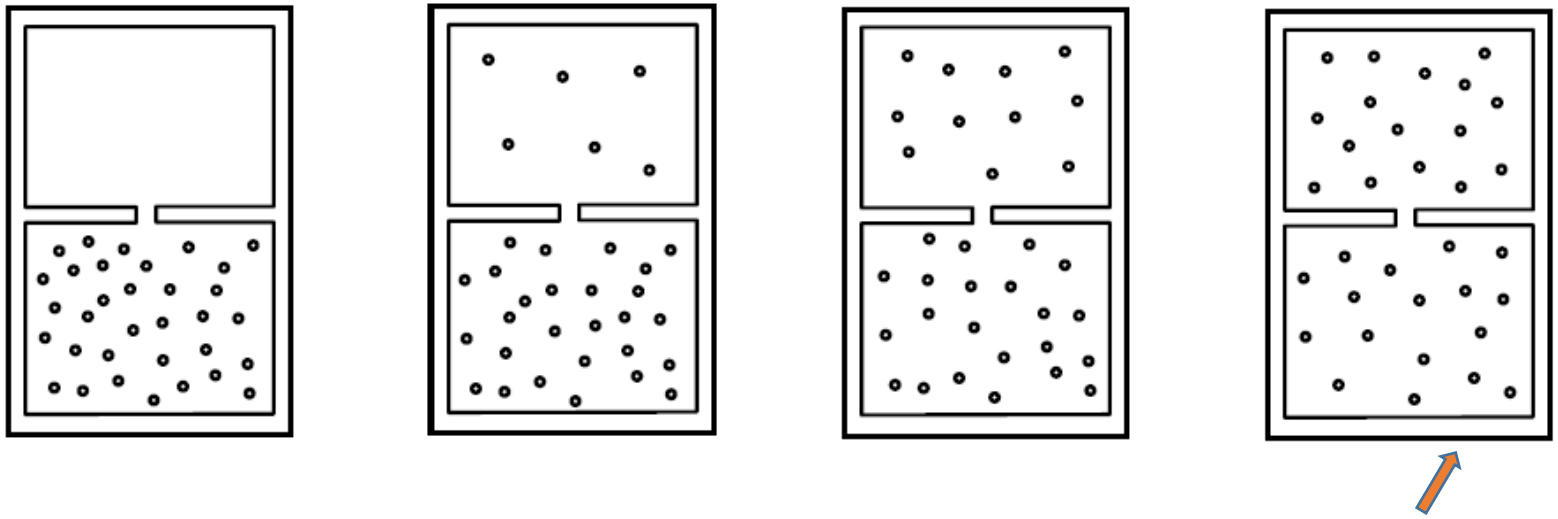
  - *In particular:* To provide a *micro-dynamical* explanation of the Second Law



Ludwig Boltzmann

Boltzmann's Approach:

• Consider different "macro-states" of a gas:



Thermodynamic equilibrium macro-state =constant thermodynamic properties (temperature, volume, pressure, etc.)

• *Why does the gas prefer to be in the equilibrium macro-state (last one)?*

Entropy in Statistical Mechanics

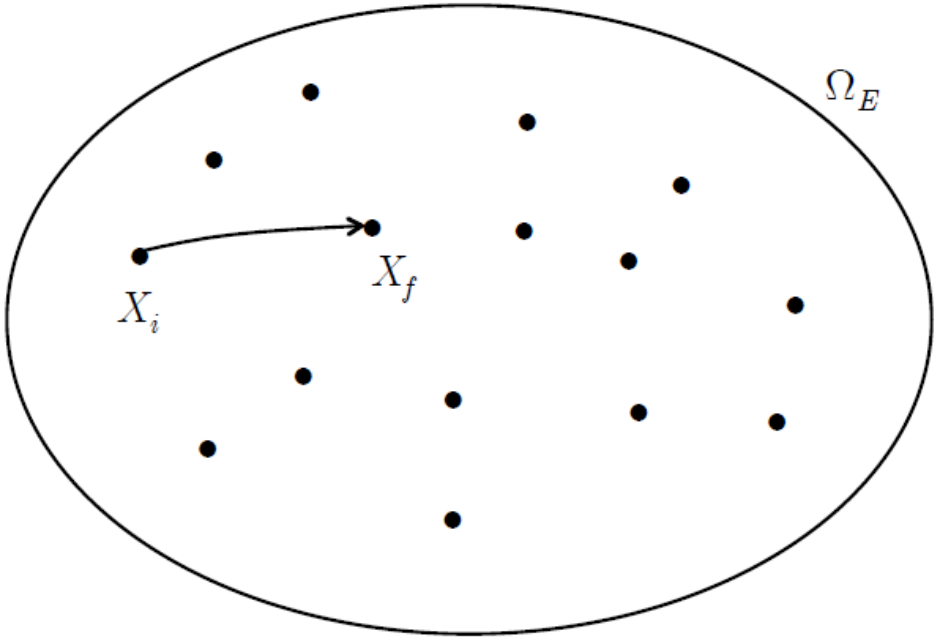
• Suppose the gas consists of *N* identical particles governed by Hamilton's equations of motion (the micro-dynamics).

Def. 1. A *micro-state X* of a gas is a specification of the position (3 degrees of freedom) and momentum (3 degrees of freedom) for each of its *N* particles.

Let Ω = *phase space* = 6*N-dim* space of all possible micro-states.
Let Ω_E = region of Ω that consists of all micro-states with constant energy *E*.

Hamiltonian dynamics maps initial micro-state X_i to final micro-state X_f.

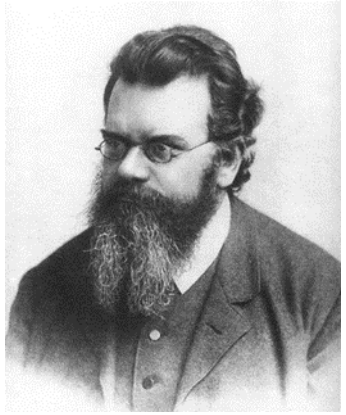
Can Second Law be explained by recourse to this dynamics?



Entropy in Statistical Mechanics

Def. 2. A *macro-state* Γ of a gas is a specification of the gas in terms of macroscopic properties (pressure, temperature, volume, etc.).

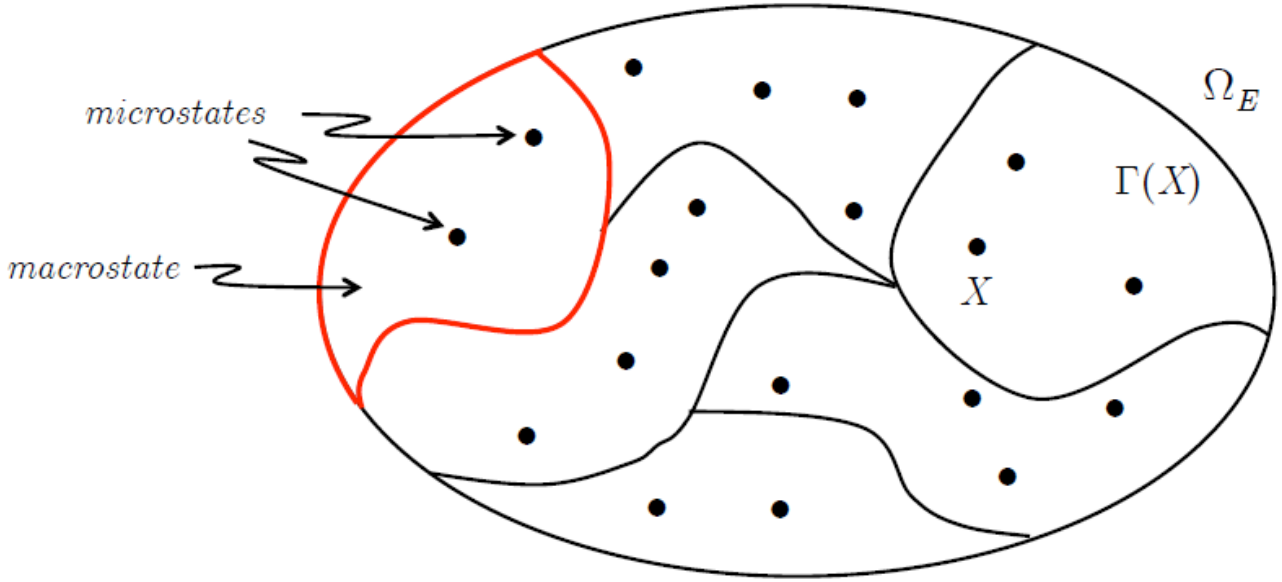
• *Relation between micro-states and macro-states:*



Macro-states supervene on micro-states!

- *To each micro-state there corresponds exactly one macro-state.*
- *Many distinct micro-states can correspond to the same macro-state.*

• *So:* Ω_E is partitioned into a finite number of regions corresponding to macro-states, with each micro-state X belonging to one macro-state $\Gamma(X)$.

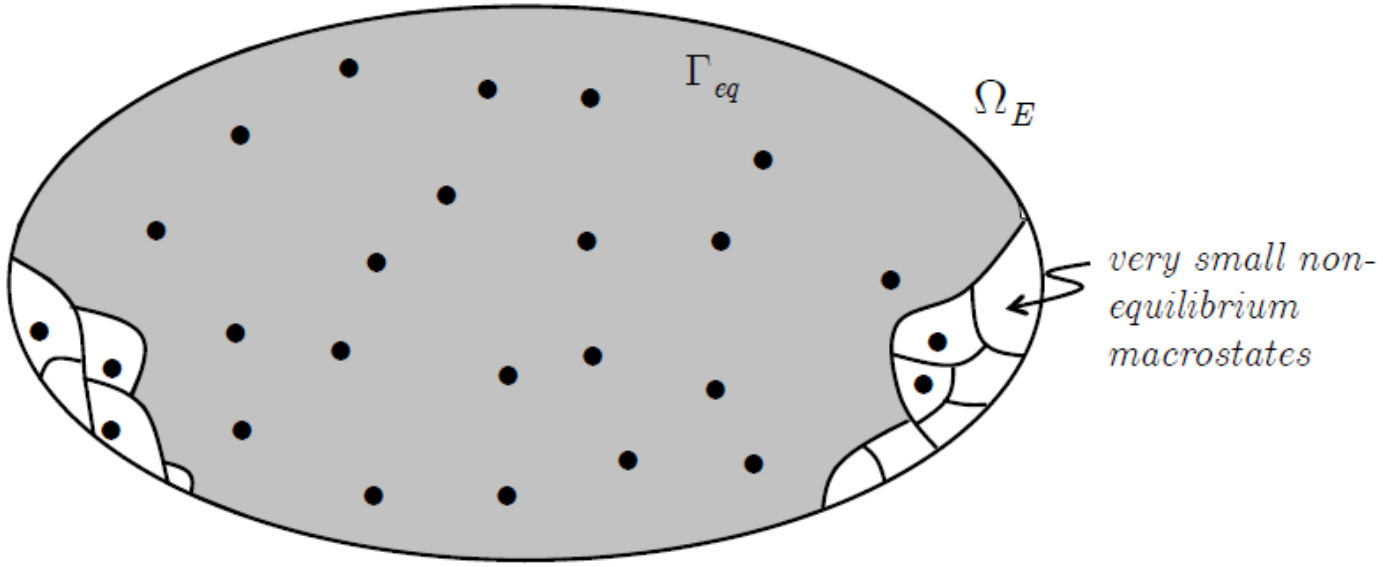


Entropy in Statistical Mechanics

Boltzmann's Claim: The equilibrium macro-state Γ_{eq} is vastly larger than any other macro-state (so it contains the vast majority of possible micro-states).

So: $S_B(\Gamma(X))$ is *a measure of the size* of $\Gamma(X)$.
And: $S_B(\Gamma(X))$ obtains its maximum value for Γ_{eq}.

Def. 3. The Boltzmann Entropy is defined by $S_B(\Gamma(X)) = k \log |\Gamma(X)|$ where $|\Gamma(X)|$ is the volume of $\Gamma(X)$.



- *Thus:* S_B increases over time because, for any initial micro-state X_i, the dynamics will map X_i into Γ_{eq} very quickly, and then keep it there for an extremely long time.

Entropy in Statistical Mechanics

Two Ways to Explain the Approach to Equilibrium:

(a) Appeal to Typicality (Goldstein 2001)

Claim: A system approaches equilibrium because equilibrium micro-states are *typical* and non-equilibrium micro-states are *atypical*.

• *Why?* For large N, Ω_E is almost entirely filled up with equilibrium micro-states. Hence they are "typical".

- But: What about the *dynamics* that evolves atypical states to typical states?
- "If a system is in an atypical micro-state, it does not evolve into an equilibrium micro-state *just because* the latter is typical." (*Frigg 2009*)
- Need to identify properties of the dynamics that guarantee atypical states evolve into typical states.
- *And:* Need to show that these properties are typical.

