Why do complex systems look critical?

Matteo Marsili
The Abdus Salam International Centre for Theoretical Physics
Trieste, Italy

+ Iacopo Mastromatteo
Yasser Roudi
Ariel Haimovici
Dante Chialvo
Silvio Franz
Claudia Battistin

Journal of Statistical Mechanics: Theory and Experiment > Volume 2013 > September 2013

Matteo Marsili *et al* *J. Stat. Mech.* (2013) P09003 doi:10.1088/1742-5468/2013/09/P09003

On sampling and modeling complex systems

Matteo Marsili[1], Iacopo Mastromatteo[2] and Yasser Roudi[3,4]

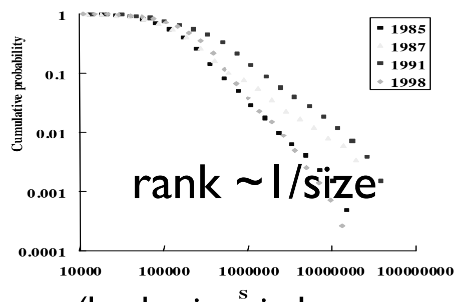
The unreasonable effectiveness of science

The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve. We should be grateful for it and <u>hope it will remain valid also in future research</u> and that it will extend, for the better of for the worse, to our pleasure, even though perhaps also to our bafflement, <u>to wide branches of learning</u> (E. P. Wigner 1960)

- Galaxies have millions of stars, a piece of material has 10^{32} molecules, ...
  Yet, we understand their behavior in terms of few <u>relevant</u> variables!
- Will this work for a cell (10^4 genes), the brain (10^7 neurons)
  an economy (10^6 individuals)... ?
- We build airplanes. Can we also cure cancer or avoid the next financial crisis?
- Even if the answer is no, what is the best we can do?
- How to find the (most) relevant variables or description of complex phenomena?

Facts and questions

- Fact 1:
  Data deluge + advanced experimental techniques (e.g. sequencing)
  Complex systems involve many variables (high-d inference, e.g. 10^4 genes)
  Strong under-sampling. Prediction is typically hard (e.g. drug design)
- Fact 2:
  We observe "Criticality", as a statistical regularity,
  in a wide variety of different systems as cities,
  the brain, languages, economy/finance, biology.



(land prices in Japan
Kaizoji & Kaizoji 2006)

- Questions:
  Are there typical properties of high-d samples of complex systems?
  Are there overarching organizing principles (e.g. SOC)?
  Can we exploit "criticality" (e.g. for model selection)?

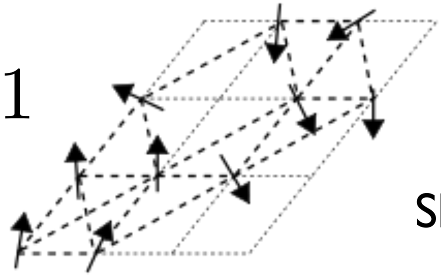
P. Bak How Nature Works (1996)
T. Mora & W. Bialek, J.Stat.Phys. (2011)
S. Ki Baek et al. N. J. Physics (2012)

Criticality in (statistical) physics

- Statistical mechanics: order and disorder

$$\underline{s} = (s_1, \ldots, s_N), \qquad s_i = \pm 1$$

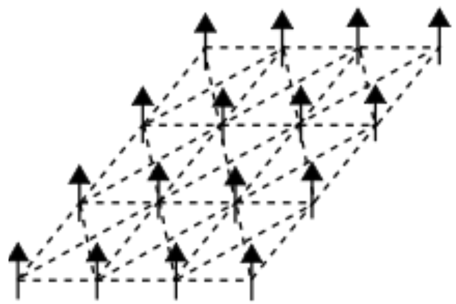
$$p\{\underline{s}|\hat{g}\} = \frac{1}{Z} e^{-E_{\hat{g}}[\underline{s}]/T}$$



$T \gg T_c$

Weak interaction
Short range correlations
Large entropy

critical point T_c



$T \ll T_c$

Strong interaction
Long range order
Small entropy

- **Critical phenomena:**
  - anomalous fluctuations (C_V)
  - scale invariance

$$C(r) \sim r^{-d-\eta}$$

Criticality everywhere

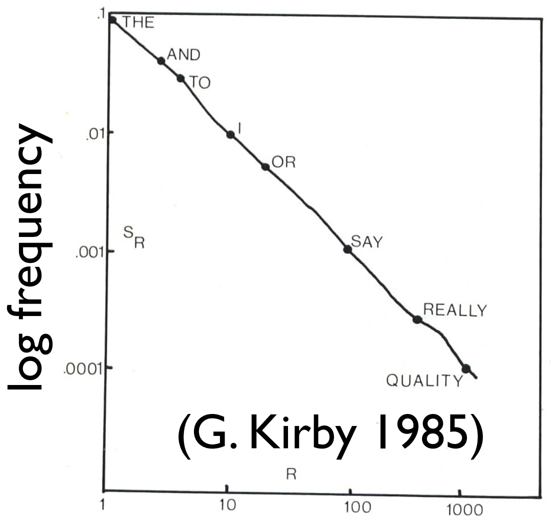
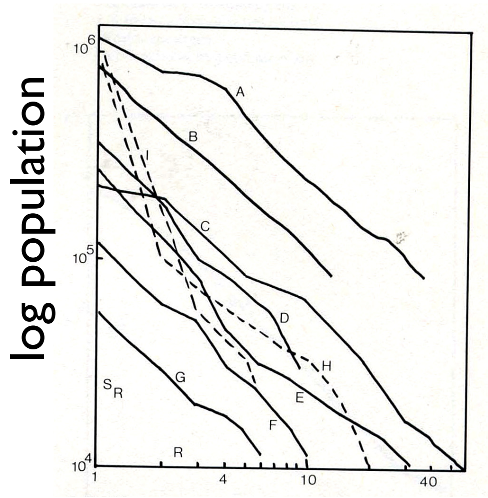


