

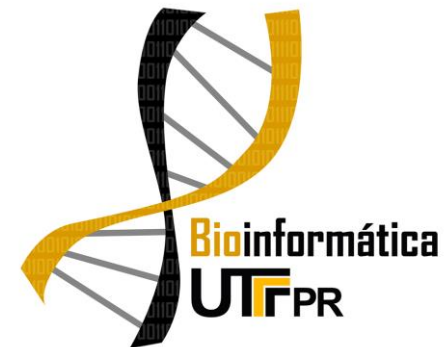
STATISTICAL MECHANICS FOR COMPLEXITY

A CELEBRATION OF THE 80TH BIRTHDAY OF CONSTANTINO TSALLIS

How to find the best q for highest information retrieval in discrete systems

Cassio H. S. Amador

Fabrício M. Lopes



Introduction

- I'm not an “*entropycist*”;
- Most of my research (PhD and post-docs) is on plasma physics and data analysis;
- I am from the north region of the Paraná State, in Brazil;
- Main economic activities are services, livestock and agriculture;
- Actually, the main cities started because of *coffee* plantations;
- It also has lots of universities and research going on.



Let's talk about coffee...

- There is intense research to improve our coffee production and processing;
- One of the technologies we use is the “freeze drying” (*liofilização*);
- The word “*liofilização*” came from the Greek (of course!);
- New technologies are developed to improve the coffee grain itself...

...some of them with the power of entropy!

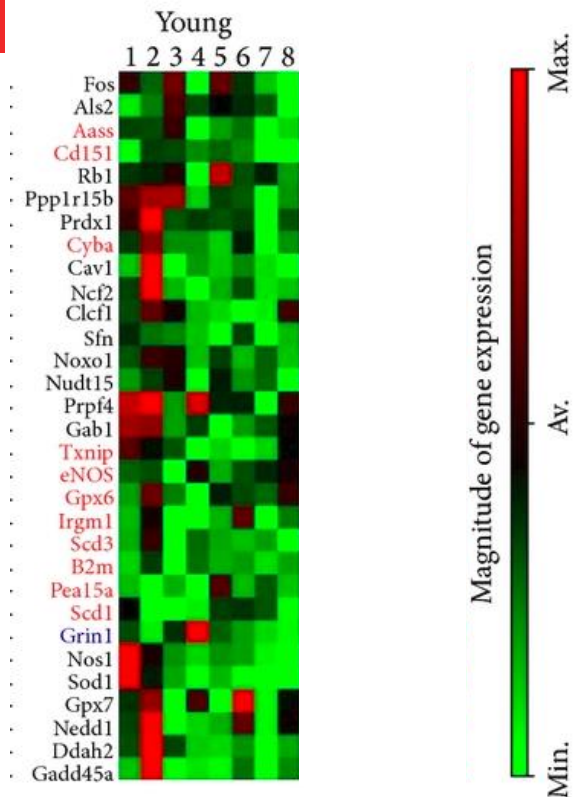
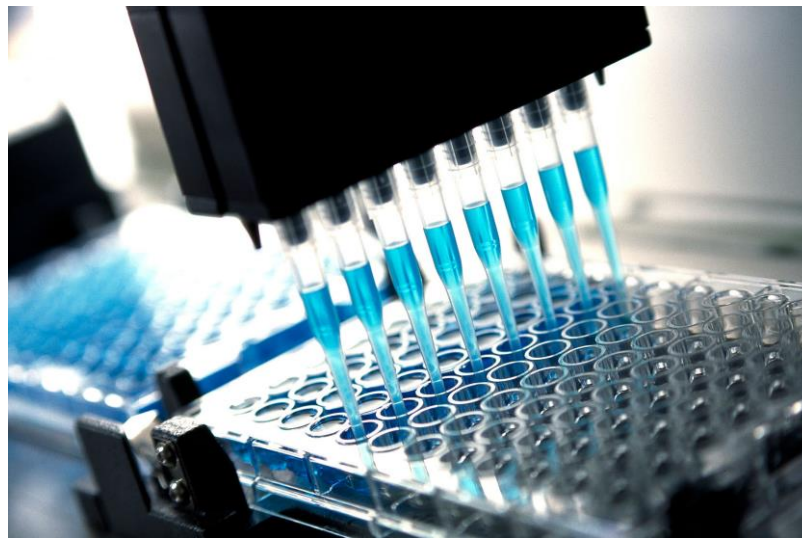


Bioinformatics

- Bioinformatics deals with data processing and analysis from biological systems;
- For example, there is active research to analyze coffee DNA and its genomic expressions;
- Near Londrina, at UTFPR-CP, there is a small but strong group of bioinformatics;
- I took my second Master's Degree with them.