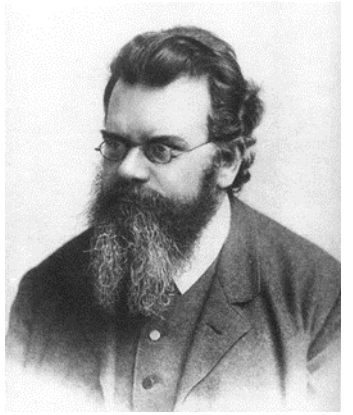
Frigg, Roman (2009), "*Probability in Boltzmannian Statistical Mechanics*", in Gerhard Ernst and Andreas Hu¨ttemann (eds.), Time, Chance and Reduction: Philosophical Aspects of Statistical Mechanics. Cambridge: Cambridge University Press.
Goldstein, Sheldon (2001), "*Boltzmann's Approach to Statistical Mechanics*", in Jean Bricmont, Detlef Durr, Maria Carla Galavotti, Gian Carlo Ghirardi, Francesco Petruccione, and Nino Zanghi(eds.), Chance in Physics: Foundations and Perspectives. Berlin: Springer, 39–54.

Entropy in Statistical Mechanics

(b) Appeal to Probabilities

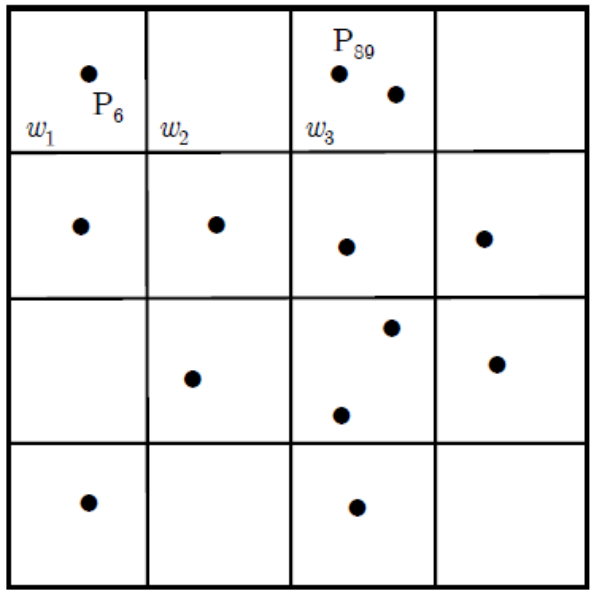
Claim: A system approaches equilibrium because it evolves from states of lower toward states of higher probability, and the equilibrium state is the state of highest probability.

- Associate probabilities with macro-states: the larger the macro-state, the greater the probability of finding a micro-state in it.



"In most cases, the initial state will be a very unlikely state. From this state the system will steadily evolve towards more likely states until it has finally reached the most likely state, i.e., the state of thermal equilibrium."

Entropy in Statistical Mechanics



Initial Arrangement: State of P_6 in ω_1, state of P_{89} in ω_3, etc.

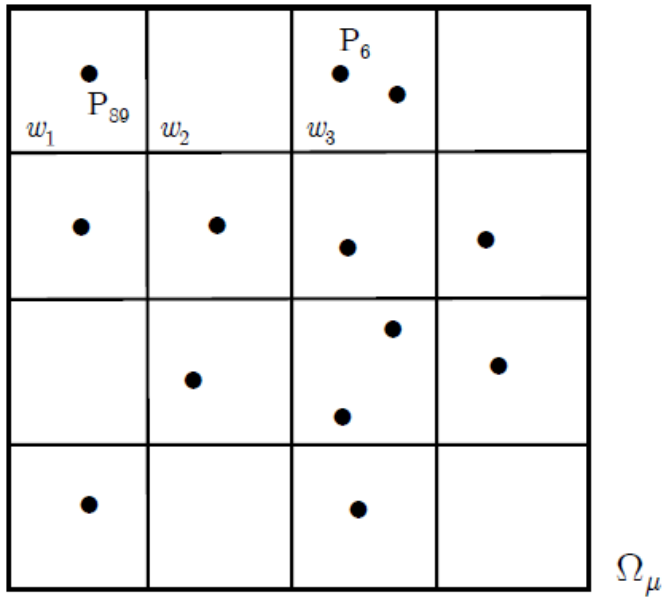
- Start with the *6-dim* phase space Ω_μ of a single particle.
- Partition Ω_μ into ℓ cells $\omega_1, \omega_2, \ldots, \omega_\ell$ of size $\delta\omega$.
- A state of an N-particle system is given by N points in Ω_μ.

Ω_E = *N copies of* Ω_μ

point in Ω_μ = single particle micro-state.

Def. 4. An arrangement is a specification of *which* points lie in which cells.

Entropy in Statistical Mechanics



Initial Arrangement: State of P_6 in ω_1, state of P_{89} in ω_3, etc.

Arrangement 2: State of P_{89} in ω_1, state of P_6 in ω_3, etc.

Distribution: (1, 0, 2, 0, 1, 1, ...)

Takes form $(n_1, n_{j2}, \ldots, n_l)$, where n_j = # of points in ω_j.

- Start with the *6-dim* phase space Ω_μ of a single particle.
- Partition Ω_μ into ℓ cells $\omega_1, \omega_2, \ldots, \omega_\ell$ of size $\delta\omega$.
- A state of an *N*-particle system is given by *N* points in Ω_μ.

Ω_E = *N copies of* Ω_μ

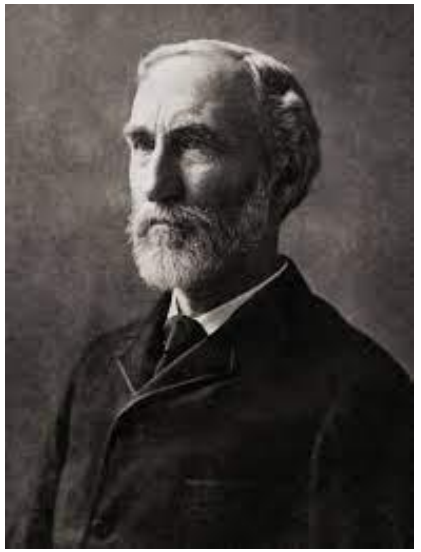
point in Ω_μ = single particle micro-state.

Def. 4. An arrangement is a specification of *which* points lie in which cells.

Def. 5. A *distribution* is a specification of how *many points* (regardless of *which* ones) lie in each cell.

- *Note:* More than one arrangement can correspond to the same distribution.

Willard Gibbs



Feb 11, 1839 – Apr 28, 1903

A Europe-trained American scientist at Yale College. His work on statistical mechanics provided a mathematical framework for quantum theory and for Maxwell's theories.

In 1870's, Gibbs gave a general expression of entropy S for a thermodynamic system

$$S = -k_B \sum_j p_j \log p_j$$

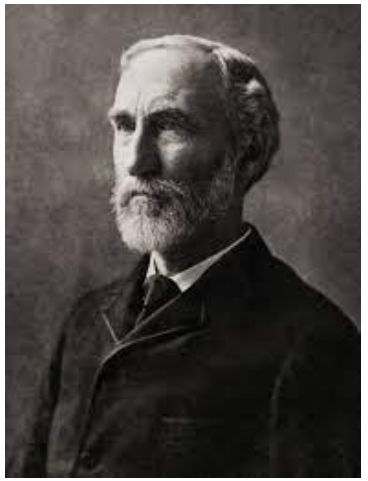
Entropy in Statistical Mechanics

- How many arrangements $G(D_i)$ are compatible with a given distribution $D_i = (n_1, n_2, \ldots, n_l)$?

- *Answer:* $$G(D_i) = \frac{N!}{n_1! n_2! \ldots n_l!}$$

Check: Let $D_1 = (N, 0, \ldots, 0)$ and $D_2 = (N-1, 1, 0, \ldots, 0)$.
- $G(D_1) = N!/N! = 1$. (Only one way for all N particles to be in ω_1.)
- $G(D_2) = N!/(N-1)! = N(N-1)(N-2)\ldots 1/(N-1)(N-2)\ldots 1 = N$.
  (There are N different ways ω_2 could have one point in it; namely, if P_1 was in it, or if P_2 was in it, or if P_3 was in it, etc...)

"The probability of this distribution $[D_i]$ is then given by the number of permutations of which the elements of this distribution are capable, that is by the number $[G(D_i)]$. As the most probable distribution, i.e., as the one corresponding to thermal equilibrium, we again regard that distribution for which this expression is maximal..."



- *Again:* The *probability* of a distribution D_i is given by $G(D_i)$.

Entropy in Statistical Mechanics

- *And:* Each distribution D_i corresponds to a macro-state Γ_{D_i}.

> *Why?* Because a system's macroscopic properties (volume, pressure, temp, *etc*) only depend on *how many* particles are in particular micro-states, and not on *which* particles are in which microstates.

- *What is the size of this macro-state?*

> - A point in Ω_E corresponds to an arrangement of Ω_μ.
> - The size of a macro-state Γ_{D_i} in Ω_E is given by the number of points it contains (the number of arrangements compatible with D_i) multiplied by a *volume element* of Ω_E.
> - A volume element of Ω_E is given by N copies of a volume element $\delta\omega$ of Ω_μ.

- *So:* The size of Γ_{D_i} is $|\Gamma_{D_i}| = \left| \begin{pmatrix} number\ of \\ arrangements \\ compatible\ with\ D_i \end{pmatrix} \times \begin{pmatrix} volume\ element \\ of\ \Omega_E \end{pmatrix} \right| = G(D_i)\ \delta\omega^N$

> *In other words:* The probability $G(D_i)$ of a distribution D_i is proportional to the size of its corresponding macro-state Γ_{D_i}.
> - *The equilibrium macro-state, being the largest, is the most probable; and a system evolves from states of low probability to states of high probability.*

Entropy in Statistical Mechanics

- *And:* Each distribution D_i corresponds to a macro-state Γ_{D_i}.

> *Why?* Because a system's macroscopic properties (volume, pressure, temp, *etc*) only depend on *how many* particles are in particular micro-states, and not on *which* particles are in which microstates.

- *What is the size of this macro-state?*

> - A point in Ω_E corresponds to an arrangement of Ω_μ.
> - The size of a macro-state Γ_{D_i} in Ω_E is given by the number of points it contains (the number of arrangements compatible with D_i) multiplied by a *volume element* of Ω_E.
> - A volume element of Ω_E is given by N copies of a volume element $\delta\omega$ of Ω_μ.

- *So:* The size of Γ_{D_i} is $|\Gamma_{D_i}| = \left| \begin{pmatrix} number\ of \\ arrangements \\ compatible\ with\ D_i \end{pmatrix} \times \begin{pmatrix} volume\ element \\ of\ \Omega_E \end{pmatrix} \right| = G(D_i)\ \delta\omega^N$

- The Boltzmann entropy of Γ_{D_i} is given by:

$$S_B(\Gamma_{D_i}) = k \log\left(G(D_i)\ \delta\omega^N\right)$$
$$= k \log\left(G(D_i)\right) + Nk \log(\delta\omega)$$
$$= k \log\left(G(D_i)\right) + constant$$

S_B is a measure of how large a macro-state is, and thus how probable the corresponding distribution of micro-states is.

Entropy in Statistical Mechanics

Other formulations of S_B

$$S_B(\Gamma_{D_i}) = k\log\big(G(D_i)\big) + constant$$
$$= k\log\left(\frac{N!}{n_1!\,n_2!\,\ldots n_l!}\right) + constant$$
$$= k\log(N!) - k\log(n_1!) - \ \ldots - k\log(n_l!) + constant$$
$$\approx (Nk\log N - N) - (n_1 k\log n_1 - n_1) - \ \ldots - (n_l k\log n_l - n_l) + constant$$
$$= -k\sum_{j=1}^{l} n_j \log n_j + constant$$

S_B in terms of micro-state occupation numbers n_j.