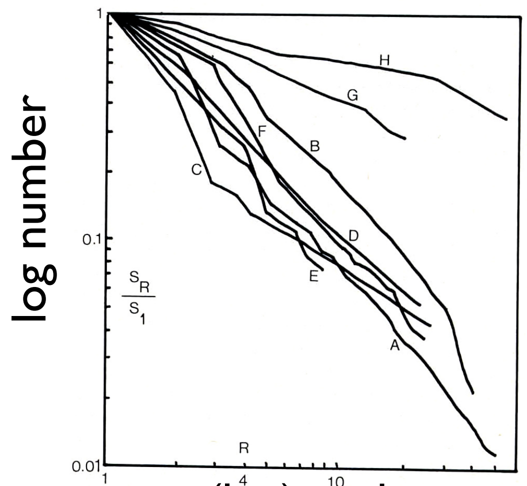
Figure 1 Frequency of word usage in English

(log) rank



(log) rank

A	United States	B	China
C	West Germany	D	Spain
E	France	F	East Germany
G	Switzerland	H	United Kingdom
I	Mexico		



(log) rank

A	Populations of all countries
B	Number of ships built by all countries
C	Students at English universities
D	Building Societies by assets
E	Populations of World's religions
F	US insurance companies by staff
G	World languages
H	English public schools by students

$$\text{rank} \propto \text{size}^{-1} \quad \Rightarrow \quad N(\text{size}) \sim \text{size}^{-2}$$

From empirical distribution to energy

$$P\{\underline{s}\} = \frac{1}{Z} e^{-\beta E\{\underline{s}\}} \qquad \Rightarrow \qquad E\{\underline{s}\} \simeq -\log \frac{K_{\underline{s}}}{M}$$

number of observations of state s

total number of observations

Criticality = linear relation between energy and entropy ~ kN(k)

Peak of Cv in learned models

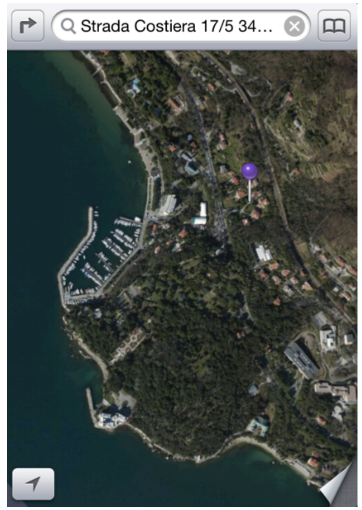
T. Mora & W. Bialek, J.Stat.Phys. (2011)

Complex system

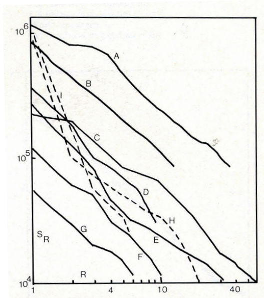
= many degrees of freedom + function

- Complex systems are not random:
  - Individuals do not live in random cities
  - A writer does not choose words at random when writing
  - Proteins are not random sequences of amino acids
  - ...
- Only part of what they do is accessible to us:
  - Variables: $\vec{s} = (\underbrace{s_1, \ldots, s_n}_{\underline{s}\ \text{knowns}}, \underbrace{s_{n+1}, \ldots, s_N}_{\bar{s}\ \text{unknowns}})\,, \qquad s_i = \pm 1,\ N \gg 1$
  - Function: $$U(\vec{s}) = \underbrace{u_{\underline{s}}}_{\text{model}} + \underbrace{v_{\bar{s}|\underline{s}}}_{\text{unknown function}}, \qquad \langle v_{\bar{s}|\underline{s}} \rangle = 0$$
  - Behavior: $$\underline{s}^* = \arg\max_{\underline{s}} \left[u_{\underline{s}} + \max_{\bar{s}} v_{\bar{s}|\underline{s}} \right]$$

How relevant are known vars?
e.g. Why do you live where you live?



- I live where I live because my zip code can be nicely decomposed in primes: 34151 = 13 x 37 x 71
- Others choose where to live depending on job, marriage, interests, etc. The zip code is not a relevant variable in this choice, whereas the city is.
- The distribution of city sizes contains information about how people choose where to live. The distribution by zip code does not.



- The distribution of population by zip code is trivial, that by city is not
- Same for language: word are the relevant variables, punctuations marks are not ...
- <u>Modeling</u>: models should contain relevant variables to be predictive
- <u>Sampling</u>: if the variables we sample are relevant, we can infer what the system is doing

Modeling:

(the direct problem)

Observables (knowns)

$$\max_{\underline{s}} \max_{\bar{s}} U(\underline{s}, \bar{s}) \Rightarrow \underline{s}^*$$

Nature

$$\max_{(\underline{s},\bar{s})} U(\underline{s}, \bar{s})$$

$$p_{\underline{s}^*} = P\{\underline{s}_0 = \underline{s}^*\}$$

Model

$$\max_{\underline{s}} E_{\bar{s}}\left[U(\underline{s}, \bar{s})\right]$$

$$= \max_{\underline{s}} u_{\underline{s}} \Rightarrow \underline{s}_0$$

$$\underline{s} = (s_1, \ldots, s_n), \qquad n = fN$$

$$\bar{s} = (s_{n+1}, \ldots, s_N)$$

Q: How many? How relevant?

$$\Rightarrow \quad P\{\underline{s}^* = \underline{s}\} = \frac{1}{Z(\beta)} e^{\beta u_{\underline{s}}}, \qquad Z(\beta) = \sum_{\underline{s}} e^{\beta u_{\underline{s}}}$$