Genomic Regulation



G1							
YG	YH	YI	YJ	YK	YL	YM	YN
G657	G658	G659	G660	G661	G662	G663	G664
10,051	10,242	11,137	11,022	8,0484	10,893	7,5977	8,2998
10,105	10,368	11,773	9,1031	7,8575	11,008	9,799	8,5782
9,3366	9,6217	10,638	10,222	8,4383	10,623	8,9566	8,684
9,4879	10,248	10,453	10,29	8,7789	10,561	9,1397	8,4264
9,5082	10,02	10,368	10,178	8,794	10,619	9,1006	8,2459
10,737	9,994	12,05	10,779	9,2065	11,141	8,6769	8,7946
10,42	9,8596	11,469	10,225	8,9883	11,262	8,5325	8,6191
7 7004	9 1313	9 5331	9 9185	8 0211	9 8152	7 441	8 8385

Time:

t0


t1

t2

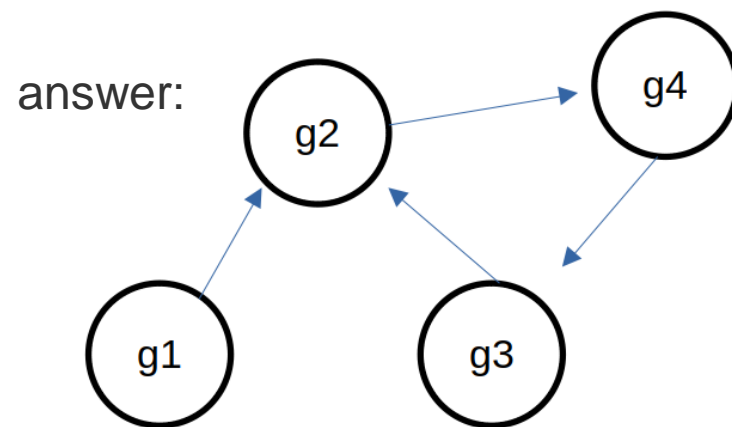
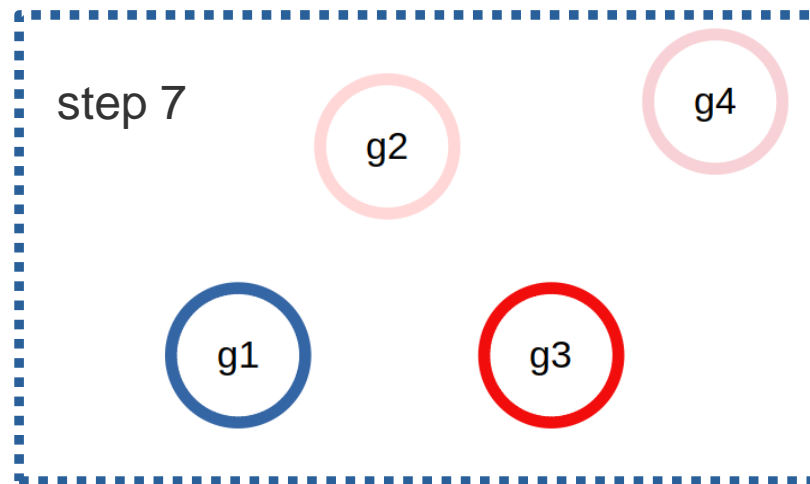
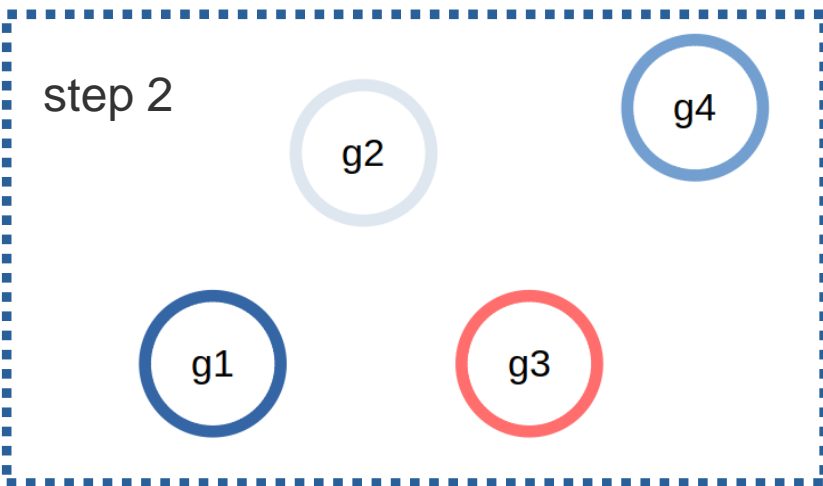
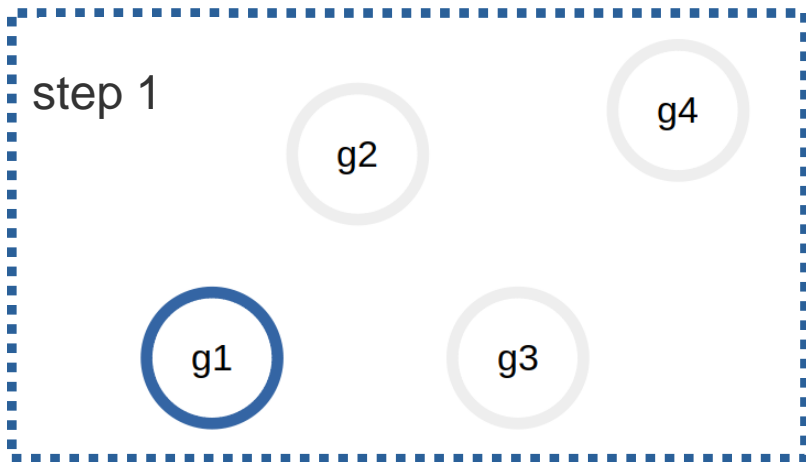
t3