Let $p_j = \frac{n_j}{N}$ = probability of finding a randomly chosen micro-state in cell ω_j

Probabilities for *micro-states*, not macro-states/distributions!

$$S_B(\Gamma_{D_i}) = -Nk\sum_{j=1}^{l} p_j \log p_j + constant$$

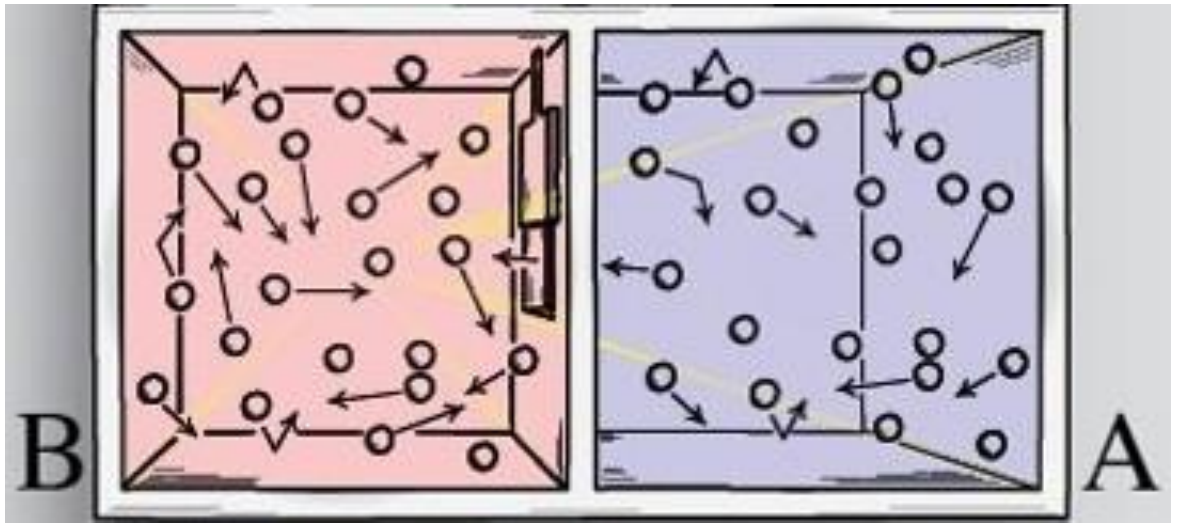
S_B in terms of micro-state probabilities p_j.

Entropy in Statistical Mechanics

Example 1: Uniform density

Problem:

An *N*-particle gas put in two chambers A and B with equal volume. Each particle in the *N*-particle gas may move freely between the two chambers. Since these chambers have equal volumes, the probability that any one particle resides in any one chamber is 1 out of 2 or 1/2. In this way, each distinct division of the *N* distinguishable particles between the two chambers models a single microstate each one of which is equiprobable. Show that the most probable macrostate is one in which the two chambers contain equal numbers of particles.



Entropy in Statistical Mechanics

Solution:

The probability $P(n)$ of a macrostate with n classical, and so distinguishable, particles in the rightmost chamber and $N - n$ particles in the leftmost chamber is

$$P(n) = \frac{N!}{n!\,(N-n)!}\left(\frac{1}{2}\right)^N$$

where the binomial coefficient $N!/\,[n!\,(N-n)!]$ is the multiplicity Ω of this macrostate and $(1/2)^N$ is the probability of any one microstate of this macrostate. Using Stirling's approximation $n! \sim n^n e^{-n}\sqrt{2\pi n}$, or $\ln n! \sim n \ln n - n$.

$$\ln P(n) = N \ln N - N - n \ln n + n - (N-n) \ln (N-n) + (N-n) - N \ln 2$$
$$= N \ln N - n \ln n - (N-n) \ln (N-n) - N \ln 2$$

Setting the derivative of $\ln P(n)$ with respect to n equal to zero produces

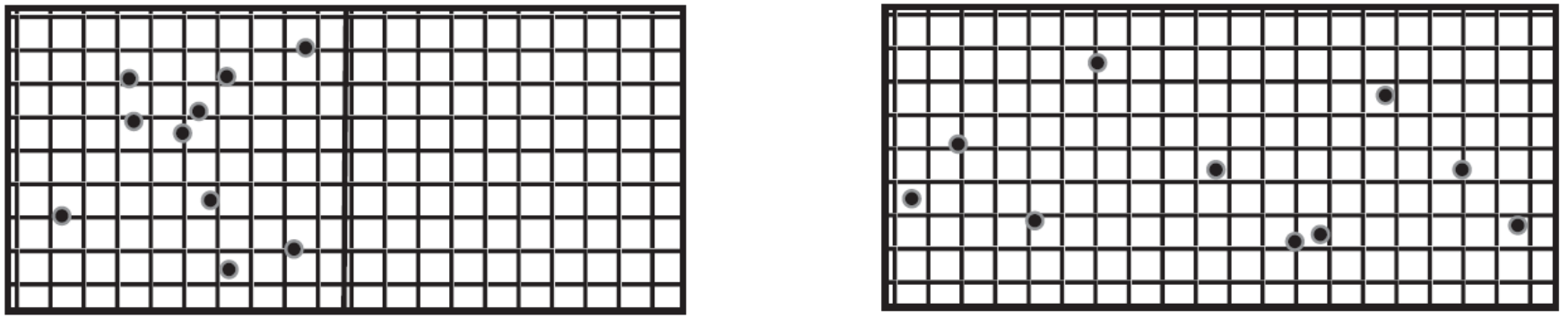
$$-\ln n + \ln(N-n) = 0 \;\; ==> n = N/2.$$

Entropy in Statistical Mechanics

Example 2: Joule Expansion

Problem:

Consider, as in Example 1, a system with two equal volume chambers containing a total of N identical but classical, and so distinguishable, particles. Suppose this system is initialized in a macrostate in which all its particles are contained within its left chamber. The particles are then allowed to move freely between both chambers by removing the barrier and achieve a final equilibrium macrostate in which the particles are distributed throughout both chambers as illustrated below. How much does the entropy of the system increase?



Entropy in Statistical Mechanics

Solution:

The number of spatial microstates available to a single particle increases by a factor of 2 during this process, and the number of spatial microstates available to two distinguishable particles increases by 2×2. The multiplicity of the N–particle system increases by 2^N during this Joule expansion. Consequently, the ratio of the final to the initial macrostate multiplicity is

$$\frac{\Omega_f}{\Omega_i} = 2^N,$$

and the entropy is incremented by

$$\Delta S = S_f - S_i = k \ln\left(\frac{\Omega_f}{\Omega_i}\right) = Nk \ln 2$$

Therefore, each particle appear to contribute $k \ln 2$ to the entropy increment.

Entropy in Statistical Mechanics

Example 3: Entropy of Mixing

Problem:

A system consisting of $2N$ classical particles is initialized in a macrostate in which N particles of one kind, e.g., nitrogen molecules, are all in the left of two equal-volume chambers and N particles of another kind, e.g., oxygen molecules, are all in the right chamber. The particles are allowed to mix by moving the barrier. What is the entropy increment?

Solution:

The microstates available to each kind of particle are doubled during this process which increases the number of microstates available to the system in its final macrostate. Therefore, the initial and final multiplicities of the $2N$ – particle system are in the ratio

$$\frac{\Omega_f}{\Omega_i} = 2^{2N},$$

and the entropy is incremented by

$$\Delta S = S_f - S_i = k \ln\left(\frac{\Omega_f}{\Omega_i}\right) = 2Nk \ln 2$$

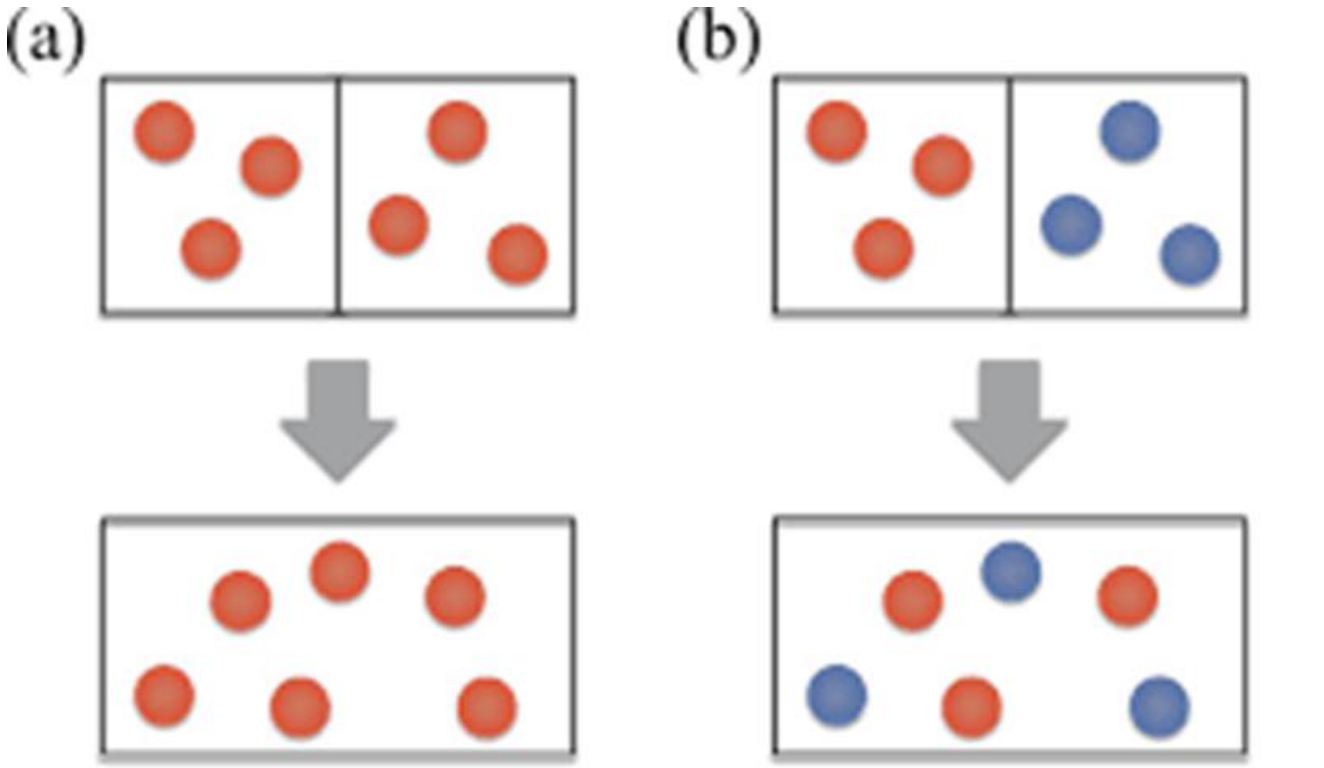
Entropy in Statistical Mechanics

Example 4: Entropy of Mixing Again

Problem:
Consider now a system of 2 N identical, classical particles initialized with N particles in the left of two equal-volume chambers and N in the right. The particles are then allowed to move freely by removing the barrier and achieve a final, equilibrium state. What is the entropy increment?

Paradox:
Two observers calculating the entropy increase during the process disagree depending on their ability to distinguish the particles. An *informed* observer, who can measure the difference between the gases, calculates $2N \ln 2$ (from the above example), while an *uninformed* observer records no entropy change. This is the so-called ***Gibbs Paradox***.

Entropy in Statistical Mechanics



⟶ The paradox is resolved by concluding the *indistinguishability* of the particles.

Recall: Entropy in Classical Information Theory

1. $H(X)$ as Maximum Amount of Message Compression

- Let $X = \{x_1, \ldots, x_\ell\}$ be a set of letters from which we construct the messages.
- Suppose the messages have N letters a piece.
- The probability distribution $P = (p_1, \ldots, p_\ell)$ is now over the letter set.

> *What this means*:
> - Each letter x_i has a probability of p_i of occuring in a message.
> - *In other words*: A typical message will contain $p_1 N$ occurrences of x_1, $p_2 N$ occurrences of x_2, *etc.*

- *Thus*:

$$\left(\begin{array}{l} \textit{The number of distinct} \\ \textit{typical messages} \end{array} \right) = \frac{N!}{(p_1 N)!(p_2 N)! \cdots (p_\ell N)!}$$

Number of ways to arrange N distinct letters into l bins with capacities p_1N, p_2N, …, p_lN.

- *So*:

$$\log_2 \left(\begin{array}{l} \textit{The number of distinct} \\ \textit{typical messages} \end{array} \right) = \log_2 \left(\frac{N!}{(p_1 N)!(p_2 N)! \cdots (p_\ell N)!} \right)$$

Comparison of Entropy in Classical Information Theory and Thermodynamics

How the message compression interpretation of H relate to S_B

Shannon

- N = # of letters in message.
- N-letter message.
- $\{x_1, \ldots, x_\ell\}$ = ℓ-letter alphabet.
- $(p_1, \ldots, p_\ell)$ = probability distribution over letters.
- p_j = probability that x_j occurs in a given message.
- Np_j = # of x_j's in typical message.

$$H(X) = -\sum_{j=1}^{\ell} p_j \log_2 p_j$$

- NH = minimum number of base 2 numerals ("bits") needed to encode a message composed of N letters drawn from set X.

Boltzmann

- N = # of single-particle microstates.
- N-microstate arrangement.
- $(n_1, \ldots, n_\ell)$ = ℓ-cell distribution.
- $(p_1, \ldots, p_\ell)$ = probabilty distribution over microstates.
- $p_j = n_j/N$ = prob that a w_j-micro-state occurs in a given arrangement.
- Np_j = # of w_j-microstates in arrangement.

$$S_B(\Gamma_{D_i}) = -Nk\sum_{j=1}^{\ell} p_j \ln p_j + const.$$

- $S_B \sim NH$ = minimum number of base c numerals ("c-bits?") needed to encode an arrangement of N single-particle microstates.