Gibbs-Boltzmann distribution

- Without further knowledge, $v_{\bar{s}|\underline{s}}$ has to be taken as an i.i.d. random variable
- As long as $\langle |v_{\bar{s}|\underline{s}}|^m \rangle < \infty \qquad \forall m$

$$\Rightarrow \max_{\bar{s}} v_{\bar{s}|\underline{s}} = a + \beta^{-1} Y, \qquad Y \sim \mathrm{Gumbel}$$

- Then

$$P\{\underline{s}^* = \underline{s}\} = \frac{1}{Z(\beta)} e^{\beta u_{\underline{s}}}, \qquad Z(\beta) = \sum_{\underline{s}} e^{\beta u_{\underline{s}}}$$

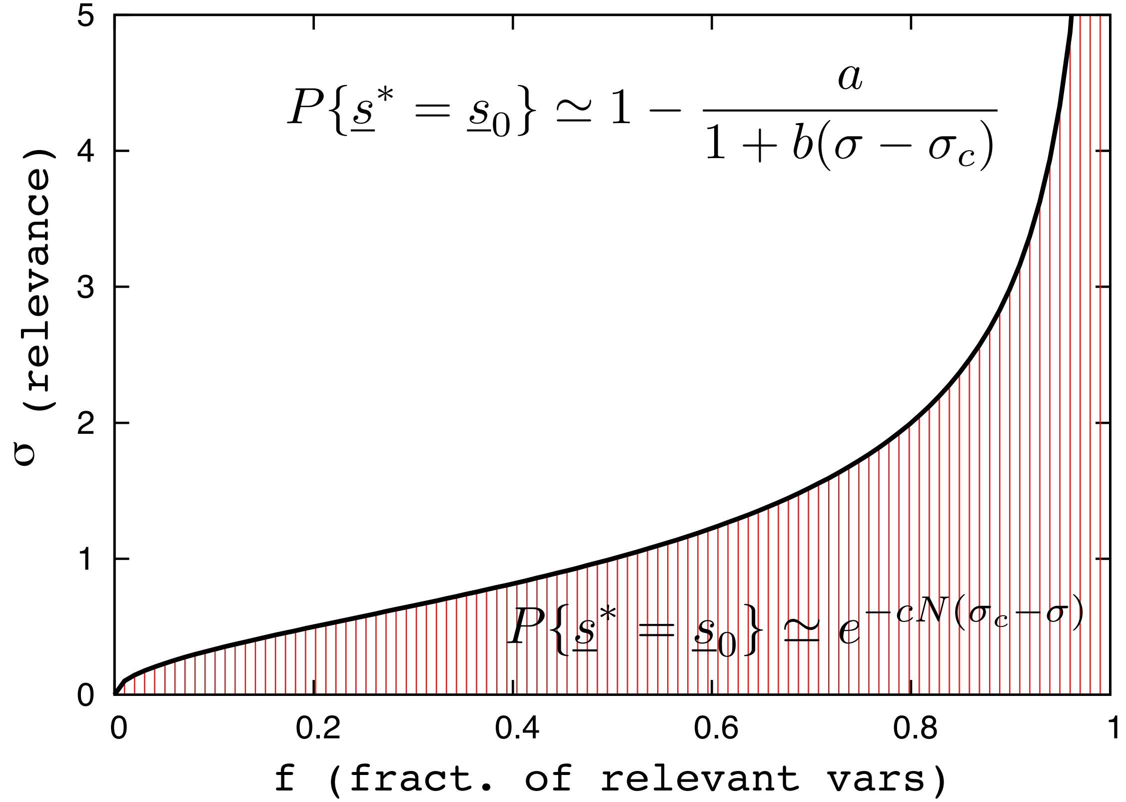
- For Gaussian(0,1) P{v}, $\beta = \sqrt{2N(1-f)\log 2}$
- Same as maximal entropy with $\langle u_{\underline{s}} \rangle = \bar{u}$

The most complex system: REM

- If $u_{\underline{s}} \sim \mathrm{Gaussian}(0, \sigma^2)$ i.i.d. then $\sigma_c = \sqrt{\frac{f}{1-f}}$

$\underline{s} = (s_1, \dots, s_n), \qquad n = fN$

$\bar{s} = (s_{n+1}, \dots, s_N)$



Known variables should be relevant enough!
(relevant = those the system cares about)

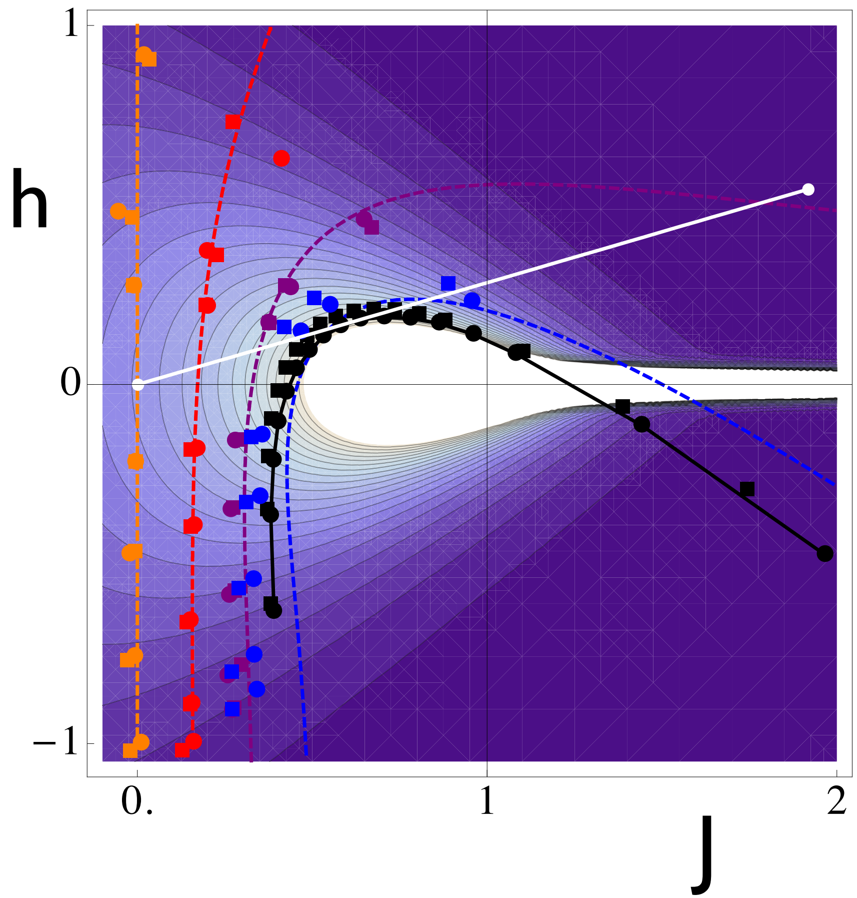
(Random Energy Model
Cook & Derrida 1991)

