Network Science: Principles and Applications[2]

CS 695 - Fall 2016

Amarda Shehu, Fei Li

Department of Computer Science
George Mason University

[2]Part of this material is based on Tom Carter's talk at SFI Complex Systems Summer School and Wikipedia.

Applications of *information theory*, *thermodynamics*, and *Lagrange multiplier* in network science

Background 1: *Information theory* and *thermodynamics*

Measuring complexity of a large system (networked system)

Problem (A class problem in the field of complexity)

How can we tell that the system we are looking at is actually a **complex** *system?*

Having chosen a system to study, we might ask "How complex is this system?"

In this more general context, we probably want at least to be able to *compare two systems*, and be able to say that system A is more complex than system B. Eventually, we probably would like to have some sort of numerical rating scale.

Various approaches to "define" a system's complexity

1. human observation and subjective rating
2. number of parts or distinct elements — how to count one as a distinct part?
3. dimension — how to measured?
4. number of parameters controlling the system
5. minimal description — which language used?
6. **information content — how to define/measure information?**
7. minimal generator/constructor — what machines/methods to use?
8. minimum energy/time to construct — how the evolution count?

Example

Counting the number of parts is likely to depend on the scale at which the phenomenon is viewed (counting atoms is different from counting molecules, cells, organs, etc.).

We should not expect to be able to come up with a single universal measure of complexity. The best we are likely to have is a measuring system useful by a particular observer, in a particular context, for a particular purpose.

The first focus will be on measures related to **how surprising or unexpected an observation or event is**. This approach has been described as *information theory*.

Example

- **A frequentist version of probability**:
  Assume we have a set of possible events, each of which we assume occurs some number of times. Thus, if there are N distinct possible events $(x_1, x_2, \ldots, x_N)$, no two of which can occur simultaneously, and the events occur with frequencies $(n_1, n_2, \ldots, n_N)$, we say that the probability of event x_i is given by
  $$\Pr(x_i) = \frac{n_i}{\sum_{j=1}^{N} n_j}$$
- **An observer relative version of probability**:
  Take a statement of probability to be an *assertion about the belief that a specific observer has of the occurrence of a specific event*.

$$
\begin{aligned}
\Pr(\bar{a}) &= 1 - \Pr(a) \\
\Pr(a \vee b) &= \Pr(a) + \Pr(b) - \Pr(a \wedge b) \\
\Pr(a, b) &:= \Pr(a|b) \cdot \Pr(b) \\
&= \Pr(b|a) \cdot \Pr(a)
\end{aligned}
$$

A mathematical methodology for drawing inferences about the world from uncertain knowledge

We could say that our observation of the coin showing heads gives us information about the world.
We will develop a formal mathematical definition of the information content of an event which occurs with a certain probability.

We would like to develop a usable measure of the *information* we get from observing the occurrence of an event having probability p.
We will want our information measure $I(p)$ to have several properties:

1. Information is a non-negative quantity:
$$I(p) \geq 0$$
2. If an event has probability 1, we get no information from the occurrence of the event:
$$I(1) = 0$$
3. If two independent events occur (whose joint probability is the product of their individual probabilities), then the information we get from observing the events is the sum of the two information:
$$I(p_1 \times p_2) = I(p_1) + I(p_2)$$
4. The information measure to be a *continuous* (and, in fact, *monotonic*) function of the probability (slight changes in probability should result in slight changes in information).

$$
\begin{aligned}
I(p^2) &= I(p \times p) = I(p) + I(p) = 2I(p) \\
I(p^n) &= n \times I(p) \\
I(p) &= I((p^{1/m})^m) = m \times I(p^{1/m}) \\
I(p^{1/m}) &= \frac{1}{m} I(p) \\
I(p^{n/m}) &= \frac{n}{m} I(p) \\
I(p^a) &= a \times I(p), \forall 0 < p \leq 1, a > 0, a \in R
\end{aligned}
$$