Recall: Entropy in Classical Information Theory

2. H(X) as a Measure of Uncertainty

• Suppose $P = (p_1, \dots, p_l)$ is a probability distribution over a set of values $\{x_1, \dots, x_l\}$ of a random variable X.

Def. 1. The *expected value* $E[X]$ of X is given by $E[X] = \sum_{j=1}^{l} p_j x_j$.

Def. 2. The *information gained* if X is measured to have the value x_j is given by $-\log_2 p_j$.

- *Motivation:* The greater p_j is, the more certain x_j is, and the less information should be associated with it.

• Then the expected value of $-\log_2 p_j$ is just the Shannon information:

$$E(-\log_2 p_j) = \sum_{j=1}^{l} p_j \log_2 p_j = H[X]$$

• *What this means*:

$H[X]$ tells us our expected *information gain* upon measuring X.

Comparison of Entropy in Classical Information Theory and Thermodynamics

How does the uncertainty interpretation of H relate to S_B

Shannon

- X = random variable.
- $\{x_1, \ldots, x_\ell\} = \ell$ values.
- $(p_1, \ldots, p_\ell)$ = probabilty distribution over values.
- p_j = probability that X has value x_j upon measurement.
- $-\log_2 p_j$ = *information* gained upon measurement of X with outcome x_j.

$$H(X) = -\sum_{j=1}^{\ell} p_j \log_2 p_j$$

- $H(X)$ = expected information gain upon measurement of X.

Boltzmann

- X = single-particle microstate.
- $(n_1, \ldots, n_\ell) = \ell$-cell distribution.
- $(p_1, \ldots, p_\ell)$ = probabilty distribution over microstates.
- $p_j = n_j/N$ = probability that a microstate occurs in cell w_j.
- $-\ln p_j$ = *information* gained upon measurement of particle to be in microstate in cell w_j.

$$S_B(\Gamma_{D_i}) = -Nk\sum_{j=1}^{\ell} p_j \ln p_j + const.$$

- S_B/N = expected information gain upon determining the microstate of a particle.

Short Summary

- Entropy is a state function in thermodynamics
- Entropy as an arrow of time
- Entropy as a measure of order and disorder is suggestive, but can mislead
- Entropy as a measure of missing information/uncertainty
- Entropy is an additive measure of the number of possibilities available to a system.

Some other types of entropy

- Relative Entropy
- Rényi Entropy
- Tsallis Entropy
- Transfer Entropy
- Kolmogorov Entropy
- Von Neumann Entropy
- ……

Relative entropy (Kullback-Leibler divergence)

Recall that mutual information $I\,[X\colon Y]$ measures the dependency between two random variables X and Y. ***Kullback–Leibler Divergence*** is a generalization of this concept, which measures *how far* a distribution q is away from the actual distribution p. For discrete probability distributions p and q defined on the same probability space, χ, the Kullback–Leibler divergence from q to p is defined to be

$$D(p||q) = \sum_{x\in\chi} p(x)\log\left(\frac{p(x)}{q(x)}\right) = -\sum_{x\in\chi} p(x)\log\left(\frac{q(x)}{p(x)}\right)$$

- The Kullback–Leibler divergence is defined only if for all x, $q(x) = 0$ implies $p(x) = 0$ (absolute continuity). Whenever $p(x)$ is zero the contribution of the corresponding term is interpreted as zero because $\lim_{x\to 0^+} x\log(x) = 0$.

- $I\,[X\colon Y] = D(p(x,y)||p(x)p(y))$ *(i.e., mutual information is the relative entropy between the joint distribution and the product of the distributions.)*

- Kullback–Leibler divergence is ***NOT*** symmetric: $D(p||q) \neq D(q||p)$

Relative entropy (Kullback-Leibler divergence)

For continuous random variable,

$$D(p||q) = \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$$

where p and q now denote the probability densities. In many respects it acts as a measure of dissimilarity or "distance" between distributions.

Note: Beyond Kullback–Leibler divergence, there are other kinds of information divergence that measures the difference between probability distributions, e.g., *f*-divergences, Bregman divergences, etc.

Relative entropy (Kullback-Leibler divergence) (cont'd)

The Kullback-Leibler divergence is always non-negative, i.e., $D(p||q) \geq 0$, a result known as *Gibb's inequality*. It is equal to zero if and only if $p = q$ for all $x \in \chi$. (** Use *Jensen's Inequality* to prove that the Kullback-Leibler divergence is always non-negative.)

$D(p||q)$ is convex in the pair of (p, q). (**Use *Log Sum Inequality* to prove.)

In their original paper, Kullback and Leibler cast their divergence measure as a tool for distinguishing between statistical populations. They refer to the quantity

$$\log\left(\frac{p(x)}{q(x)}\right)$$

as "*the information in x for discrimination between*" the distributions p and q. Their divergence is then the mean information for discrimination per observation from p. In many respects it acts as a measure of dissimilarity or "distance" between distributions.

Kullback, S.; Leibler, R.A. (1951). "*On information and sufficiency*". Annals of Mathematical Statistics. 22 (1): 79–86.

Relative entropy (Kullback-Leibler divergence) (cont'd)

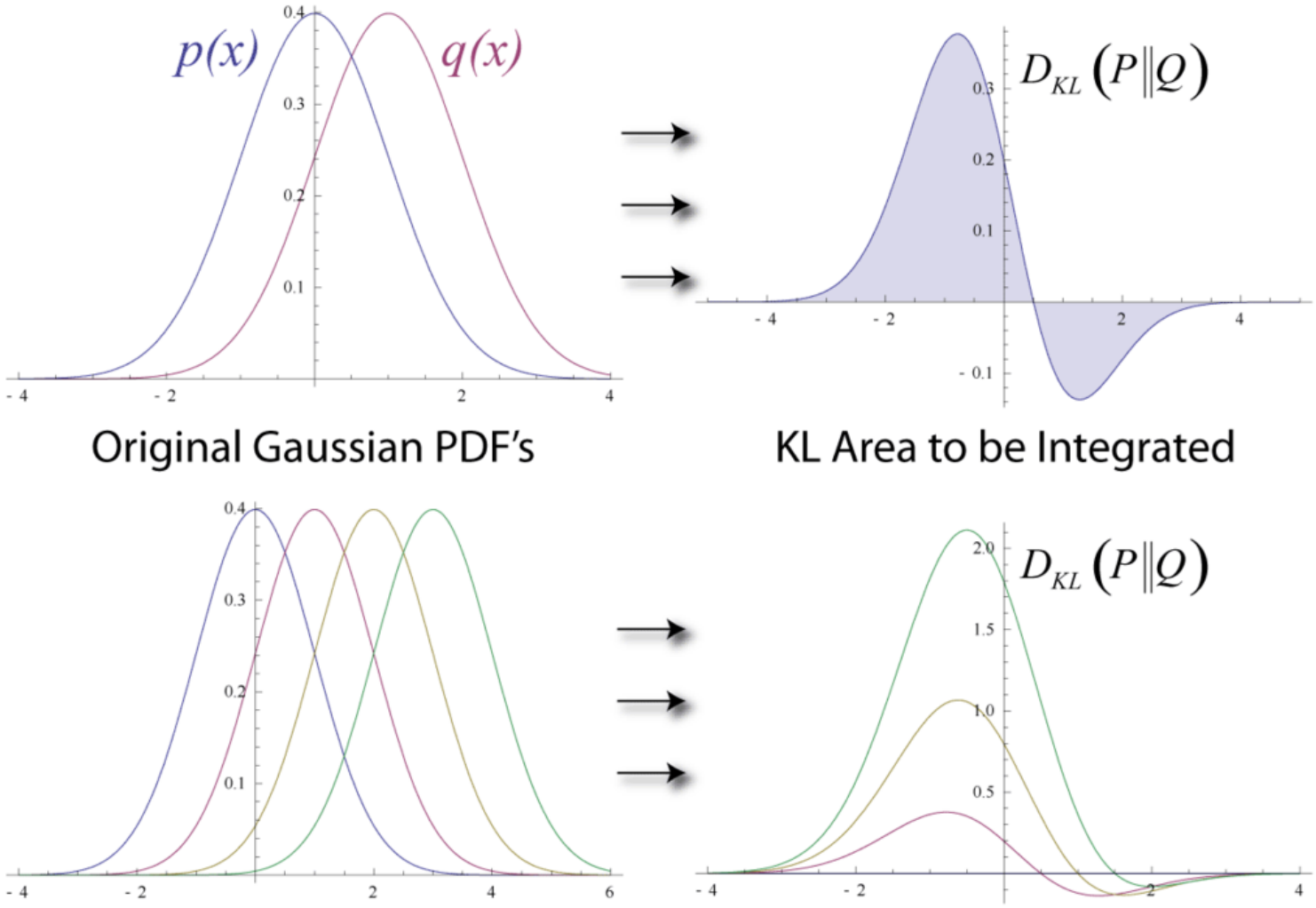


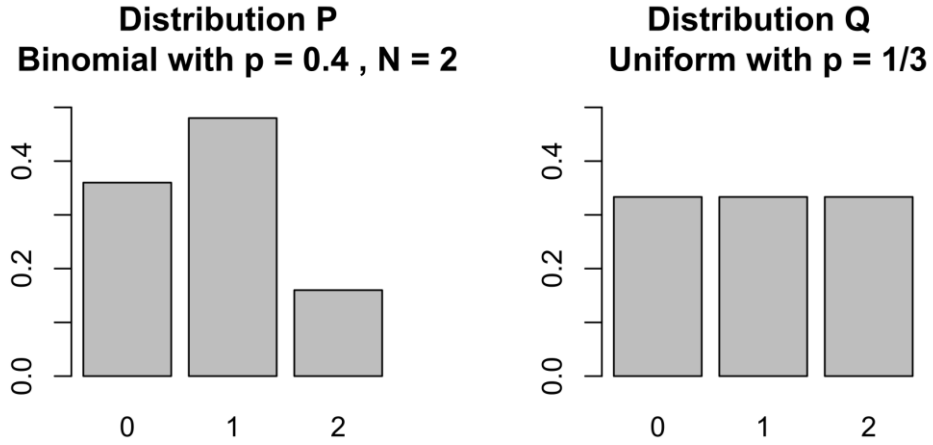
Illustration of the *Kullback–Leibler* (*KL*) divergence for two normal distributions. The asymmetry for the *KL* divergence is clearly visible.

Relative entropy (Kullback-Leibler divergence) (cont'd)

A Simple Example:
(**This example uses the natural log with base e, designated* ln *to get results*)

$$D(p||q)$$
$$= \sum_{x \in \chi} p(x) \ln\left(\frac{p(x)}{q(x)}\right) = 0.36 \ln\left(\frac{0.36}{0.333}\right) + 0.48 \ln\left(\frac{0.48}{0.333}\right)$$
$$+ 0.16 \ln\left(\frac{0.16}{0.333}\right) = 0.0852996$$

$$D(q||p)$$
$$= \sum_{x \in \chi} q(x) \ln\left(\frac{q(x)}{p(x)}\right) = 0.333 \ln\left(\frac{0.333}{0.36}\right) + 0.333 \ln\left(\frac{0.333}{0.48}\right)$$
$$+ 0.333 \ln\left(\frac{0.333}{0.16}\right) = 0.097455$$



x	0	1	2
Distribution *P*(x)	0.36	0.48	0.16
Distribution *Q*(x)	0.333	0.333	0.333

Kullback, S. (1959), *Information Theory and Statistics*, John Wiley & Sons.

A Derivation of Relative Entropy

Suppose you are observing a random variable X, and have a theory that predicts its probability distribution Q_X. Let the probability to observe $X = x_i$ to be $q_i = Q_X(x_i)$, which differs from the true probability distribution P_X to observe $X = x_i$ with $p_i = P_X(x_i)$. After observing N times, how sure could you be that the initial theory is wrong?

If the correct distribution is P_X, after observing N time, the probability of the outcome $X = x_i$ is approximately p_iN times. Our theory will have a probability of

$$Pr = \prod_{i=1}^{N} {q_i}^{p_i N} \frac{N!}{\prod_{j=1}^{N}(p_i N)!} \sim 2^{-N\sum_i p_i(\log p_i - \log q_i)} = 2^{-ND(P_X||Q_X)} \quad (*)$$

(Recall Stirling's formula, $N! \approx N \ln N - N$). The relative entropy (Kullback-Liebler divergence) is defined as

$$D(P_X||Q_X) = \sum_i p_i(\log\ p_i - \log q_i)$$

From (*), it is clear that $D(P_X||Q_X)$ is non-negative, and equals zero only if $P_X = Q_X$.