Maximally informative models are critical

(Mastromatteo+Marsili JSTAT 2012)

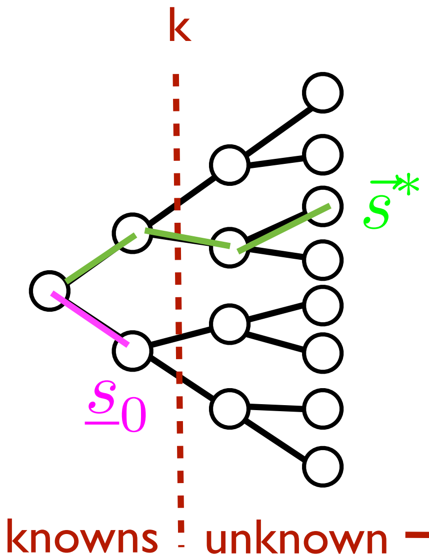
- e.g. s = n binary variables (e.g. spikes from salamander retina)
- Parametric models:
  p(s) = p(s|h,J) = Ising model
- Uniform P{p(s)} maps in a non-uniform P{h,J} that concentrates around critical points
- Intuition (Cramer-Rao):

$$\chi = \frac{\delta s}{\delta h} = \frac{\delta \text{data}}{\delta \text{params}}$$



Extensions:

- What is the analogous of Boltzmann for fat tailed P{v}?
- How relevant and how many should known variables be when P{v} is sub-exponential?
- GREM (directed polymers on trees) optimal resolution/discounting



$$U(\vec{s}) = u^1_{\underline{s}_1} + u^2_{\underline{s}_2|\underline{s}_1} + u^3_{\underline{s}_3|\underline{s}_2,\underline{s}_1} + \ldots + u^m_{\underline{s}_m|\underline{s}_{m-1},\ldots,\underline{s}_1}$$

Discounting: $u^k_{\underline{s}_k|\underline{s}_{k-1},\ldots,\underline{s}_1} \sim \delta^{k-1}, \qquad \delta < 1$

$$\bar{s} \equiv \underline{s}_{\geq k} = (\underline{s}_k, \ldots, \underline{s}_m)$$

$$\underline{s} \equiv \underline{s}_{<k} = (\underline{s}_1, \ldots, \underline{s}_{k-1})$$

Sampling:

(the inverse problem)

Observables (knowns)

$$\max_{\underline{s}} \max_{\bar{s}} U(\underline{s}, \bar{s}) \Rightarrow \underline{s}^*$$

Nature

$$\max_{(\underline{s},\bar{s})} U(\underline{s}, \bar{s})$$

Data M observations

$$\hat{s} = \left(\underline{s}^{(1)}, \ldots, \underline{s}^{(M)}\right)$$

Q: What can I say on $u_{\underline{s}} = E_{\underline{s}}[U(\underline{s},\bar{s})]$?

When is M large enough?

What do samples (typically) look like when M is small?

Where is the information on $u_{\underline{s}}$ in the sample?

- Sample of M observations $\hat{s} = \left(\underline{s}^{(1)}, \ldots, \underline{s}^{(M)}\right)$
- $K_{\underline{s}} = \sum_{1=1}^{M} \delta_{\underline{s}^{(i)},\underline{s}}$ gives a noisy estimate of $u_{\underline{s}}$

$$u_{\underline{s}} \approx c + \beta^{-1} \log K_{\underline{s}}$$

- The information contained in the sample is H[K]

$$H[K] = -\sum_{k} \frac{kN(k)}{M} \log_2 \frac{kN(k)}{M}$$

N(K)=n. of cities of size K

The information content of the city size distribution: how many bits to find Mr X?

- M people in the US, need $\log_2 M$ bits to find Mr X
- If you knew the size K_X of the city where X lives then you'd need $\log_2 [K_X N(K_X)]$ binary questions (i.e. bits).
- If you knew which city s_X X lives in, then you'd need $\log_2 K_X$ bits
- If all individuals live in the same city $K_X=M$ then you don't gain any information either way
- If each individual lives in a different city ($K_X=1$) you don't gain anything if you know K_X you know everything if you know s_X
- Information gain depends on N(K) and the amount of information is given by H[K]

Information gain and entropy

$$H[K] = -\sum_k \frac{kN(k)}{M} \log_2 \frac{kN(k)}{M}$$

$$H[\underline{s}] = -\sum_k \frac{kN(k)}{M} \log_2 \frac{k}{M}$$

$$H[K] = H[\underline{s}] = 0$$

$$H[K] = 0, \quad H[\underline{s}] = \log_2 M$$

What is the most informative N(k) for $0 < H[s] < \log_2 M$?