t4

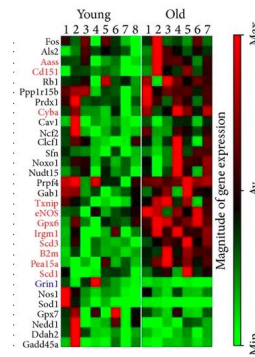
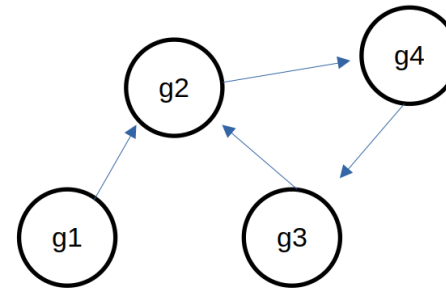
Genomic expression in brains of rats.

- 
- How to discover (*infer*) the genomic network in biological systems?

Gene Network inference



Gene network inference



- It is assumed that there exists an order in the system;
- To *infer* this order means to find out which structure exists underlying a measured genomic expression;
- It is common to have thousands of connected genes, and just a handful of measured time steps;
- Information is very scarce and every bit of it is valuable!
- There is definitely a long range interaction at play here.

Gene network inference

- One can build a *criterion function*, based on entropy:

$$C_q(Y|X) = \frac{\alpha(M - n)}{\alpha M + d} H_q(Y) + H_q(Y|X)$$

- $H_q(Y|X)$ is a function of the conditional q-entropy for predictor X to act on target Y :