If the initial hypothesis is wrong, we will be sure of this once

$$N\ D(P_X||Q_X) \gg 1$$

The probability of falsely excluding a correct hypothesis due to a large fluctuation that causes the data to be more accurately simulated by P_X than by Q_X, decays for large N as

$$2^{-ND(P_X||Q_X)}$$

when N is large.

** Notice that we have ignored noise in the observations here.

Now, consider a pair of random variables X, Y and we consider two different probability distribution functions $P_{X,Y}(x_i,y_i)$ and $Q_{X,Y}(x_i,y_i)$. The separate distributions for X and Y are

$$P_X(x_i) = \sum_j P_{X,Y}(x_i, y_j),\ P_Y(y_j) = \sum_i P_{X,Y}(x_i, y_j)$$

We define

$$Q_{X,Y}(x_i, y_j) = P_X(x_i)\, P_Y(y_j)$$

We now calculate the relative entropy of these two distributions

$$D(P_{X,Y}||Q_{X,Y}) = \sum_{i,j} P_{X,Y}(x_i, y_j)\left(\log P_{X,Y}(x_i, y_j) - \log P_X(x_i)\, P_Y(y_j)\right)$$

$$= \sum_{i,j} P_{X,Y}(x_i, y_j)\left(\log P_{X,Y}(x_i, y_j) - \log P_X(x_i) - \log P_Y(y_j)\right) = H[X] + H[Y] - H[X,Y]$$

$$= I[X:Y]$$

Thus, $I[X:Y] \geq 0$, with equality iff X and Y are independent of each other. The property $H[X] + H[Y] - H[X,Y] \geq 0$ is called ***Subadditivity of entropy***.

Relative entropy (Kullback-Leibler divergence) (cont'd)

Chain Rule for Relative Entropy

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

Proof:

$$D(p(x,y)||q(x,y)) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} = \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)}$$

$$= \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)}$$

$$= D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

Relative entropy (Kullback-Leibler divergence) (cont'd)

Monotonicity of Relative Entropy

Again consider a pair of random variables X, Y and we consider two different probability distribution functions $P_{X,Y}(x_i,y_i)$ and $Q_{X,Y}(x_i,y_i)$. Start with a hypothesis $Q_{X,Y}$ for the joint probability. After many trials in which we observe X and Y, the chance that we are wrong will be controlled by $D(P_{X,Y}||Q_{X,Y})$. Suppose we only observe X, the reduced distributions P_X and Q_X for X only are given by

$$P_X(x_i) = \sum_j P_{X,Y}(x_i, y_j), \qquad Q_X(x_i) = \sum_j Q_{X,Y}(x_i, y_j).$$

Now, after many trials in which we observe X only, the chance that we are wrong will be controlled by $D(P_X||Q_X)$.

Relative entropy (Kullback-Leibler divergence) (cont'd)

Monotonicity of Relative Entropy (cont'd)

It is harder to disprove the initial hypothesis if we observe only X, therefore

$$D(P_{X,Y}||Q_{X,Y}) \geq D(P_X||Q_X).$$

This is called ***Monotonicity of Relative Entropy***.

Relative entropy (Kullback-Leibler divergence) (cont'd)

Monotonicity of Relative Entropy (cont'd)

Assuming we now observe a sequence $x_{i1}, x_{i2}, \dots, x_{iN}$ in N trials. To estimate how unlikely this is, we can imagine a sequence of y's that minimizes the unlikelihood of the joint sequence,

$$(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \dots, (x_{iN}, y_{iN}) .$$

Any real sequence of y's that we may observe can only be more unlikely than this. Thus, observing Y as well as X can only increase our estimate of how unlikely the outcome was, given the sequence of x's. Hence, the relative entropy decreases upon "integrating out" some of the variables.

Relative entropy (Kullback-Leibler divergence) (cont'd)

Monotonicity of Relative Entropy (cont'd)

For a more formal proof of the inequality, rewrite $D(P_{X,Y}||Q_{X,Y}) \geq D(P_X||Q_X)$ as

$$\sum_{i,j} P_{X,Y}(x_i, y_j)\left(\log\left(\frac{P_{X,Y}(x_i, y_j)}{Q_{X,Y}(x_i, y_j)}\right) - \log\left(\frac{P_X(x_i)}{Q_X(x_i)}\right)\right) \geq 0\,. \qquad (1)$$

Or,

$$\sum_i P_X(x_i) \sum_j \frac{P_{X,Y}(x_i, y_j)}{P_X(x_i)}\left(\log\left(\frac{P_{X,Y}(x_i, y_j)/P_X(x_i)}{Q_{X,Y}(x_i, y_j)/Q_X(x_i)}\right)\right) \geq 0\,. \qquad (2)$$

Now, define

$$P_{Y|X=x_i}(y_j) = \frac{P_{X,Y}(x_i, y_j)}{P_X(x_i)}\,, Q_{Y|X=x_i}(y_j) = \frac{Q_{X,Y}(x_i, y_j)}{Q_X(x_i)}\,.$$

Equation (2) becomes, $\sum_i P_X(x_i)\, D(P_{Y|X=x_i}||Q_{Y|X=x_i})$, which is a sum of positive terms and so this establishes the *monotonicity of relative entropy*.

Relative entropy (Kullback-Leibler divergence) (cont'd)

Monotonicity of Relative Entropy (cont'd)

An important special case is the ***Strong Subadditivity of Entropy***. Consider three random variables *X*, *Y*, *Z*. The combined system has a joint probability distribution $P_{X,Y,Z}(x_i, y_j, z_k)$. We define another probability distribution $Q_{X,Y,Z}$ by neglecting the correlations between *X* and *YZ*,

$$Q_{X,Y,Z}(x_i, y_j, z_k) = P_X(x_i) P_{Y,Z}(y_j, z_k),$$

where

$$P_X(x_i) = \sum_{j,k} P_{X,Y,Z}(x_i, y_j, z_k) \ , \qquad P_{Y,Z}(y_j, z_k) = \sum_{i} P_{X,Y,Z}(x_i, y_j, z_k) \ .$$

The relative entropy of the combined system is: $D(P_{X,Y,Z} || Q_{X,Y,Z})$.

Relative entropy (Kullback-Leibler divergence) (cont'd)

Monotonicity of Relative Entropy (cont'd)

Imagine we only observe the system XY, we then replace $P_{X,Y,Z}$ and $Q_{X,Y,Z}$ by probability distributions $P_{X,Y}$ and $Q_{X,Y}$ with

$$P_{X,Y}(x_i, y_j) = \sum_k P_{X,Y,Z}(x_i, y_j, z_k)\,, Q_{X,Y}(x_i, y_j) = \sum_k Q_{X,Y,Z}(x_i, y_j, z_k) = P_X(x_i)P_Y(y_j)$$

We again define the relative entropy $D(P_{X,Y}||Q_{X,Y})$ and monotonicity of relative entropy gives

$$D(P_{X,Y,Z}||Q_{X,Y,Z}) \geq D(P_{X,Y}||Q_{X,Y})\,.$$

Relative entropy (Kullback-Leibler divergence) (cont'd)

Monotonicity of Relative Entropy (cont'd)

The relation between relative entropy and mutual information for this system gives

$$D(P_{X,Y,Z}||Q_{X,Y,Z}) = I[X:YZ] = H[X] - H[X,Y,Z] + H[Y,Z]$$

and

$$D(P_{X,Y}||Q_{X,Y}) = I[X:Y] = H[X] - H[X,Y] + H[Y].$$

Therefore

$$H[X] - H[X,Y,Z] + H[Y,Z] \geq H[X] - H[X,Y] + H[Y]$$

or

$$H[X,Y] + H[Y,Z] \geq H[Y] + H[X,Y,Z]\,,$$

which is called the ***Strong Subadditivity of Entropy***. **This statement turns out to be also true in quantum mechanics, which is both powerful and surprising.

Relative entropy (Kullback-Leibler divergence) (cont'd)

Monotonicity of Relative Entropy (cont'd)

From the above, we have

$$I[X:Y,Z] \geq I[X:Y] \,,$$

which is known as *monotonicity of mutual information* (refer to the p.28 of Lecture 9 on mutual information). The interpretation is that what one learns about a random variable X by observing both Y and Z is at least as much as one could learn by observing Y only.

Rényi entropy

In information theory, Rényi entropy generalizes Hartley entropy, Shannon entropy, collision entropy and min-entropy. Rényi entropy of order α, where $\alpha \geq 0$ and $\alpha \neq 1$, is defined as

$$H_\alpha(x) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} p_i^\alpha \right)$$

where x is a discrete random variable with possible outcomes 1, 2, …, n and corresponding probabilities $p_i = \Pr(x = i)$ for $i = 1, 2, .., n$.

The limit when $\alpha \to 1$ gives the Shannon entropy

$$H_1(x) = \lim_{\alpha \to 1} H_\alpha(x) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Rényi entropy (cont'd)

Example: Renyi entropies of a uniform distribution

Consider the probability density $\mu(x) = 1$ for $x \in [0, 1]$ and 0 elsewhere. Partitioning the unit interval into N intervals of length $\varepsilon = 1/N$ yields

$$\hat{H}_\alpha(\varepsilon) = (1 - \alpha)^{-1} \ln(N\, \varepsilon^\alpha) = -\ln \varepsilon = \ln N.$$

Therefore, all order-α entropies are the same == > a consequence of the homogeneity of the probability distribution. Its numerical value is the logarithm of the number of partition elements, which means that the better we resolve the real numbers by the partition the more information we gain.

Rényi entropy (cont'd)

Taking various values of α, will correspond to different entropies:

- $\alpha = 0$: (Hartley or max-entropy)

$$H_0(x) = \log n = \log|x|$$

- $\alpha = 1$: (Shannon entropy)

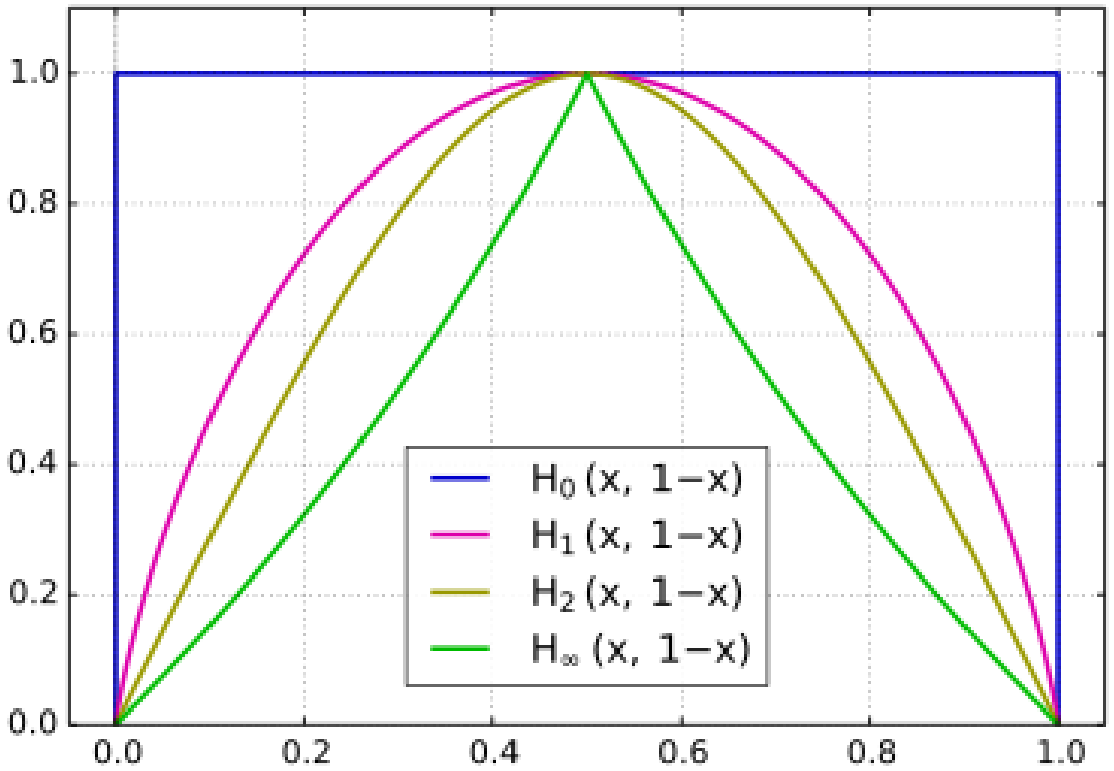
$$H_1(x) = \lim_{\alpha \to 1} H_\alpha(x) = -\sum_{i=1}^{n} p_i \log p_i$$

- $\alpha = 2$: (Collision or Rényi entropy)

$$H_2(x) = -\log \sum_{i=1}^{n} p_i^2$$

- $\alpha \to \infty$: (min-entropy)

$$H_\infty(x) = \min_i(-\log p_i) = -\left(\max_i \log p_i\right) = -\log \max_i p_i$$



Rényi entropy of a random variable with two possible outcomes against p_1, where $P = (p_1, 1 - p_1)$. Shown are H_0, H_1, H_2 and H_∞, in units of bits.