Maximally informative samples (upper bound)

$$N(k): \quad \max_{\{N(k)\}} H[K]$$

$$\text{s.t.} \quad H[\underline{s}] = H_0$$

$$\sum_k k N(k) = M$$

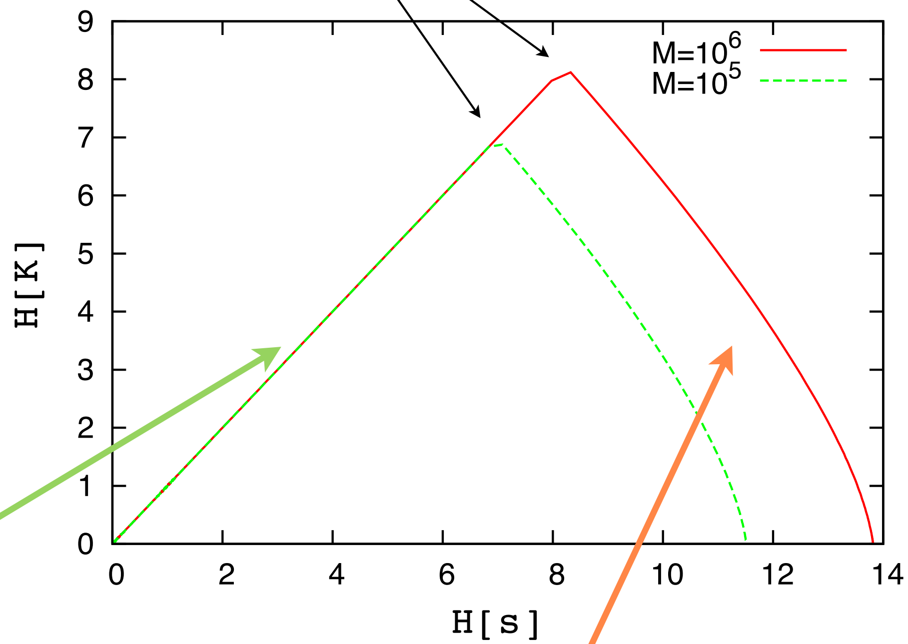
Data processing inequality:

$$H[\underline{s}] - H[K] = \sum_k \frac{kN(k)}{M} \log N(k)$$

$$\geq 0$$

$$N(k) = 1 \qquad \sim \forall k$$

Zipf: $\mu = 2$



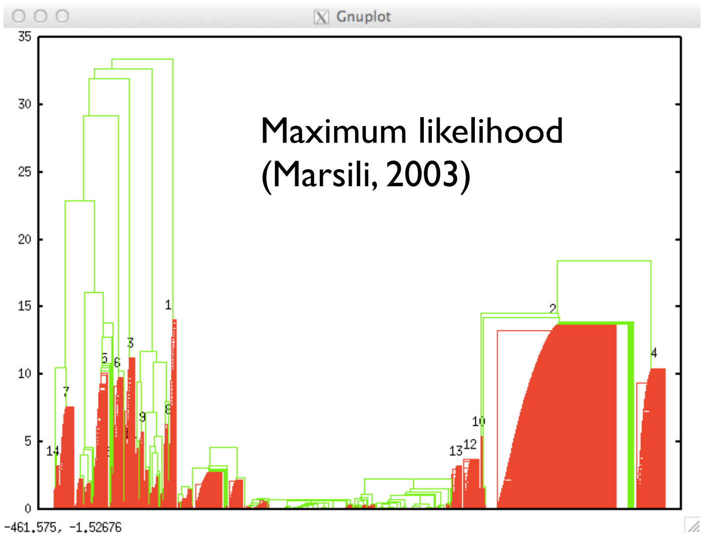
$$N(k) \sim k^{-\mu}$$

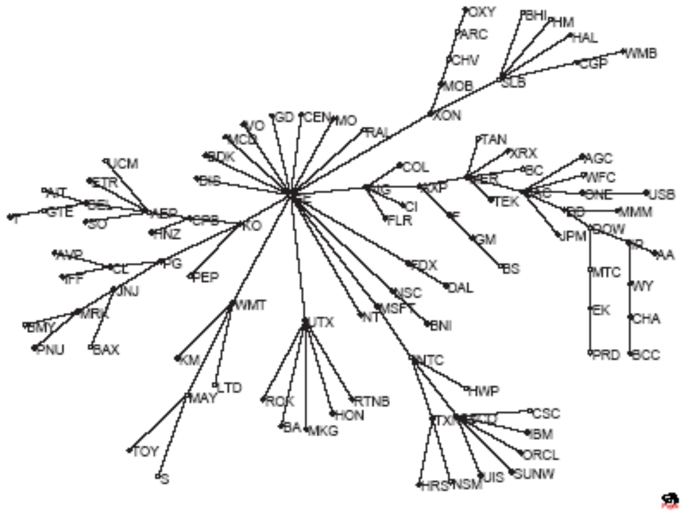
Applications/examples

- Data clustering: Classifying financial stocks
- Keywords in the "Origin of the Species"
- Finding relevant positions in proteins
- Optimal description of the dynamics of a complex system

Finding relevant variables 1:
Classifying 4000 NYSE stocks

- Time series for M=4000 stocks,
  daily returns (1 Jan 1990 - 30 Apr 1999)
- $\underline{s}^{(i)}$ = label of stock i in hierarchical data clustering with N clusters
- Which method?

