Entropy and Compression for Sparse Graphs with Vertex Names

David Aldous

February 7, 2012

Consider a graph with

- N vertices
- $O(1)$ average degree
- vertices have distinct "names", strings of length $O(\log N)$ from a fixed finite alphabet.

Envisage some association between vertex names and graph structure, as in

- phylogenetic trees on species
- road networks.

I claim this is the "interesting" setting for data compression of sparse graphs, because the "entropy" of both the graph structure and of the names has the same order, $N \log N$. [Will say more later].

Lots of existing work is somewhat related to the setting above – see the survey by Szpankowski - Choi *Compression of graphical structures: Fundamental limits, algorithms, and experiments*.

This is a conceptual talk, arguing that (at some position on the theory-applications spectrum) this is the right setting to study.

I have no big theorem, just one technical lemma.

The first half of the talk gives my take on classical Shannon theory, both math and conceptual. But to digress for 2 slides, I teach a course "probability and the real world" with the following style.

The course consists of 20 lectures, on topics chosen to be maximally diverse. Here are my desiderata for an ideal topic.

- It is appropriate for the target audience: those interested in the relation between mathematics and the real world, rather than those interested in the mathematics itself.
- There is some concrete bottom line conclusion, which can be said in words . . .
- . . . but where mathematics has been used to derive conclusions . . .
- . . . . . and where mathematics leads to some theoretical quantitative prediction that my students can test by gathering fresh data.
- There is available "further reading", both non-technical and technical, that I can recommend to students.

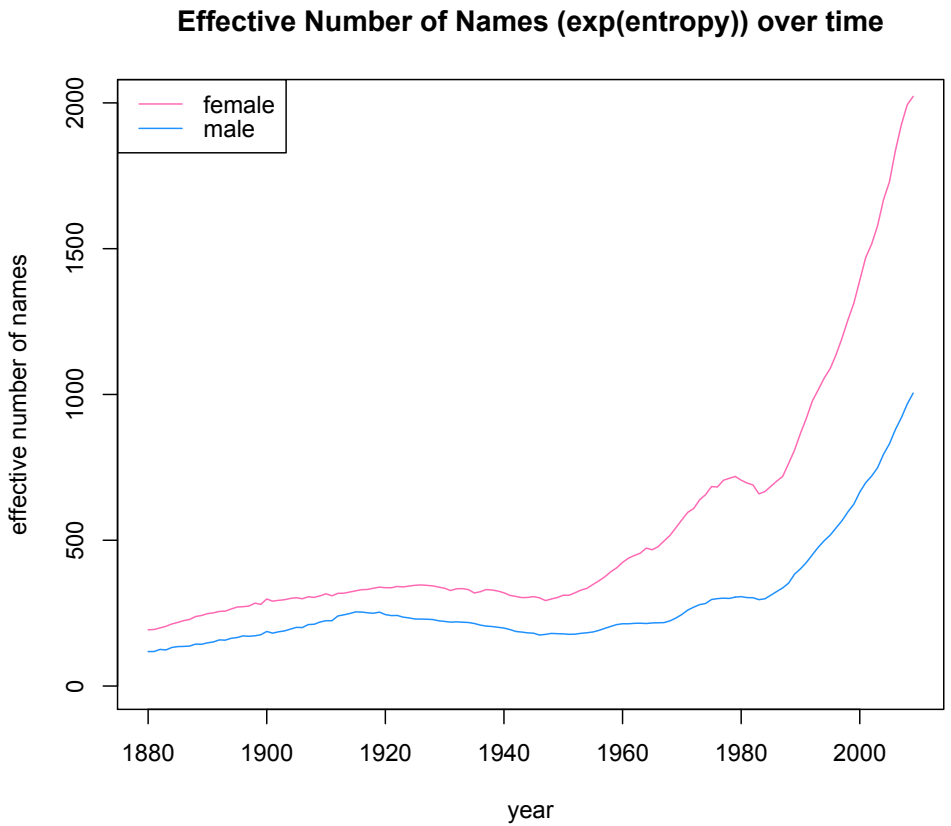
Very few topics permit all this, so the actual lectures fail to attain the ideal!

Here are the topics from 2011, and student feedback "like minus dislike".