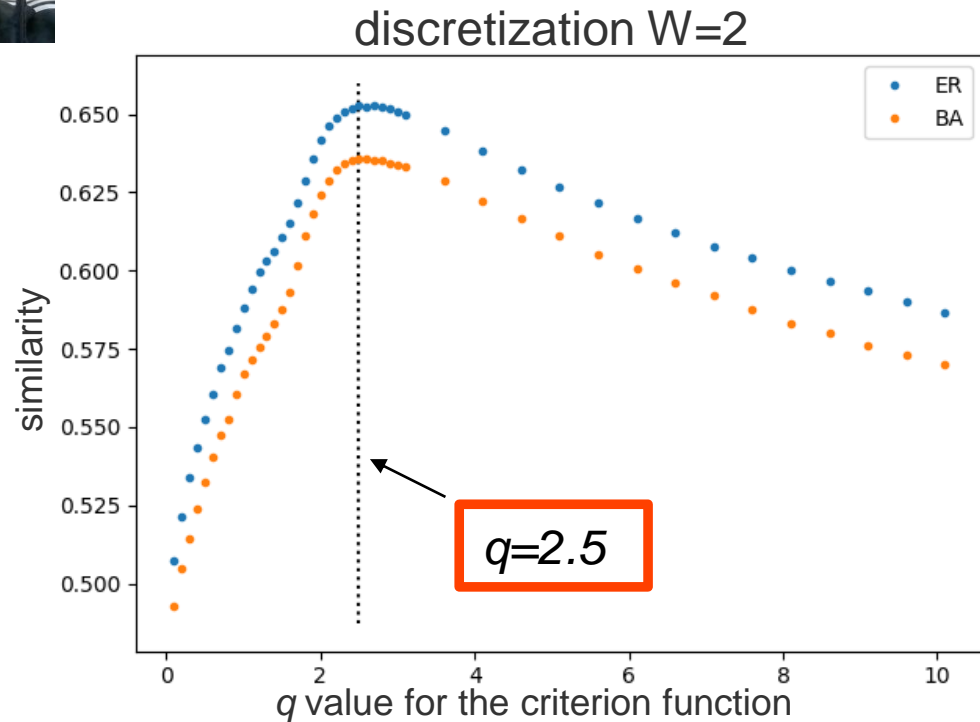
$$H_q(Y|X) = \sum_{x \in X} P(x) S_q(Y|x) \qquad S_q(Y|x) = \frac{1 - \sum_{y \in Y} P(y|x)^q}{q - 1}$$

- The structure is guessed for all possible combinations, and the one with the *lowest criterion function* value is chosen.
- The genomic expression values must be discrete, for example binary ($W=2$), ternary ($W=3$) or higher;

Previous work

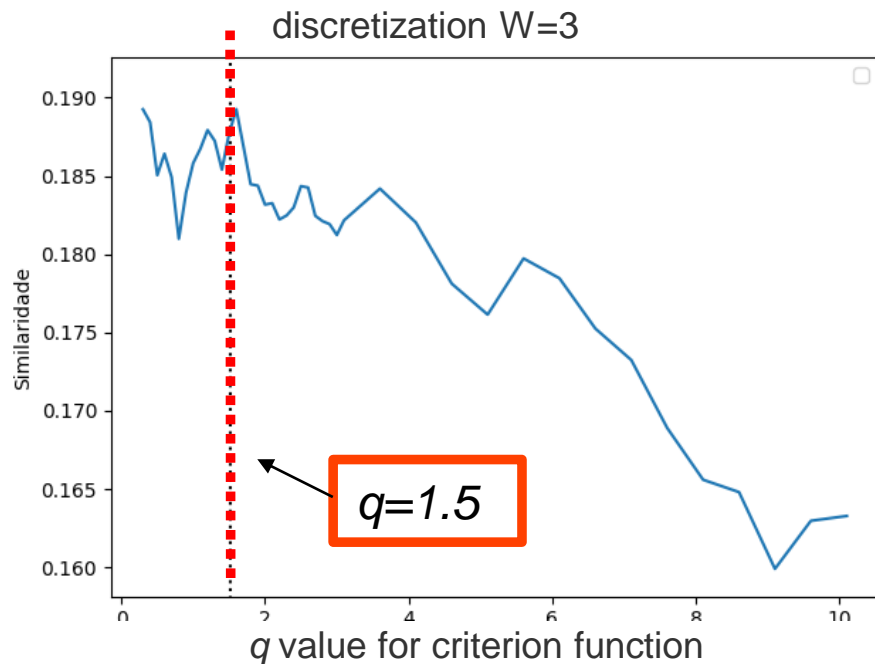



- Fabrício M. Lopes, during his PhD thesis in Computer Sci. at Unicamp, attended a seminar from a Physics researcher, called Tsallis, around 2007~2008;
- He is not a physicist, but he tried to apply this *Tsallis entropy* within the inference method for gene networks, and tested it with a large genomic expression data set.
- He found out that the results were much better compared with the BG entropy!
- His PhD won best thesis of the year at Unicamp in the informatics department.