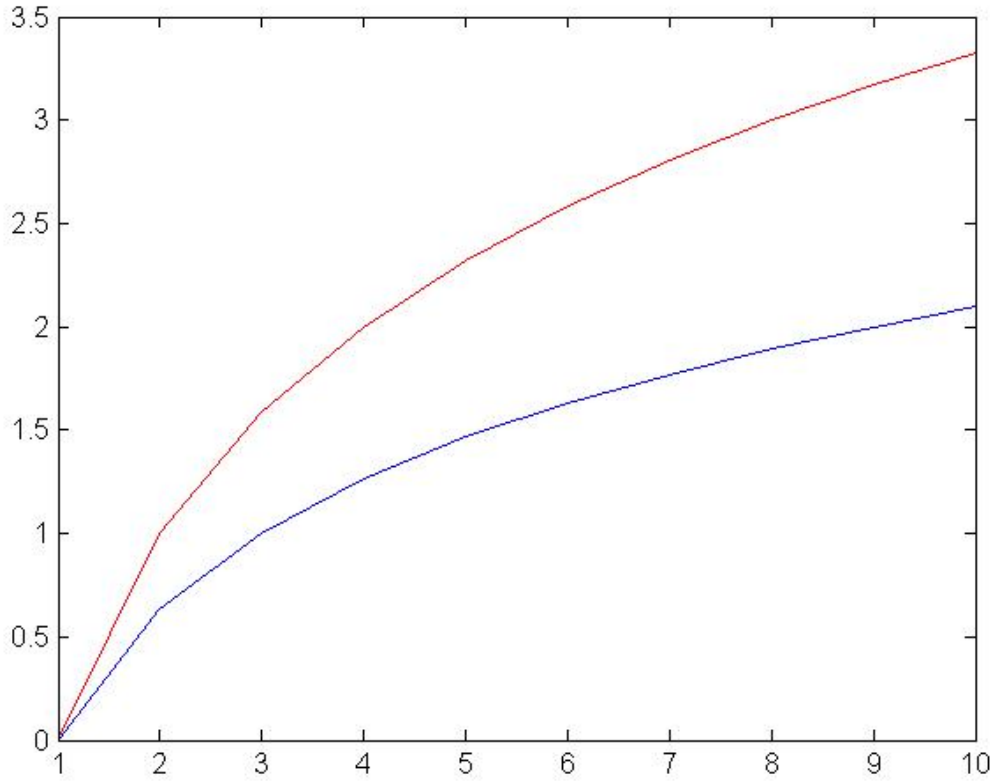
Theorem (Property of $I(p)$)

For some base b which determines the units we are using,

$$I(p) = -\log_b(p) = \log_b \frac{1}{p}$$

Using different bases for the logarithm results in information measures which are just constant multiples of each other, corresponding with measurements in different units.

Set $p = 1, 1/2, 1/3, \ldots, 1/10$. Set $b = 2, 3$



1. For an event with a smaller probability, we need more information (bits) to describe it. If such an event happens, it tells us more information.
2. If we use a larger unit to measure an event, we get less information.

Suppose that we have n symbols $\{a_1, a_2, \ldots, a_n\}$, and some source is providing us with a stream of these symbols. Suppose further that the source emits the symbols with probabilities $\{p_1, p_2, \ldots, p_n\}$, respectively. Assume that the symbols are emitted independently.

Problem

What is the average amount of information we get from each symbol we see in the stream?

$$
\begin{aligned}
I &= \sum_{i=1}^{n} (N \cdot p_i \cdot \log(1/p_i)) \\
\frac{I}{N} &= \frac{1}{N} \sum_{i=1}^{n} (N \cdot p_i \cdot \log(1/p_i)) \\
&= \sum_{i=1}^{n} (p_i \cdot \log(1/p_i))
\end{aligned}
$$

Definition (Entropy)

Suppose that we have a set of probabilities (a probability distribution) $P = \{p_1, p_2, \ldots, p_n\}$. We define the *entropy* of the distribution P by:

$$H(P) = \sum_{i=1}^{n} p_i \cdot \log(1/p_i)$$

$$H(P) = \int \Pr(x) \cdot \log(1/\Pr(x))dx$$

The entropy of a probability distribution is just the expected value of the information of the distribution.