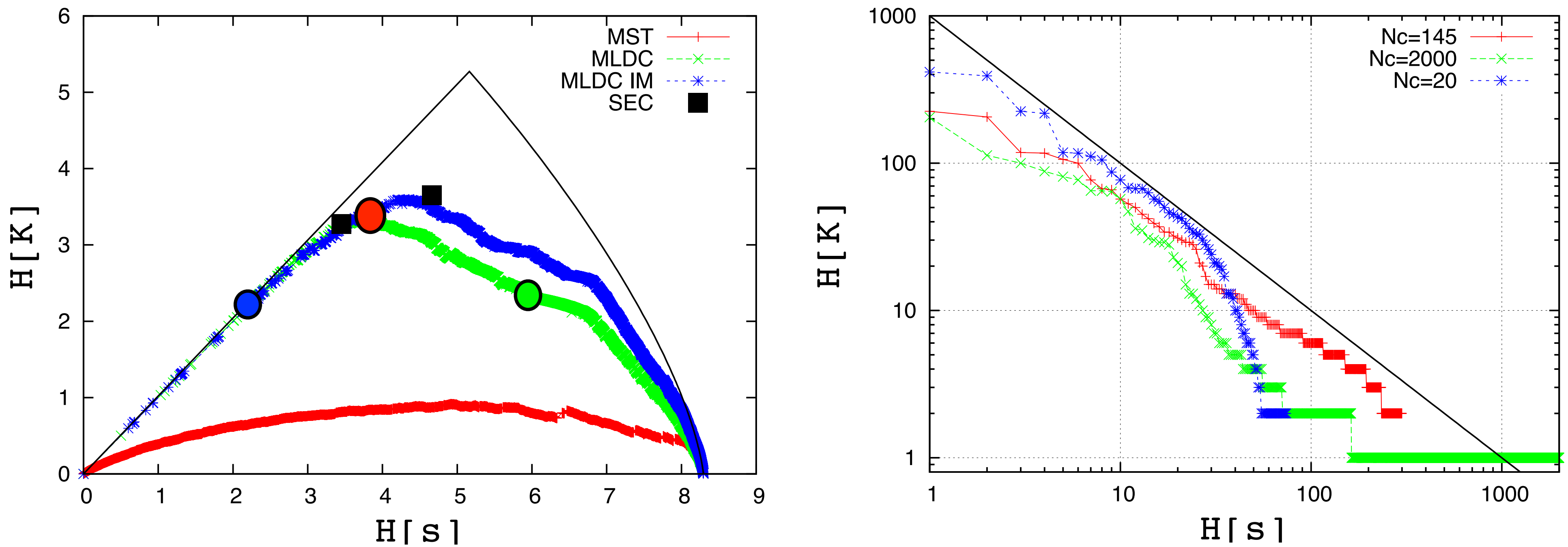




Minimal Spanning Tree (MST)
(Bonanno et. al. 2004, Tumminello et al. 2006)

H[K] can be used to score clustering methods

Data: $x_i(t)$ = (log)return of stock i=1,...,4000 in day t =1/1/90 - 30/4/99



MST = Minimal Spanning Tree
MLDC = Maximum Likelihood Data Clustering
MLDC IM = MLDC on internal modes
SEC = US Security Exchange Commission classification

Finding relevant variables II:
Keywords in text

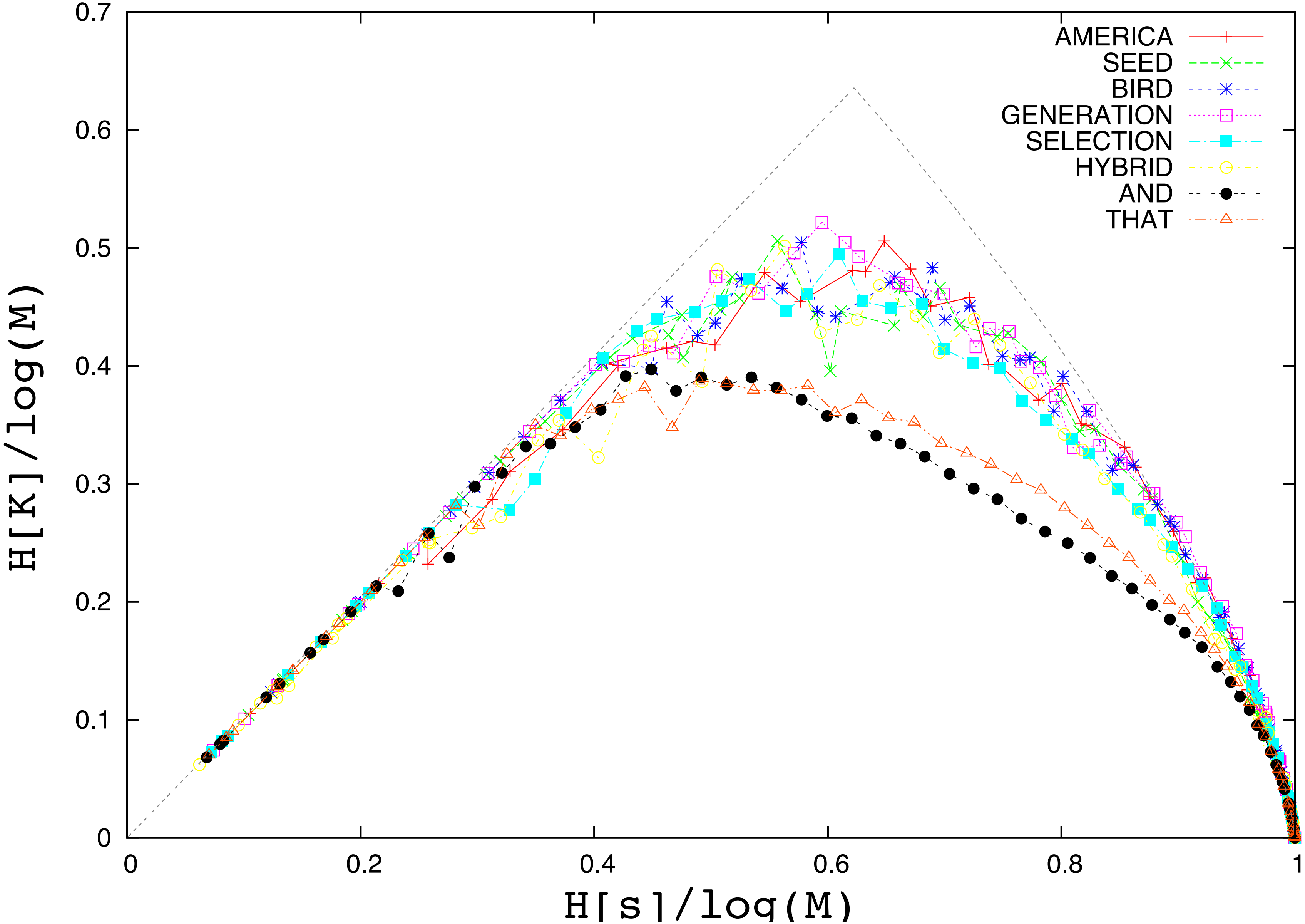
- Text = (w_1,w_2, w_3, ... , w_L) in blocks of B words