Previous works


- Other data sets showed similar results, and he also tried with $W=3$:



- 
- ~10 years later, now he is supervising a physicist
 - My mission was to answer a simple question:
 - *WHY?*
 - *Why $q=2.5$ for binary discretization ($W=2$)?*
 - *Is it possible to arrive at this value from the gene network dynamics?*

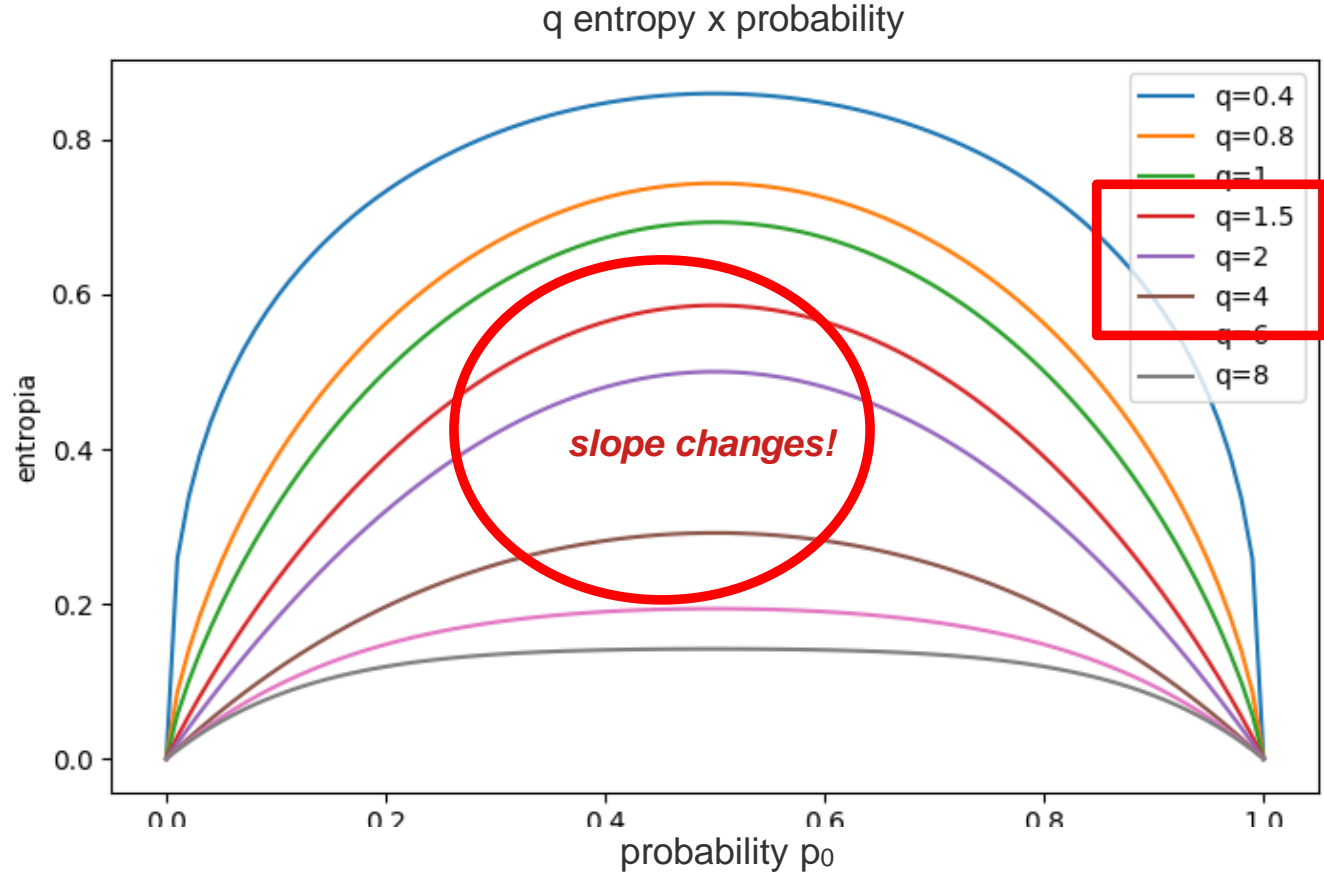
Initial possibilities and some dead ends

- Some hypotheses were raised:
 - Effect of “additivity” on the sum of network sections
$$S_q(A + B) = S_q(A) + S_q(B) + (1 - q)S_q(A)S_q(B)$$
 - Complexity of activation/inactivation between genes
 - Network topology (small world? Barabási-Albert? random?)
 - Influence of number of predictor genes and “hub” genes?