Theorem (Gibbs inequality)

Given two probability distributions, $P = \{p_1, p_2, \ldots, p_n\}$ *and* $Q = \{q_1, q_2, \ldots, q_n\}$*, where* $p_i, q_i \geq 0$ *and* $\sum_i p_i = \sum_i q_i = 1$*. Then*

$$\sum_{i=1}^{n} p_i \ln\left(\frac{q_i}{p_i}\right) \leq 0$$

with equality only when $p_i = q_i$ *for all* i*.*

Proof.

$$\begin{aligned}
\ln x &\leq x - 1 \\
\ln 1 &= 1 - 1 = 0 \\
\sum_{i=1}^{n} p_i \ln\left(\frac{q_i}{p_i}\right) &\leq \sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1\right) \\
&= \sum_{i=1}^{n} (q_i - p_i) = \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i = 1 - 1 = 0
\end{aligned}$$

□

Corollary (A probability distribution that maximizes the entropy function)

$\log n$ *is the maximal entropy of a system* ($H(P) \leq \log n$) *and* $P = \{1/n, 1/n, \ldots, 1/n\}$.

Proof.

$$
\begin{aligned}
H(P) - \log n &= \sum_{i=1}^{n} p_i \log(1/p_i) - \log n \\
&= \sum_{i=1}^{n} p_i \log(1/p_i) - \log n \sum_{i=1}^{n} p_i \\
&= \sum_{i=1}^{n} p_i (\log(1/p_i) - \log n) \\
&= \sum_{i=1}^{n} p_i \log\left(\frac{1/n}{p_i}\right) \\
&\leq 0
\end{aligned}
$$

□

Example (Clarification)

- Suppose a book contains ascii characters. If the book is to provide us with information at the maximum rate, then each ascii character will occur with equal probability — it will be a random sequence of characters.
- The definitions of *information and entropy* depend only on the probability distribution.
  Two people listening to the same lecture can get very different information from the lecture. For example, without appropriate background, one person might not understand anything at all, and therefore have as probability model a completely random source, and therefore get much more information than the listener who understands quite a bit, and can therefore anticipate much of what goes on, and therefore assigns non-equal probabilities to successive words.

Definition (1st law of thermodynamics)

For change in energy dE, heat change dQ, work done dW,

$$dE = dQ - dW.$$

Definition (2nd law of thermodynamics)

- It is impossible for a system operating in a cycle and in contact with one thermal reservoir to perform positive work in the surroundings.
- It is impossible for a system operating in a cycle to produce positive heat flow from a colder body to a hotter body.

Definition (3rd law of thermodynamics)

As temperature goes to 0, the entropy S approaches a constant S_0. Furthermore, it guarantees that the entropy of a pure, perfectly crystalline substance is 0 if the absolute temperature is 0.

Explanation of the second law of thermodynamics

It is overwhelmingly probable that as time passes, macroscopically, **the system will increase in entropy until it reaches the maximum**. In many respects, these general arguments can be thought of as a "proof" (or at least an explanation) of a version of the second law of thermodynamics.

Given any macroscopic system which is free to change configurations, and given any configuration with entropy less than the maximum, there will be overwhelmingly many more accessible configurations with higher entropy than lower entropy, and thus, with probability indistinguishable from 1, the system will (in macroscopic time steps) successively change to configurations with higher entropy until it reaches the maximum.

Definition (Maximum entropy principle)

- Suppose we have a system for which we can measure certain macroscopic characteristics.
- Suppose further that the system is made up of many microscopic elements, and that the system is free to vary among various states.
- Let us assume that with probability essentially equal to 1, the system will be observed in states with *maximum entropy*.

We will then sometimes be able to gain understanding of the system by applying a maximum information entropy principle.

Example

Consider two rooms a, b and four people w, x, y, z. The entropy that these 4 people are in the same room is

$$H(P) = H(\{w, x, y, z\} \in a \vee \{w, x, y, z\} \in b) = \frac{1}{2^4} \log 2^4 + \frac{1}{2^4} \log 2^4 = 0.5$$

The entropy that two people in one room is

$$H(P) = \sum_{i=1}^{6} \frac{1}{2^4} \log 2^4 = \frac{3}{4} = 0.75$$

Background 2: Lagrange multiplier

We consider an optimization problem having the standard form

$$\begin{aligned} \text{min or max} \quad & f(x) \\ \text{subject to} \quad & h(x) = b \\ & x \in X \end{aligned}$$

Definition (Feasible set)

Define set $X(b) = \{x \in X : h(x) = b\}$ the **feasible set**.

Example

Minimize $x_1^2 + x_2^2$ subject to $a_1 \cdot x_1 + a_2 \cdot x_2 = b$ and $x_1, x_2 \geq 0$ for some given a_1, a_2 and b.