(22) Psychology of probability: predictable irrationality
(18) Global economic risks
(17) Everyday perception of chance
(16) Luck
(16) Science fiction meets science
(14) Risk to individuals: perception and reality
(13) Probability and algorithms.
(13) Game theory.
(13) Coincidences and paradoxes.
(11) So what do I do in my own research? (spatial networks)
(10) Stock Market investment, as gambling on a favorable game
(10) Mixing and sorting
(9) Tipping points and phase transitions
(9) Size-biasing, regression effect and dust-to-dust phenomena
(6) Prediction markets, fair games and martingales
(6) Branching processes, advantageous mutations and epidemics
(5) Toy models of social networks
(4) The local uniformity principle
(2) Coding and entropy
(-5) From neutral alleles to diversity statistics.

For any probability distribution $\mathbf{p} = (p_s) = (p(s))$ on any finite set S, its entropy is the number

$$\mathrm{ent}(\mathbf{p}) = -\sum_s p_s \log p_s.$$



For any probability distribution $\mathbf{p} = (p_s) = (p(s))$ on any finite set S, its entropy is the number

$$\mathrm{ent}(\mathbf{p}) = -\sum_s p_s \log p_s.$$

What's the point of this definition?

Write $\mathbf{B}$ for the set of binary strings $\mathbf{b} = b_1 b_2 \dots b_m$ and $\mathrm{len}(\mathbf{b}) = m$. Then (Huffman code) given X with distribution $\mathbf{p}$, there exists a coding $f_{\mathbf{p}} : S \to \mathbf{B}$ such that

$$\mathbb{E}\ \mathrm{len}(f_{\mathbf{p}}(X)) \approx \mathrm{ent}_2(\mathbf{p}).$$

Note

- "a coding" just means a $1-1$ function.
- S is arbitrary.
- $f_{\mathbf{p}}$ depends (very much) on $\mathbf{p}$.

A quick rant

90 years of bad teaching of freshman Statistics, and shorthand like "the data is statistically significant", has led to the widespread blunder of thinking that statistical significance is an attribute of observed data; of course it's really an attribute of the hypothetical probability model that may or may not be generating the data.

I suspect this all started with the bad choice of using the same words (mean, s.d., etc) for data and for models.

Here's an example with entropy.

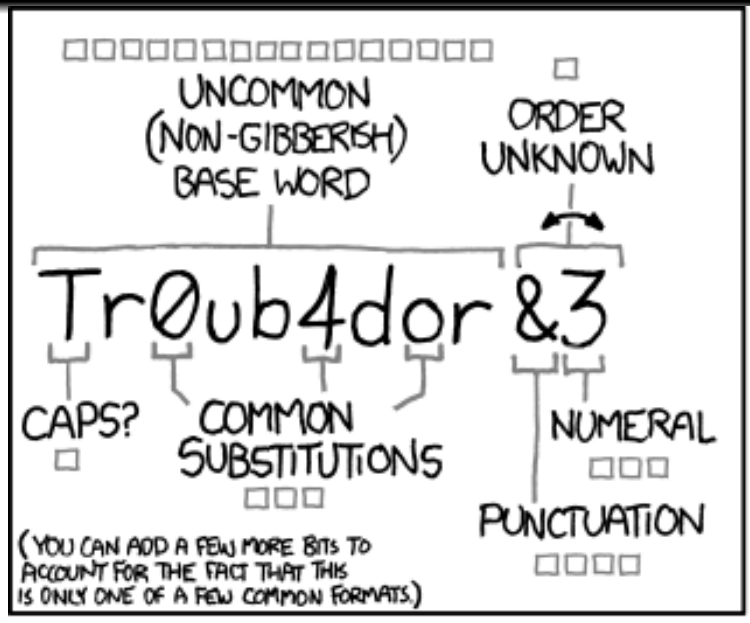
Suppose, from a library of about one million books, you pick one book uniformly at random and write out the entire text of that book. What is the entropy of what you write?

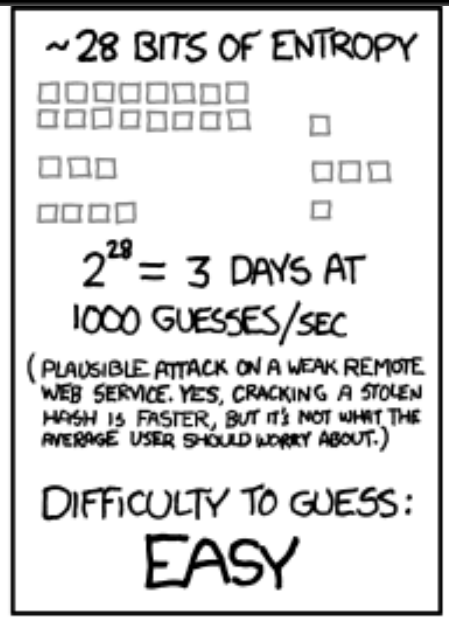
Answer [according to our definition]: 20 bits.