- 
- *More than 1 year of testing had passed...*
 - *and the possible explanation had nothing to do with gene networks!*

What does it mean to have “higher information”?

$$S_q = \frac{1 - \sum_{i=1}^W p_i^q}{1 - q}$$



What does it mean to retrieve more information?

- Maximum entropy indicates total disorder, or lack of information from an specific part of a given system;
- It is desirable that any available information shows a measurable change on entropy compared to the disorder state;
- The curve with the steepest slope should give more entropy variation in exchange for a small quantity of information;
- In my proposal I defined a normalized curve, called *Relative Entropy*:

$$S_q^{rel} = \frac{S_q}{S_q^{max}}$$

Relative entropy (or normalized entropy)

- For Tsallis entropy, the relative entropy S^{rel} is:

$$S_q^{rel} = \frac{1 - \sum_{i=1}^W p_i^q}{1 - W^q}$$

(W is the number of possible discrete values)

- In the case $q \rightarrow 1$:

$$S_1^{rel} = \frac{\sum_{i=1}^W p_i \ln p_i}{\ln W}$$

Finding the best value for q

- With the normalized curves, we look for the one with the greatest concavity:

$$\frac{\partial S_q^{rel}}{\partial q} = 0$$

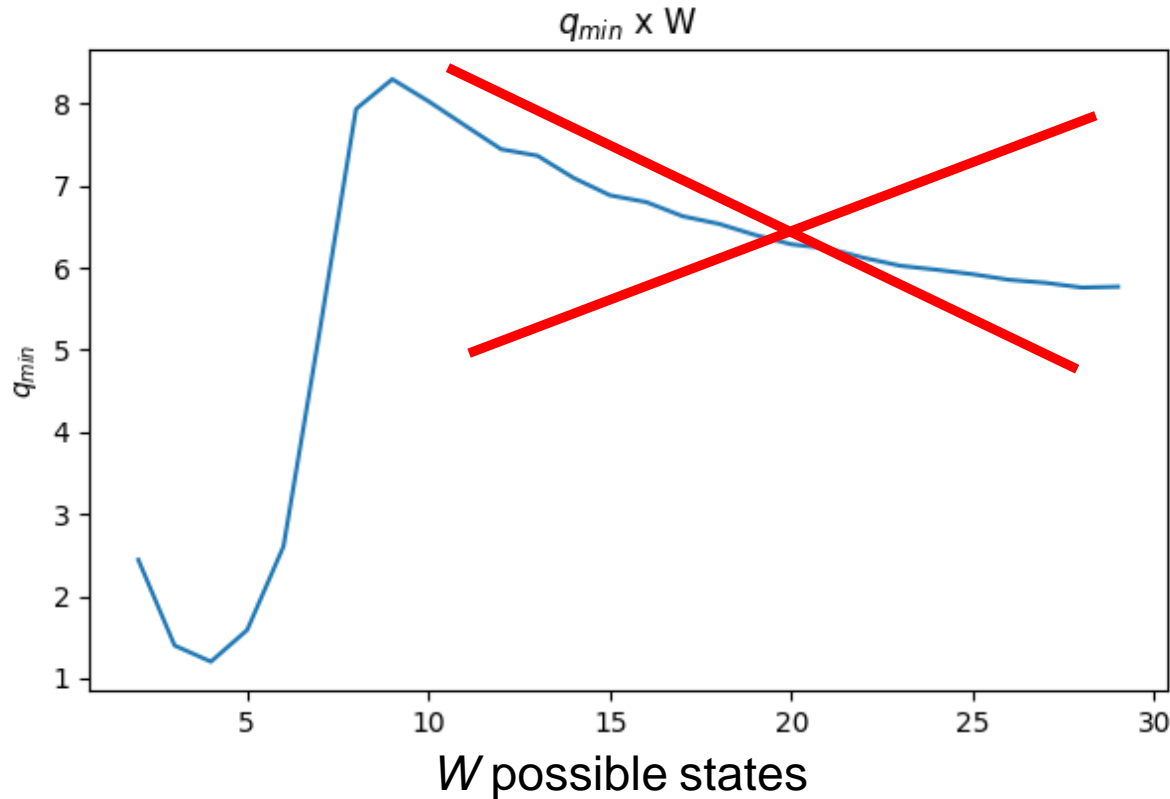
- Which results in:

$$0 = \left(\sum_{i=1}^W p_i^q \ln p_i \right) (W^{q-1} - 1) + \left(1 - \sum_{i=1}^W p_i^q \right) \ln W$$

- *In other words, for a given probability distribution, discretized into W states, it is possible to numerically find the value of a q_{\min} that maximizes the amount of information that can be inferred!*

Case $W > 2$

$$0 = \left(\sum_{i=1}^W p_i^q \ln p_i \right) (W^{q-1} - 1) + \left(1 - \sum_{i=1}^W p_i^q \right) \ln W$$



Average q	
W	q
3	1,52
4	1,22
8	8,16
15	6,91
20	6,30
29	5,73

Case $W=2$

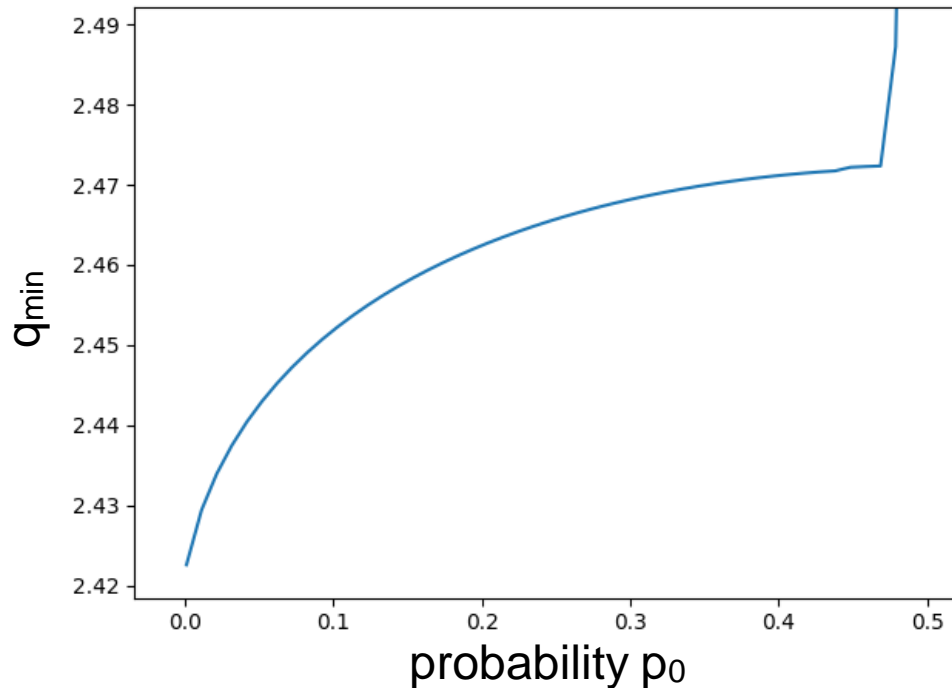
- There are only two probabilities, p_0 e p_1 :

$$0 = 2^{q-1} - 1[p_0^q \ln(p_0) + (1 - p_0)^q \ln(1 - p_0)] + \ln(2)[1 - p_0^q - (1 - p_0)^q]$$

- Two extreme cases:
 - If one of the probabilities is zero, the logarithm diverge to minus infinity, therefore only the case $q=1$ is numerically possible.
 - If $p_0 \approx p_1 \approx 0.5$, the above equation is null for any value of q , as expected, since it is the case of maximum entropy.