A beautiful and powerful method for solving constrained optimization problems is that of *Lagrange multipliers*. The idea is to **reduce constrained optimization to unconstrained optimization**, and to take the (functional) constraints into account by augmenting the objective function with a weighted sum of them.

Definition (Lagrangian)

$$L(x, \lambda) = f(x) - \lambda^T (h(x) - b)$$

where $\lambda \in R^m$ is a vector of Lagrange multipliers.

Theorem (Lagrangian sufficiency theorem)

Suppose $x \in X$ and $\lambda \in R^m$ such that $L(x, \lambda) = \inf_{x' \in X} L(x', \lambda)$ and $h(x) = b$. Then x is an optimal solution.

Proof.

$$\begin{aligned} \min_{x' \in X(b)} f(x') &= \min_{x' \in X(b)} [f(x') - \lambda^T (h(x') - b)] \\ &\geq \min_{x' \in X} [f(x') - \lambda^T (h(x') - b)] \\ &= f(x) - \lambda^T (h(x) - b) = f(x) \end{aligned}$$

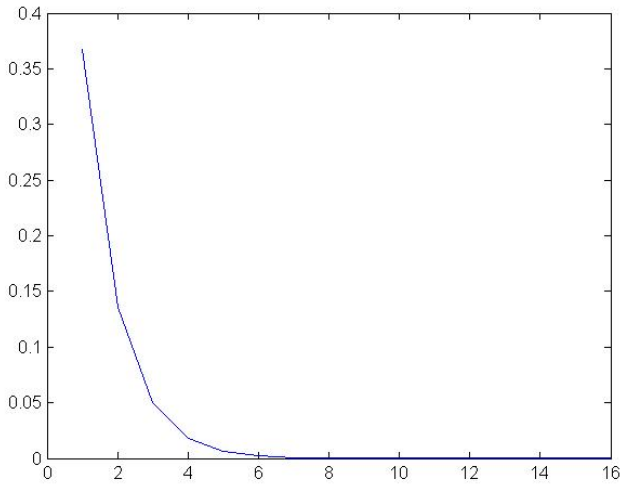
In the third line we finally use that x minimizes L (and $h(x) = b$). □

Example (Boltzmann economy 1: an agent's money distribution)

Suppose there is a fixed amount of money (M dollars), and a fixed number of agents (N) in the economy. Suppose that during each time step, each agent randomly selects another agent and transfers one dollar to the selected agent. An agent having no money does not go in debt. **What will the long term (stable) distribution of money be**?

Theorem (Boltzmann-Gibbs distribution)

Probabilities $\{p_i\}$ *denote an agent has the amount of money* i. $p_i = \frac{N}{M} e^{-\frac{i \cdot N}{M}}$.



- n_i: the number of agents who have i dollars, $i = 0, 1, \ldots, M$
- p_i: the probability that an agent has the amount of money i, $i = 0, 1, \ldots, M$

$$\begin{aligned} \sum_{i=0}^{M} (n_i \cdot i) &= M \\ \sum_{i=0}^{M} n_i &= N \\ p_i &= \frac{n_i}{N}. \end{aligned}$$

Proof.

Using an indicator variable $X_j = \begin{cases} 0, & j \text{ does not have money } i \\ 1, & j \text{ has money } i \end{cases}$, we have

$$n_i = \sum_{j=1}^{N} X_j = \sum_{j=1}^{N} p_i = p_i \cdot N.$$

So, we have

$$\begin{aligned}\sum_{i=1}^{M}(p_i \cdot i) &= \sum_{i=0}^{M}\frac{n_i}{N}\cdot i = \frac{\sum_{i=0}^{M}(n_i \cdot i)}{N} = \frac{M}{N}\\ \sum_{i=0}^{M} p_i &= 1\end{aligned}$$