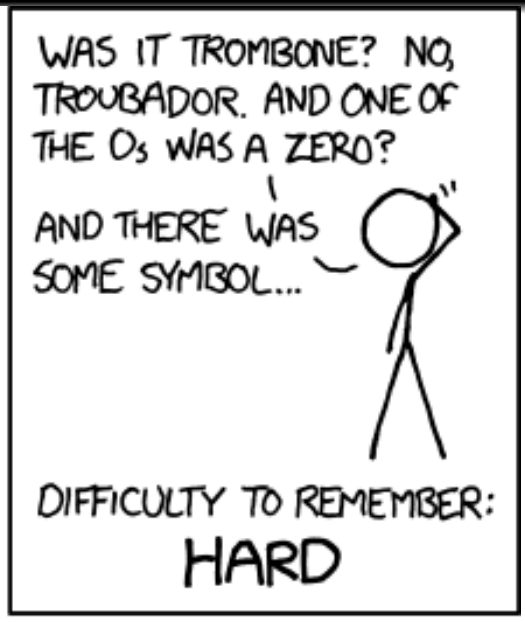
On the other hand, widely claimed "entropy of English language is about 1.3 bits per letter", so the text of a book with 300,000 letters should have entropy about 400,000 bits.

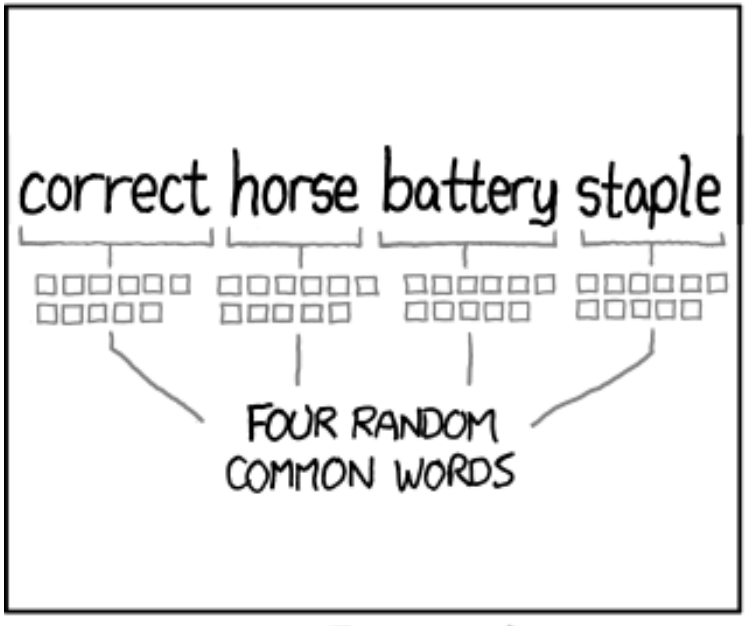
Point (again): there is a distinction between actual data and hypothetical probability models by which the data might have been generated.

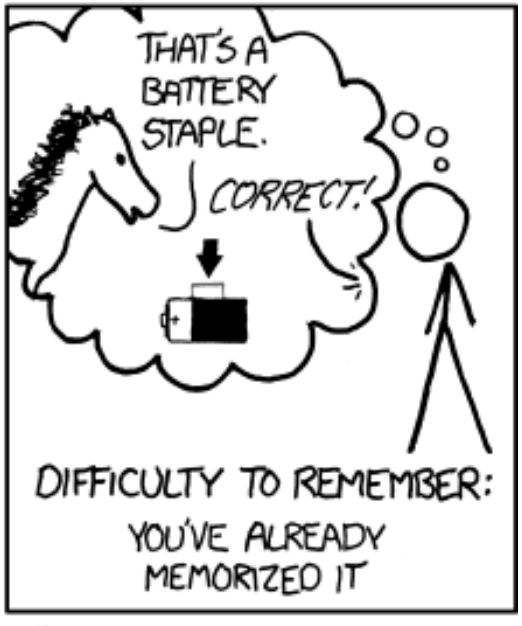
Here's a less artificial example. From the Hall of Fame for back-of-an-envelope calculations.