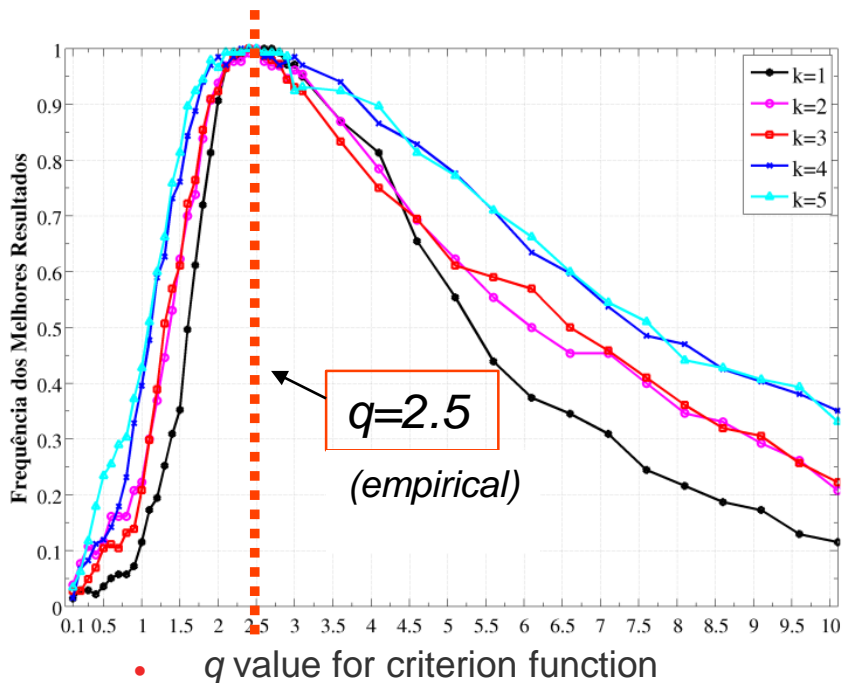
Case $W=2$

- Solving numerically for the probabilities $0 < p_0 < 0,5$:

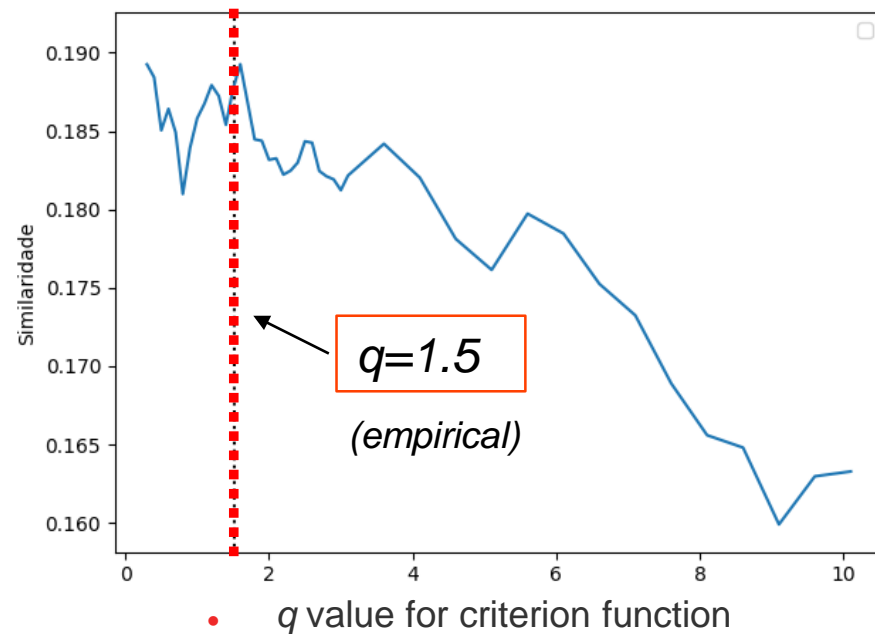


- Meaningful values are between
2.42 and 2.47
- Weighted average is $q \approx 2.46$

In summary for $W=2$ and $W=3$



W	q_{min}
2	2.46



W	q_{min}
3	1.52

Final remarks

$$S_q^{rel} = \frac{S_q}{S_q^{max}} \quad \frac{\partial S_q^{rel}}{\partial q} = 0$$

- These results are still unpublished;
- The results for the $W=2$ case are very good;
- The proposed method must be improved for $W>3$;
- Research is ongoing to apply to other information systems, for example, in digital transmission signal sources.

Thank you for your attention!

cassioamador@utfpr.edu.br

and Happy Birthday Tsallis!