Rényi entropy (cont'd)

Preservation of Additivity for Independent Events

Consider two random variables x, y that are independent with possible outcomes 1, 2, …, n ; 1, 2, …, m and corresponding probabilities $p_i = \Pr(x = i)$ for $i = 1, 2, .., n$; $q_i = \Pr(y = j)$ for $j = 1, 2, .., m$, respectively . Then

$$H_\alpha(xy) = H_\alpha(x) + H_\alpha(y) .$$

Proof:

$$H_\alpha(xy) = \frac{1}{1-\alpha}\log\left(\sum_{i,j} p_i^\alpha q_j^\alpha\right) = \frac{1}{1-\alpha}\log\left((\sum_i p_i^\alpha)(\sum_j q_j^\alpha)\right)$$

$$= \frac{1}{1-\alpha}\left(\log\left(\sum_i p_i^\alpha\right) + \log\left(\sum_j q_j^\alpha\right)\right) = H_\alpha(x) + H_\alpha(y) .$$

Rényi entropy (cont'd)

$H_\alpha(x)$ is non-increasing in α for any given distribution of probabilities p_i

Proof:

$$-\frac{dH_\alpha(x)}{d\alpha} = \frac{1}{(1-\alpha)^2}\sum_{i=1}^{n} z_i \log\left(\frac{z_i}{p_i}\right), \qquad z_i = \frac{{p_i}^\alpha}{\sum_{j=1}^{n} {p_j}^\alpha} .$$

The right hand side of the derivative is proportional to Kullback-Leibler divergence ($D(z_i||p_i)$), and therefore is always non-negative. This can also be proven by using Jensen's Inequality in particular cases

$$\log n = H_0(x) \geq H_1(x) \geq H_2(x) \geq H_\infty(x) .$$

For values of α > 1, inequalities in the other direction also hold. In particular, one has

$$H_2(x) \leq 2H_\infty(x) .$$

Rényi entropy (cont'd)

One can also defined a spectrum of divergence measures generalizing the Kullback–Leibler divergence called *Rényi divergence*. The Rényi divergence of order α or alpha-divergence of a distribution p from a distribution q is defined to be

$$D_\alpha(p||q) = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{n}\frac{p_i^\alpha}{q_i^{\alpha-1}}\right)$$

where $0 < \alpha < \infty$ and $\alpha \neq 1$. One can define the Rényi divergence for the special values $\alpha = 0, 1, \infty$ by taking a limit, and in particular the limit $\alpha \to 1$ gives the Kullback–Leibler divergence.

Rényi entropy is important in ecology and statistics as *index of diversity*. The Rényi entropy is also important in quantum information, where it can be used as a measure of entanglement and the limit of the Rényi entropy as $\alpha \to 1$ is the ***von Neumann entropy***.

Tsallis entropy

In physics, the *Tsallis entropy* is a generalization of the standard Boltzmann–Gibbs entropy. Given a discrete set of probabilities $\{p_i\}$ with the condition $\sum_i p_i = 1$, and q any real number, the Tsallis entropy is defined as

$$H_q(x) = \frac{1}{1-q}\left(1 - \sum_{i=1}^{n} p_i^q\right)$$

where q is a real parameter and sometimes called *entropic index*. In the limit $q \to 1$, we recover the usual Boltzmann-Gibbs entropy,

$$H_B(x) = H_1(x) = -k \sum_i p_i \ln p_i$$

For continuous probability distributions, the Tsallis entropy takes the form,

$$H_q(x) = \frac{1}{1-q}\left(1 - \int (p(x))^q dx\right)$$

where $p(x)$ is a probability density function.

Tsallis entropy (cont'd)

The discrete Tsallis entropy satisfies

$$H_q(x) = -\lim_{x\to 1} D_q \sum_i p_i^x$$

where D_q is the q-derivative with respect to x, which can be compared to the standard entropy formula

$$H(x) = -\lim_{x\to 1} \frac{d}{dx} \sum_i p_i^x$$

Given two independent systems A and B, for which the joint probability density satisfies $p(A,B) = p(A)p(B)$, the Tsallis entropy of this system satisfies

$$H_q(A,B) = H_q(A) + H_q(B) + (1-q)H_q(A)H_q(B)$$

In the limit $q \to 1$

$$H(A,B) = H(A) + H(B)$$

Tsallis entropy (cont'd)

Non-additivity

Given two independent systems A and B, for which the joint probability density satisfies $P(A, B) = P(A)P(B)$. Recall

$$H_q(x) = \frac{1}{1-q}\left(1 - \sum_{i=1}^{n} p_i^q\right)$$

$$H_q(A, B) = H_q(A) + H_q(B) + (1 - q)\, H_q(A)\, H_q(B)$$

The parameter |1 - q| is therefore a measure of the departure from additivity. In the limit when q = 1,

$$H_q(A, B) = H_q(A) + H_q(B)$$

which is what is expected for an additive system. This property is sometimes referred to as ***pseudo-additivity.***

Transfer entropy

Transfer entropy is a non-parametric method to measure the amount of directed transfer of information between two random processes. Transfer entropy from a process Y to another process X is the amount of uncertainty reduced in future values of X by knowing the past values of Y given past values of X.
Recall mutual information,

$$I[X:Y] = \sum_{\substack{x\in\chi \\ y\in\Upsilon}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

One can also give it a *time lag* in either one of the variables x and y,

$$I[X(t):Y(t-1)] = \sum_{\substack{x\in\chi \\ y\in\Upsilon}} p(x(t),y(t-1))\log\frac{p(x(t),y(t-1))}{p(x(t))p(y(t))}$$

To give it a directional sense in an ad hoc way.

Schreiber, Thomas (1 July 2000). "Measuring information transfer". Physical Review Letters. 85 (2): 461–464.

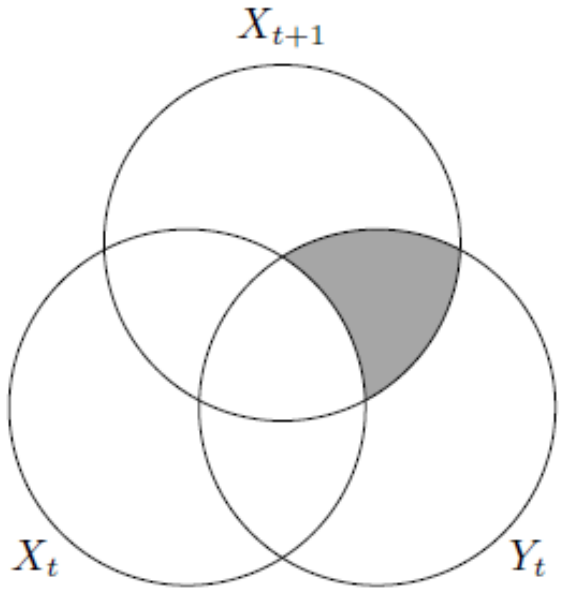
Transfer entropy (cont'd)

Transfer entropy is defined as

$$H_{y\to x}(x,y)=\sum_{\substack{x\in\chi\\ y\in\Upsilon}} p\big(x(t+1),x^{\alpha}(t),y^{\beta}(t)\big)\log\frac{p(x(t+1)|x^{\alpha}(t),y^{\beta}(t))}{p(x(t+1)|x^{\alpha}(t))}$$

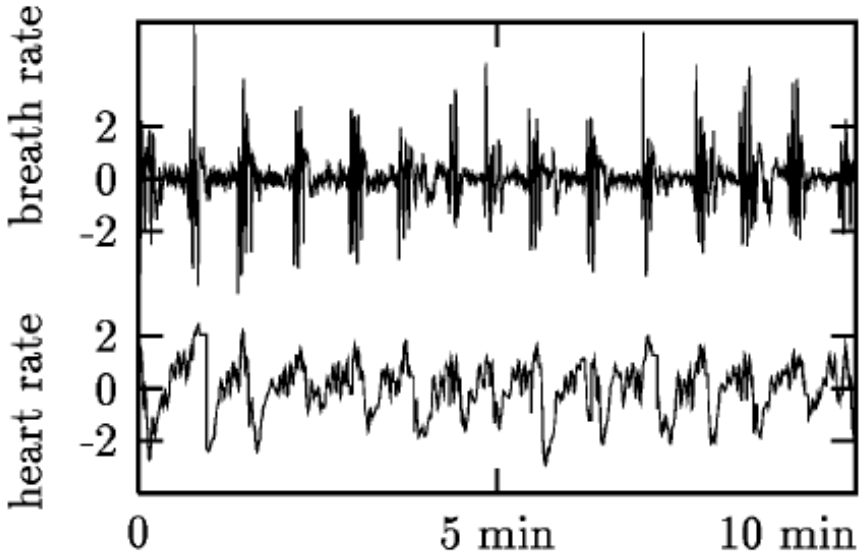
where α and β refer to Markov processes of orders α and β. In many cases, one chooses α and β to be equal to one. Transfer entropy reduces to *Granger causality* for vector auto-regressive processes. Hence, it is advantageous when the model assumption of Granger causality does not hold, for example, analysis of non-linear signals.

Venn diagram of transfer entropy, which is represented by the gray area

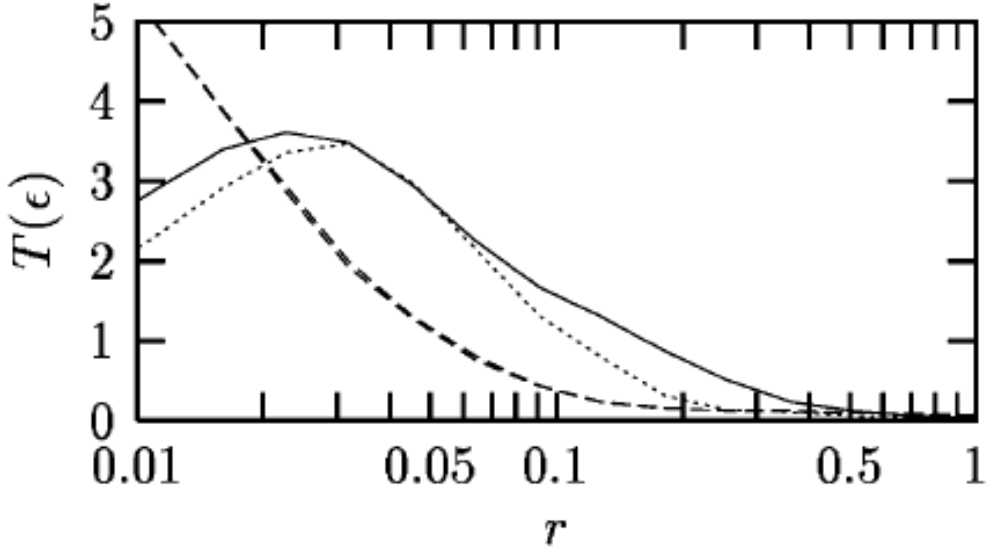


Transfer entropy (cont'd)

Example: Sleep Apnea



Bivariate time series of the breath rate (upper) and instantaneous heart rate (lower) of a sleeping human. The data is sampled at 2 Hz. Both traces have been normalized to zero mean and unit variance.



Transfer entropies *H*(heart → breath) (solid line), *H*(breath → heart) (dotted line), and time delayed mutual information *I* (0.5 second) (directions indistinguishable, dashed line) for the physiological time series shown in the left figure.

Schreiber, Thomas (1 July 2000). "Measuring information transfer". Physical Review Letters. 85 (2): 461–464.

Network entropy

There are different types of *network entropy* used nowadays. We here introduce one of them and demonstrate with some applications. We define the network entropy as the *entropy of the degree distribution*, which is given by

$$H = -\sum_{k=1}^{N-1} P(k) \log P(k)$$

where N is the total number of nodes in the network, $P(k)$ is degree distribution, which gives the probability of having a node with k links.

Network entropy (cont'd)

Using the formula above, one can calculate the entropy of several classes of networks, including the Erdös-Rényi (ER) random, Barabási-Albert (BA) scale-free and Watts-Strogatz (WS) small-world networks.