THROUGH 20 YEARS OF EFFORT, WE'VE SUCCESSFULLY TRAINED EVERYONE TO USE PASSWORDS THAT ARE HARD FOR HUMANS TO REMEMBER, BUT EASY FOR COMPUTERS TO GUESS.

In fact I claim there is a good analogy between issues of password strength and issues of data compression of sparse graphs.

It's easy/fun to invent probability models of how
- people might create passwords
- sparse labeled graphs might arise

and compute entropy for the model; this tells us (roughly speaking how easy it is for
- an adversary to find the password
- us to compress the graph

for a typical realization of a **known** model.

But in neither case are we engaging real data.

Classical Shannon setting.

A stationary ergodic sequence $\mathbf{X} = (X_i)$, values in finite alphabet, has an *entropy rate* H, characterized in several ways.

$$H = \lim_{k\to\infty} k^{-1}\mathrm{ent}(X_1, \ldots, X_k) \quad (1)$$

$$H = -\mathbb{E}\log p(X_0 \in \cdot | X_{-1}, X_{-2}, \ldots) \quad (2)$$

The former implies that, given the distribution of $\mathbf{X}$ and ε, we can use *block codes* to code $X_1, \ldots, X_n$ as a binary sequence of length L_n with

$$\mathbb{E}L_n \le (H + \varepsilon + o(1))n \text{ as } n \to \infty.$$

What if the distribution of $\mathbf{X}$ is unknown? As a "horrible hack" one could use some initial portion to estimate the distribution of k-blocks and then use the optimal code for that estimated distribution. In the 1970s more elegant *Lempel-Ziv* type algorithms were devised which are "universal" in the sense that

$$\mathbb{E}L_n = (H + o(1))n \text{ as } n \to \infty.$$

The preceding is just mathematics, but . . . . . .

The magic of classic Shannon theory

Specific to "text" – length-N strings from a known finite alphabet. Real text comes from some source which does not fit the dictionary definition of the word *random* –
proceeding, made, or occurring without definite aim, reason, or pattern.
But we can pretend source is random, in the sense of stationary ergodic, and apply Lempel-Ziv algorithm to compress. One cannot check the claim

$$\mathbb{E}L_n = (H + o(1))n \text{ as } n \to \infty$$

because we have no prior definition of H.
A simple checkable theoretical prediction is that if you take a long piece of text, split it into two halves of equal uncompressed length, and compress each half separately, then the two compressed halves will be approximately the same length.

	uncompressed	compressed
first half of Don Quixote	1109963	444456
second half of Don Quixote	1109901	451336

After this long ramble we come to the point of this talk:

how much of this theory for sequences can we push over to the setting of sparse graphs with vertex-names?

There are two technical reasons why I chose this specific setting.

Reason 1. Because **sparse** we may assume there is a **local weak limit** of the graph structure, a random infinite rooted unlabeled graph with an analog of the stationarity property. One could develop a theory for unlabeled graphs, or with labels from a fixed finite alphabet, but this doesn't seem so relevant to actual data.

Recall: entropy is about constants, not orders of magnitude

For typical English text with N letters, the most naive way to store in binary uses $c_1 N$ bits, the most efficient way uses $c_2 N$ bits.

Similarly, for a typical sparse graph on vertices $1, \ldots, N$, the most naive way to store in binary uses $c_1 N \log N$ bits, the most efficient way uses $c_2 N \log N$ bits.

And (unless very sparse) the same holds for unlabeled sparse graphs. So there is order $N \log N$ entropy from graph structure.

Having vertex-names be strings of length $O(\log N)$ from a fixed finite alphabet means there is also order $N \log N$ entropy from the names.

Reason 2. Having the same (order of) entropy from both ingredients is the "interesting" case where you need to pay attention to both.

As already mentioned, easy/fun to invent probability models and show

$$\mathrm{ent}(\mathcal{G}_N) \sim cN \log N$$

for some entropy rate c.
(Cute observation: in this $N \log N$ world, c does not depend on base of logarithms).

(Backing up one slide)

how much of classical Shannon theory for sequences can we push over to the setting of sparse graphs with vertex-names?

Because **sparse** we may assume there is a **local weak limit** of the graph structure, a random infinite rooted unlabeled graph with an analog of the stationarity property.

Here's a simple example to show the most we can hope to get out of the LWC approach.