- Montemurro, Zanette (2009): relevant words are those whose frequency distribution in blocks differs most from the random distribution.

- K_s=number of times w occurs in block s=1,..,L/B

- Words with larger H[K] are the most relevant (those that are chosen for specific reasons)

The Origin of the Species



Finding relevant variables III:
Choosing relevant positions in proteins

- Protein: amino-acid sequence $\vec{s} = (s_1, \ldots, s_N)$
- Function (e.g. response regulator receptor) is related to sequence (e.g. structure/contacts, active sites, etc)
- Data: Families of homologous proteins in PFAM database. Same function different organisms, different sequences $\vec{s}^{(1)} \ldots \vec{s}^{(M)}$

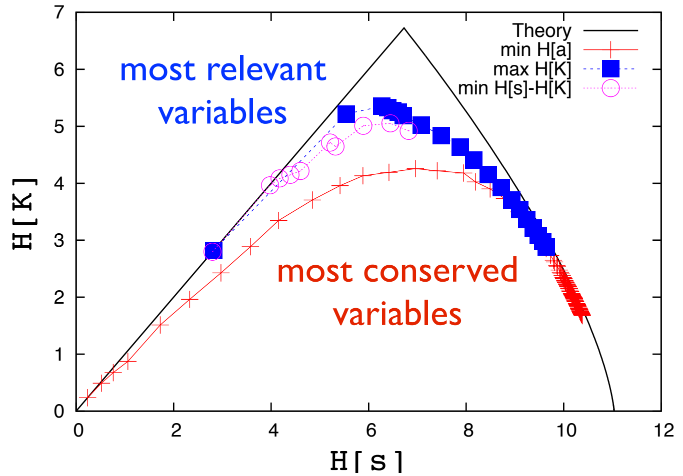
$$\vec{s}^{(i)} = \left(\underline{s}^{(i)}, \overline{s}^{(i)}\right)$$

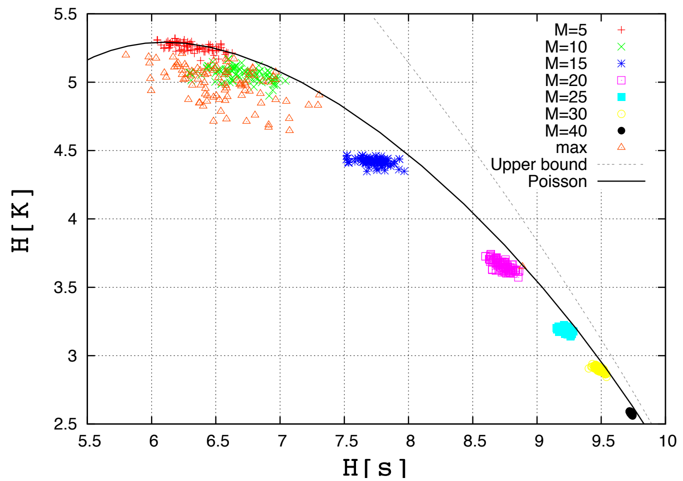
- How to find relevant variables?

1. subsequence of n most conserved amino-acids
2. subsequence that maximizes H[K]

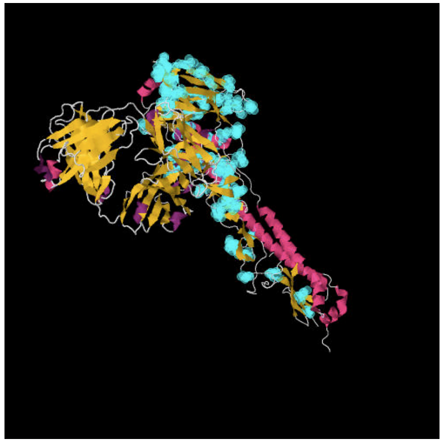
“Most relevant” subsequences

- Relevant variables are not only the most conserved ones
- Over-fitting?