Apply Lagrange multipliers:

$$L = \sum_{i=0}^{M}(p_i \cdot \ln(1/p_i)) - \lambda\left[\sum_{i=0}^{M}(p_i \cdot i) - \frac{M}{N}\right] - \mu\left[\sum_{i=0}^{M} p_i - 1\right],$$

from which we get

$$\begin{aligned}\frac{\partial L}{\partial p_i} &= -[1 + \ln p_i] - \lambda \cdot i - \mu = 0\\ \ln p_i &= -\lambda \cdot i - (1+\mu)\\ p_i &= e^{-(1+\mu)} \cdot e^{-\lambda \cdot i}\end{aligned}$$

$$1 = \sum_{i=0}^{M} p_i = \sum_i (e^{-(1+\mu)} \cdot e^{-\lambda \cdot i}) = e^{-(1+\mu)} \sum_{i=0}^{M} e^{-\lambda \cdot i}$$

$$\frac{M}{N} = \sum_i (p_i \cdot i) = \sum_i (e^{-(1+\mu)} \cdot e^{-\lambda \cdot i} \cdot i) = e^{-(1+\mu)} \sum_{i=0}^{M} (e^{-\lambda \cdot i} \cdot i)$$

We approximate (for large M)

$$\sum_{i=0}^{M} e^{-\lambda \cdot i} \approx \int_0^M e^{-\lambda \cdot x} dx \approx \frac{1}{\lambda}$$

$$\sum_{i=0}^{M} (e^{-\lambda \cdot i} \cdot i) \approx \int_0^M x \cdot e^{-\lambda \cdot x} dx \approx \frac{1}{\lambda^2}$$

We have

$$\begin{aligned} e^{-(1+\mu)} &= \frac{1}{\lambda} \\ e^{-(1+\mu)}\frac{M}{N} &= \frac{1}{\lambda^2} \end{aligned}$$

We now have

$$\begin{aligned} \lambda &= \frac{N}{M} = e^{-(1+\mu)} \\ p_i &= e^{-(1+\mu)}e^{-\lambda \cdot i} = \frac{N}{M}e^{-\frac{N \cdot i}{M}} \end{aligned}$$

References[29] and[30] are useful.

[29] "Statistical mechanics of money" by Adrian Dragulescu and Victor M. Yakovenko, http://arxiv.org/abs/cond-mat/0001432

[30] "Statistical mechanics of money: How saving propensity affects its distribution" by Anirban Chakraborti and Bikas K. Chakrabarti http://arxiv.org/abs/cond-mat/0004256

Suppose that a (simple) economy is made up of many agents "a", each with wealth at time t in the amount of $w(a,t)$. The total wealth in the economy is $W(t) = \sum_a w(a,t)$.

Problem (Boltzmann economy 2: all agents' money distribution — a power law?)

We are interested in looking at the distribution of wealth in the economy, so we will assume there is some collection $\{w_i\}$ of possible values for the wealth an agent can have, and associated probabilities $\{p_i\}$ that an agent has wealth w_i. We are hoping to develop a model for the collection $\{p_i\}$.

Look at global (aggregate/macro) observable of the system that reflect (or are made up of) characteristics of (micro) elements of the system.
Look at the growth rate of the economy. A reasonable way to think about this is to let

$$\begin{aligned} R_i &= \frac{w_i(t_1)}{w_i(t_0)} \\ R &= \frac{W_1}{W_0} \end{aligned}$$

(t_0 and t_1 represent time steps of the economy.)
The growth rate will then be $\ln(R)$. We then have the two constraints on the p_i:

$$\begin{aligned} \sum_{i=0}^{M}(p_i \cdot \ln(R_i)) &= \ln(R) \\ \sum_{i=0}^{M} p_i &= 1 \end{aligned}$$

We now apply Lagrange multipliers:

$$L = \sum_{i=0}^{M}(p_i \ln(1/p_i)) - \lambda \left[\sum_i (p_i \ln(R_i) - \ln(R))\right] - \mu \left[\sum_i p_i - 1\right],$$

from which we get

$$\frac{\partial L}{\partial p_i} = -[1 + \ln(p_i)] - \lambda \ln(R_i) - \mu = 0.$$

We can solve this for p_i:

$$p_i = e^{-(1+\mu)} e^{-\lambda \ln(R_i)} = e^{-(1+\mu)} R_i^{-\lambda}$$

Solving it, we get $1 + \mu = \ln(\sum_i R_i^{-\lambda})$. From this we see the power law (for $\lambda > 1$):

$$p_i = \frac{R_i^{-\lambda}}{\sum_i R_i^{-\lambda}}.$$

The reference is[34].

[34] "A statistical equilibrium model of wealth distribution" by Mishael Milakovic, 2001