Recall that for the Erdös-Rényi (ER) random network, the degree distribution is

$$P(k) = \frac{(np)^k e^{-np}}{k!}$$

Similarly, for the Watts-Strogatz (WS) small-world networks, with $\langle k \rangle = K$

$$P(k) = \sum C_{K/2}^{n} (1-p)^n p^{\frac{K}{2}-n} \frac{\left(\frac{pK}{2}\right)^{k-\frac{K}{2}-n}}{\left(k-\frac{K}{2}-n\right)!} e^{-pK/2}$$

Barabási-Albert (BA) scale-free network with N nodes, $p(k) \sim k^{-\alpha}$, and

$$H = \left(\log(\alpha - 1) + \frac{\alpha}{1-\alpha}\right)\frac{1-N}{N} + \frac{\alpha}{1-\alpha}\frac{\log N}{N}$$

Network entropy: An example

We define the stock network entropy as the entropy of the degree distribution, which is given by

$$H = -\sum_{k=1}^{N-1} P(k) \log P(k),$$

where N is the total number of nodes in the network and $P(k)$ is the degree distribution, which gives the probability of having a node with k links.

Similarly, we define the relative entropy between two node degree distributions as

$$H^r = -\sum_{k=1}^{N-1} P(k) \log \frac{P(k)}{Q(k)},$$

where $Q(k)$ is the node degree distribution during the stable state period.

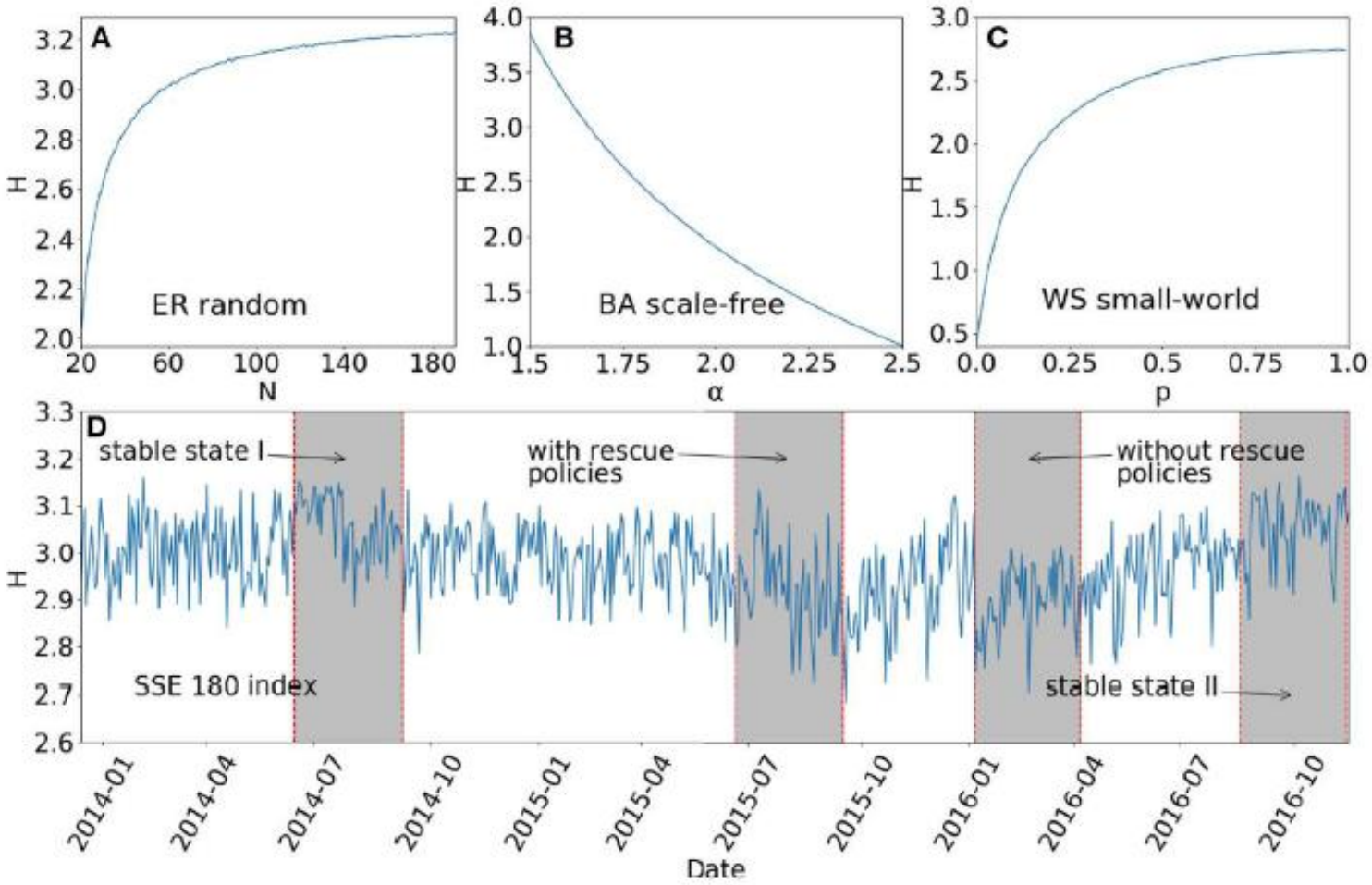
M.Y. Yang et al., "Stock network stability after crashes based on entropy method", Front. Phys. doi: 10.3389/fphy.2020.00163, Frontiers in Physics (2020).

TABLE 1 | Statistical description of the ER random, BA scale-free, and WS small-world networks and the stock network.

	ER random	BA scale-free	WS small-world	Stock network
N	180	180	180	180
E	537	531	540	534
C	0.03	0.10	0.17	0.73
$\langle l \rangle$	3	2.80	3.50	3.58
$\langle k \rangle$	6	6	6	5.93
$\langle s \rangle$	6	6	6	1.63
$\langle E_w \rangle$	1	1	1	0.28

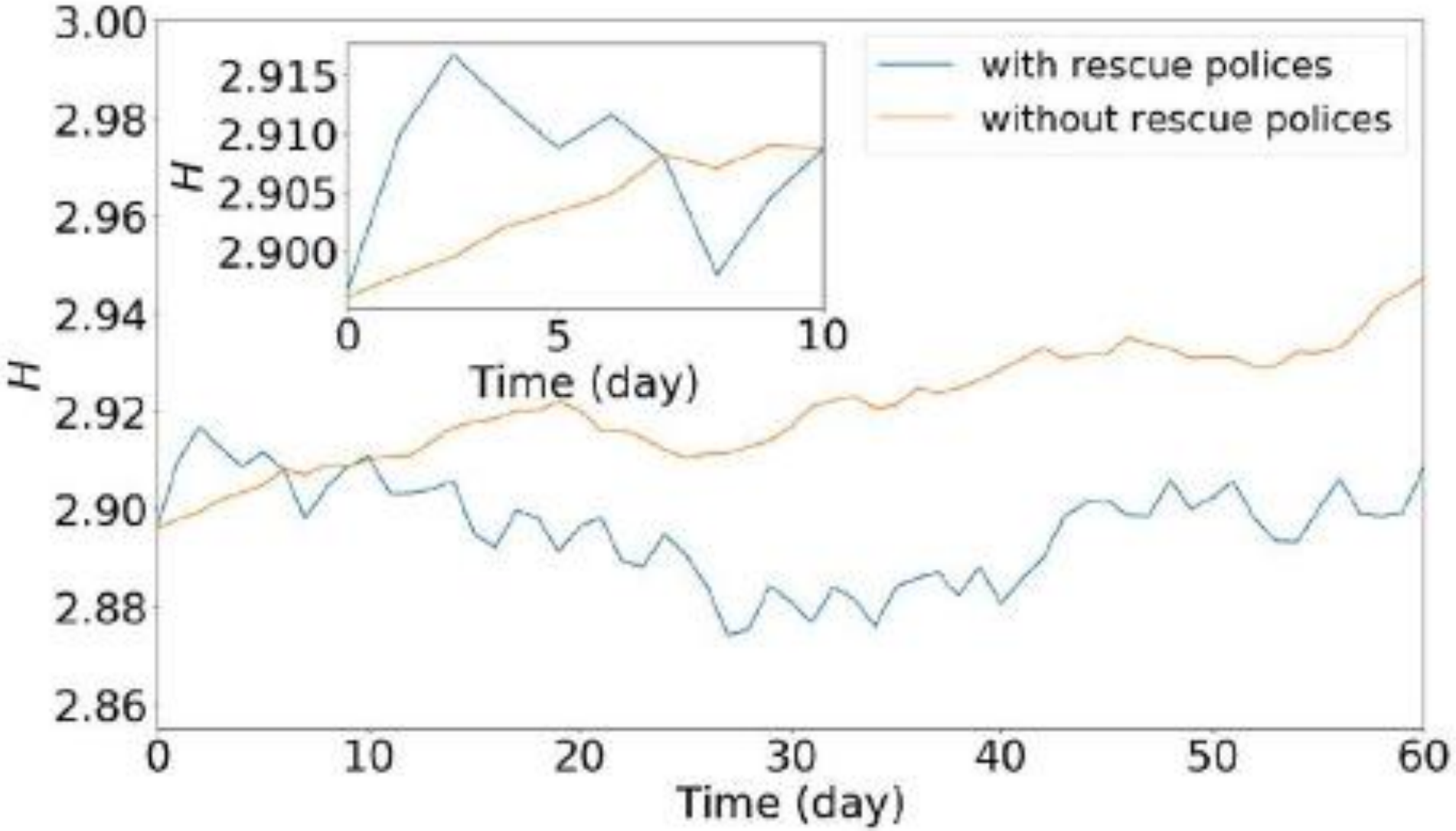
N represents the number of nodes, E the number of links, C the clustering coefficient, $\langle l \rangle$ the average shortest path length, $\langle k \rangle$ the average node degree, $\langle s \rangle$ the average node strength, and $\langle E_w \rangle$ the average link weight. The weights for the links in the ER random, BA scale-free and WS small-world networks are equal to 1. In the stock network, each link weight is defined as the correlation coefficient $\rho_{i,j}$ for stocks i and j [31].

Network entropy: An example



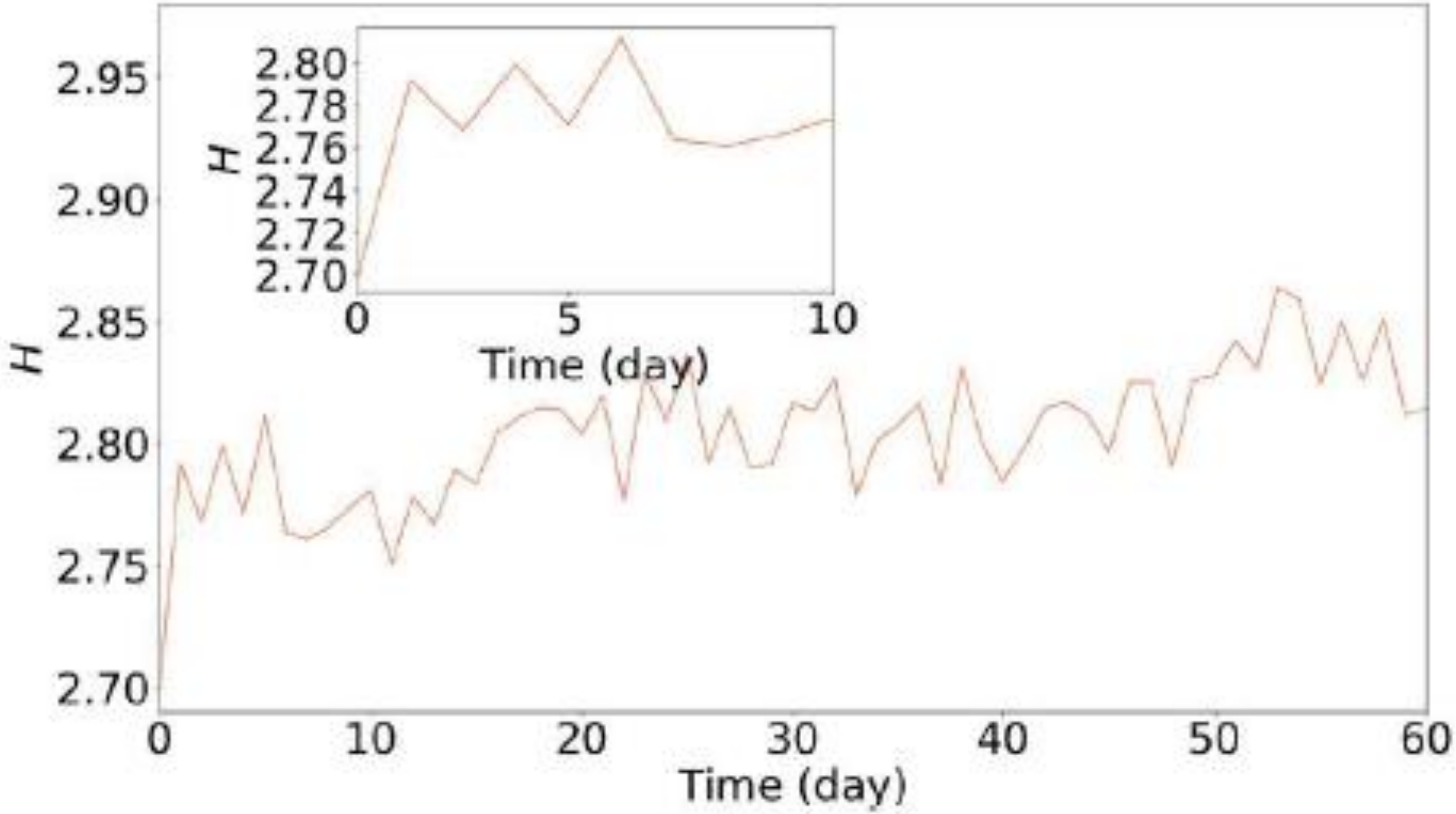
Entropy of (A) the ER random network with Np = 6 for various N; (B) the BA scale-free network for N = 180, m = 1 and various α; (C) the WS small-world network for K = 6, N = 180 and various p; and (D) the stock network for the sampling period from 16 December 2013 to 22 November 2016 in the Shanghai stock market.