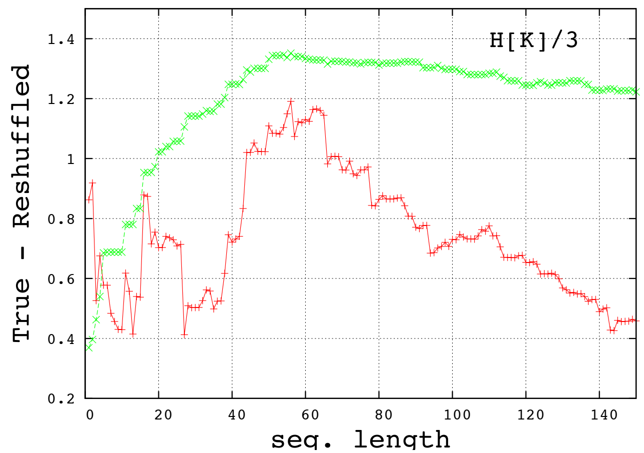


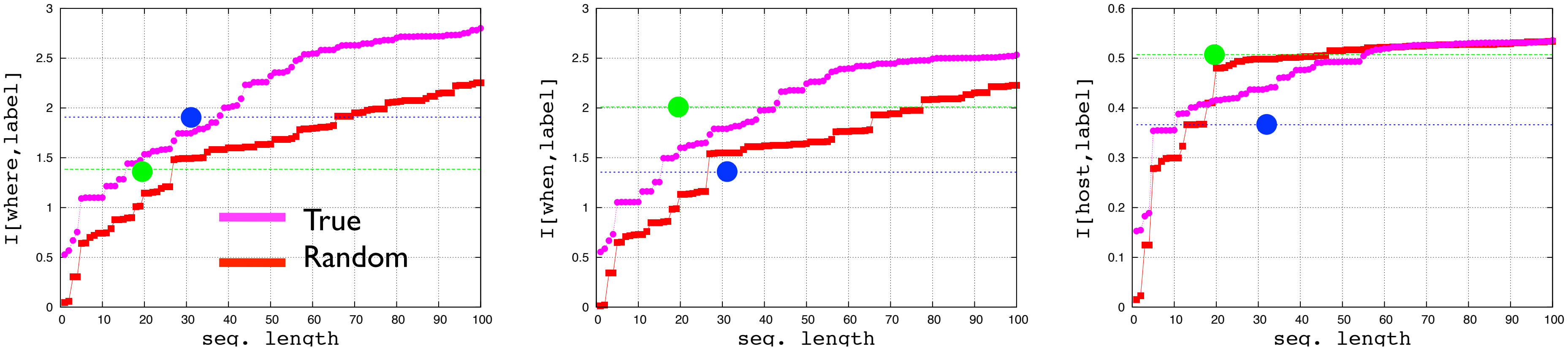


HA1 of H3N2



M=6573, N=328 amino acids
n most relevant positions
- no correlation with known structural or functional sites
- mutual information with annotation=(where, when, host) is comparable to expert classification
- difference with random sequence peaks where H[K] peaks

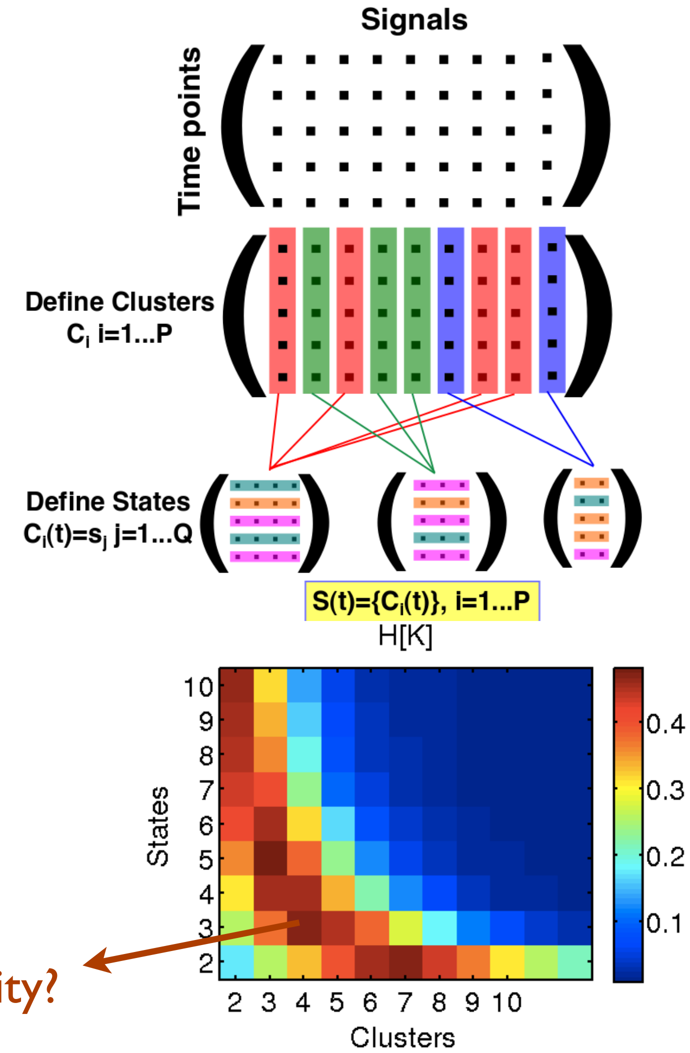




Expert classification: Fitch et al. 1999 (18 sites) Dushoff et al. 2003 (32 sites)

Finding relevant variables IV:
On the dynamics of complex systems

- High dimensional data:
  Brain: 40k voxels, 10k time points
  Finance: 4k stocks, 2k days
- Dimensionality reduction:
  clusters and states
- What resolution?
  How many clusters/states?
- Which are the relevant clusters?



max
predictability?

(work in Progress, Ariel Haimovici, Dante Chialvo, MM)

Summary

- Models may be predictive only when known variables are relevant
- Relevant variables are those for which samples "look critical" (i.e. most informative samples in the under-sampling regime are power laws)
- Zipf's law separates the under-sampling from well sampled regimes
- H[K] vs H[s] plot can be useful
  - to find relevant variables, keywords
  - to score clustering methods
  - ...
- Model free method

Thanks

arXiv.org > physics > arXiv:1301.3622

Physics > Data Analysis, Statistics and Probability

On sampling and modeling complex systems

Matteo Marsili, Iacopo Mastromatteo, Yasser Roudi

(Submitted on 16 Jan 2013 (v1), last revised 1 Nov 2013 (this version, v4))

Journal of Statistical Mechanics: Theory and Experiment > Volume 2013 > September 2013

Matteo Marsili *et al J. Stat. Mech.* (2013) P09003 doi:10.1088/1742-5468/2013/09/P09003

On sampling and modeling complex systems

Matteo Marsili[1], Iacopo Mastromatteo[2] and Yasser Roudi[3,4]