Network entropy: An example



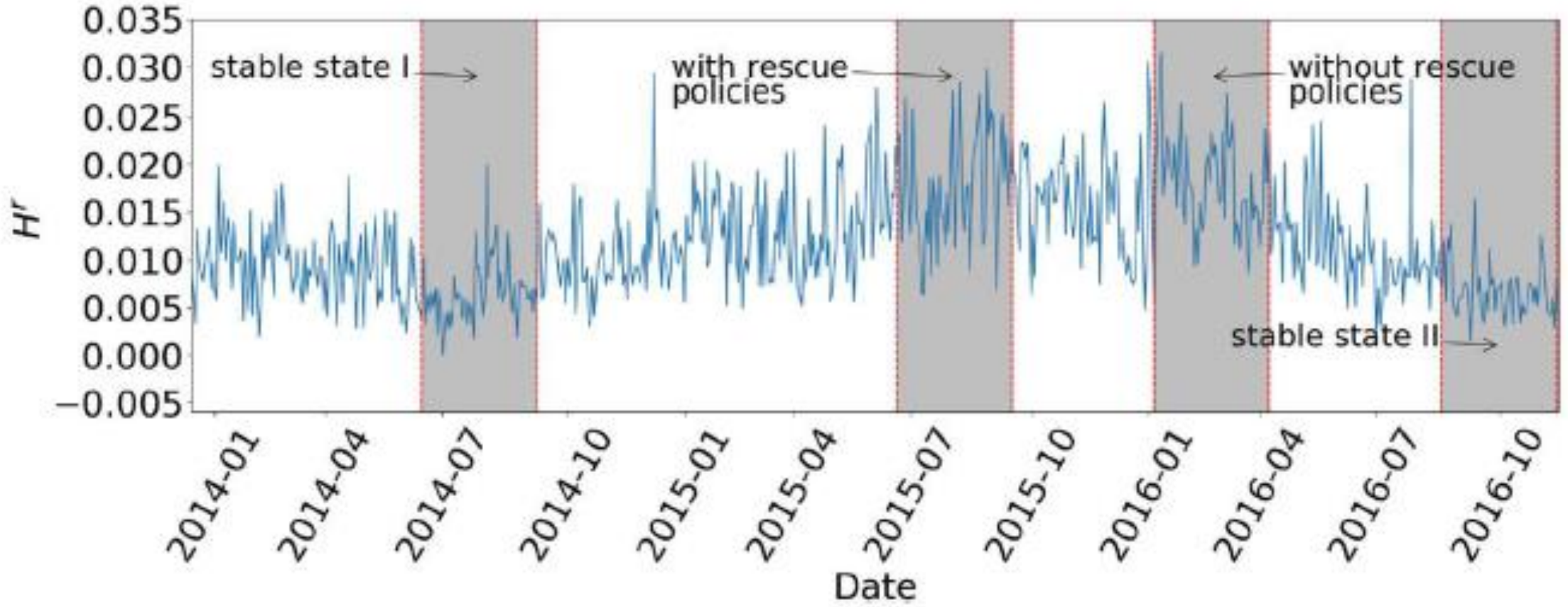
Entropy evolution of the stock network during sub-periods with and without rescue policies after crashes in the Shanghai stock market.

Network entropy: An example



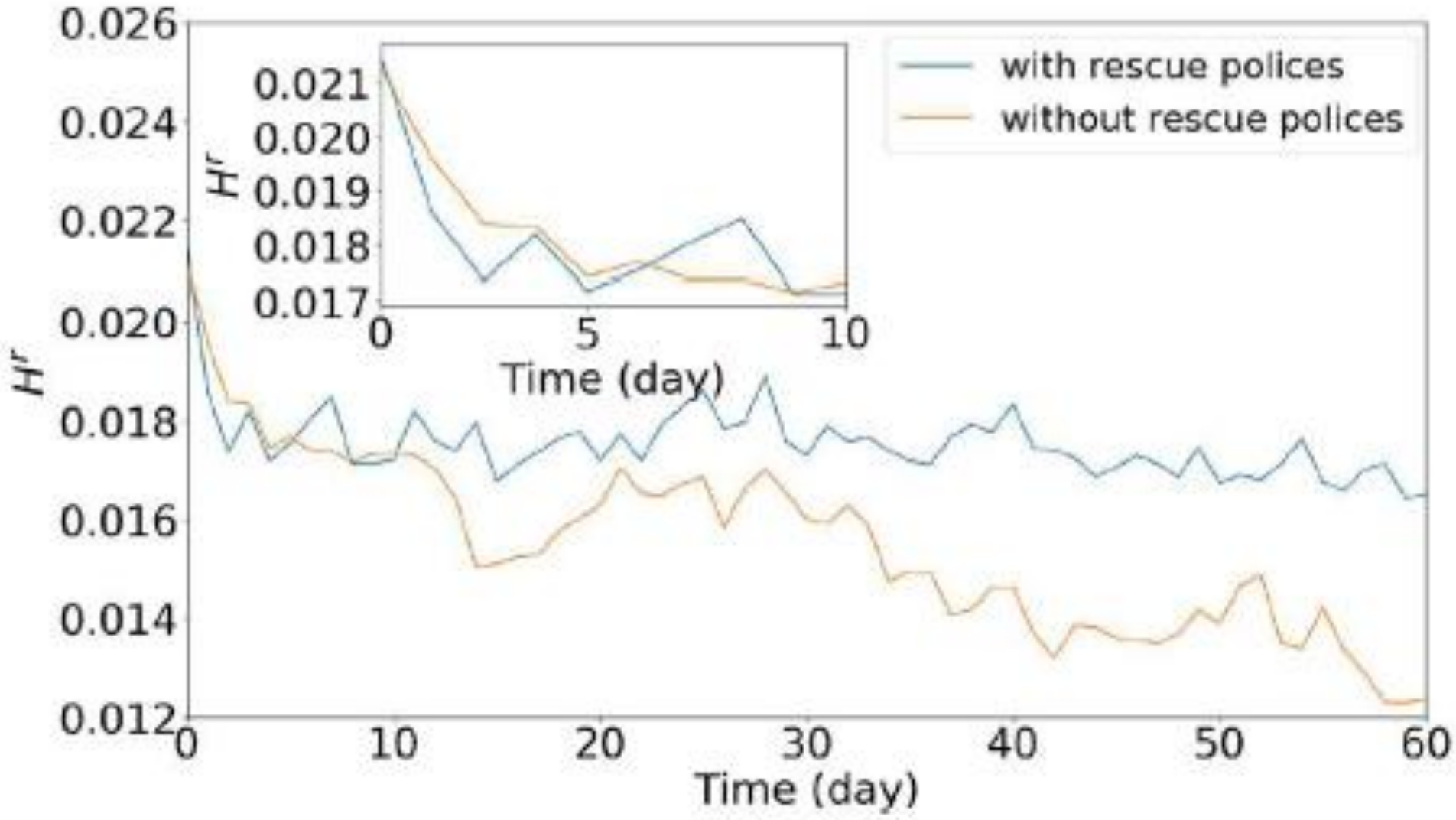
Entropy evolution of the stock network during the 60-day period following the crash on 10 March 2015 in the Hong Kong stock market.

Network entropy: An example



Relative entropy of the stock network for the sampling period from 16 December 2013 to 22 November 2016 in the Shanghai stock market.

Network entropy: An example



Relative entropy evolution of the stock network during sub-periods with and without rescue policies after crashes in the Shanghai stock market.

Von Neumann entropy

The extension of classical Gibbs entropy concepts to the field of quantum mechanics is called the *von Neumann entropy*, named after John von Neumann.

Recall: Quantum mechanics deals with probabilities, but the real quantum analog of a classical probability distribution is not a quantum state but a *density matrix*.

For a quantum mechanical system described by a density matrix ρ, the von Neumann entropy is given by

$$S = -tr(\rho \ln \rho)$$

where *tr* denotes the trace and ln denotes the (natural) matrix logarithm. If ρ is written in terms of the eigenvectors |1>, |2>, …, |n> as $\rho = \sum_j p_j |j><j|$, then

$$S = -\sum_j p_j \ln p_j$$

** Notice that the density matrix is ***an operator***.

Examples of Density Matrix

Consider a matrix state of |0> and |1> with probability ½ each. Thus,

$$|0\rangle\langle 0| = \begin{pmatrix}1\\0\end{pmatrix}(1 \quad 0) = \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}$$

$$|1\rangle\langle 1| = \begin{pmatrix}0\\1\end{pmatrix}(0 \quad 1) = \begin{pmatrix}0 & 0\\0 & 1\end{pmatrix}$$

Thus,

$$\rho = \frac{1}{2}(|0\rangle\langle 0| + |1\rangle\langle 1|) = \begin{pmatrix}\frac{1}{2} & 0\\0 & \frac{1}{2}\end{pmatrix}$$

Examples of Density Matrix (cont'd)

Now consider a matrix state of |+> and |–> with probability ½ each, where

$|\pm\rangle = \frac{1}{\sqrt{2}}(|0\rangle \pm |1\rangle)$. Thus,

$$|+\rangle\langle+| = \frac{1}{2}\begin{pmatrix}1\\1\end{pmatrix}(1 \quad 1) = \begin{pmatrix}1 & 1\\1 & 1\end{pmatrix}$$

$$|-\rangle\langle-| = \frac{1}{2}\begin{pmatrix}1\\-1\end{pmatrix}(1 \quad -1) = \begin{pmatrix}1 & -1\\-1 & 1\end{pmatrix}$$

Thus,

$$\rho = \frac{1}{2}(|+\rangle\langle+| + |1\rangle\langle1|) = \begin{pmatrix}\frac{1}{2} & 0\\0 & \frac{1}{2}\end{pmatrix}$$

Therefore, we can see that both cases give the same density matrix.

== > Two mixed states can be distinguished if and only if the density matrix ρ are different.

Von Neumann entropy (cont'd)

An immediate consequence is that, just as for the Shannon entropy, the von Neumann entropy has the property, $S \geq 0$, with equality only for a pure state (one of the p's being 1 and the others 0). This also implies the same upper bound that we had classically for a system with k states

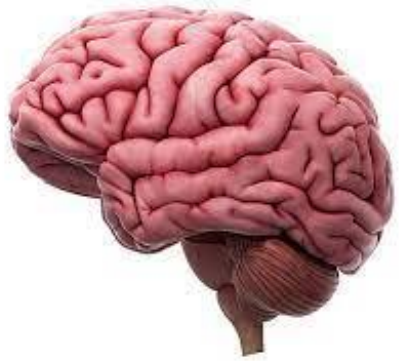
$$S \leq \log k \,,$$

with equality only if ρ is a multiple of the identity

$$\rho = \frac{1}{k}\begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & . \end{pmatrix}$$

In this case, the system is in a maximally mixed state.

From Physics to Biology:
What Is the Memory Capacity of the Human Brain?



- The human brain consists of about one billion neurons.
- Each neuron forms about 1,000 connections to other neurons, amounting to more than a trillion connections.
- Neurons combine so that each one helps with many memories at a time, exponentially increasing the brain's memory storage capacity to something closer to around 2.5 petabytes.
- If your brain worked like a digital video recorder in a television, 2.5 petabytes would be enough to hold three million hours of TV shows.

https://www.scientificamerican.com/article/what-is-the-memory-capacity/

Overall Summary

Information Entropy: $H[X] = -\sum_{x\in\chi} p(x)\log p(x)$

Relative Entropy: $D(p||q) = \sum_{x\in\chi} p(x)\log\left(\frac{p(x)}{q(x)}\right)$

Mutual Information: $I[X:Y] = \sum_{\substack{x\in\chi \\ y\in\Upsilon}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$

Information Inequality: $D(p||q) \geq 0$

Asymptotic Equipartition Property: $-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H[X]$

Data Compression: $H_D[X] \leq L^* < H_D[X] + 1$

Channel Capacity: $C = \max_{p(x)} I[X:Y]$