Fix a finite alphabet, and for each n let $\mathbf{X}_n = (X_{n,i}, 1 \le i \le n)$ be a random string, and make (only) the following assumption. Take U_n uniform on $\{1, \ldots, n\}$ and suppose that for each $k \ge 1$

$$(X_{n,U_n+i}, -k \le i \le k) \stackrel{d}{\to} (Z_i, -k \le i \le k) \text{ as } n \to \infty \qquad (3)$$

for some doubly-infinite random sequence (Z_i). This is of course equivalent to LWC of $\mathbf{X}_n$ considered as the linear path graph with vertex-labels in the alphabet. Then (Z_i) is stationary and has some entropy rate H. Think of H as a "local" property of the random string.

A straightforward consequence of subadditivity of entropy is that

$$\limsup_n n^{-1} \mathrm{ent}(\mathbf{X}_n) \le H.$$

And we can remove assumption (3) by taking $H^* :=$ max of H over all subsequential limits.

$$\limsup_n n^{-1}\mathrm{ent}(\mathbf{X}_n) \le H^*. \tag{4}$$

Note that, for the base-b expansion of a real number x, the assertion "x is normal in base b" is exactly the assertion that, taking $\mathbf{X}_n$ as the first n digits, (3) holds with limit process i.i.d. uniform on $\{0, 1, \ldots, b-1\}$.

This dramatically shows that in general we do not have equality in (4).

Roughly speaking, this is what Lempel-Ziv and any conceivable "universal" algorithm does – compress to length nH but not to length $\mathrm{ent}(\mathbf{X}_n)$.

This is the "real story" of Shannon, without assuming data is from an infinite stationary sequence.

Informal statement of actual minor result.

Given $(\mathcal{G}_N)$ – some model of random sparse graphs with vertex-names – we are interested in entropy rate

$$c := \limsup_N \frac{\mathrm{ent}(\mathcal{G}_N)}{N \log N}$$

Consider the subgraph on a s-vertex neighborhood of a uniform random vertex, and suppose the entropy of this random subgraph grows as $c_s \log N$, as $N \to \infty$ for fixed s. Then

$$c \leq \limsup_{s \to \infty} s^{-1} c_s \tag{5}$$

provided the graph is not expander-like, in the sense that we can choose large $K(\varepsilon)$ and partition vertices into clusters of size $\approx K(\varepsilon)$ so that the proportion ε of all edges linking different clusters is $\leq \varepsilon$ for large N.

Not surprising! The argument rests on the following "size-biasing" lemma.

Consider a collection $\{C_i(A_i), i \geq 1\}$ where (A_i) is a partition of $\{1, \ldots, N\}$ and $C_i(A_i)$ is an object (from some specified finite set of possible objects) comprising A_i and some extra structure. Now let $\mathbf{C}$ be a random such collection, regarded as an unordered set. Take U uniform random on $\{1, \ldots, N\}$, independent of $\mathbf{C}$, and for the $I = I(U)$ such that $U \in A_I$ write $A^* = A_I$ and $C^*(A^*) = C_I(A_I)$. (In words, $C^*(A^*)$ is a size-biased selection from $\mathbf{C}$). Let $q_1(s) = P(|C^*(A^*)| = s)$ and let Z_1^s have the conditional distribution of $C^*(A^*)$ given $|A^*| = m$.

Lemma

$$\mathrm{ent}(\mathbf{C}) \leq (N+1) \sum_{s \geq 1} q_1(s) \frac{\Gamma - \log q_1(s) + \mathrm{ent}(Z_1^s)}{s+1} + N \log \frac{N}{N-M} \quad (6)$$

where M is the expected number of components in $\mathbf{C} = (C_i(A_i), i \geq 1)$ and Γ is a numerical constant.

This lemma is key to the proof (for non-expander graphs) that

$$c_{global} \leq c_{local}$$

where

$$c_{global} := \limsup_N \frac{\mathrm{ent}(\mathcal{G}_N)}{N \log N}$$

$$c_{local} := \lim_{s \to \infty} s^{-1} \limsup_N \frac{\mathrm{ent}(B_N(s))}{N}$$

for $B_N(s)$ the subgraph on a s-vertex n'hood of a random vertex. That is, an inequality between "global" and "local" definitions of entropy rate. Extending this inequality to expander-type graphs is a "do-able" open problem.

Thesis: a general-purpose algorithm cannot do better than c_{local}. So don't try.

Two Research Projects

1. Give some abstract condition on a model, analogous to "ergodic" in Shannon theory, for c_{global} to exist as a $N \to \infty$ limit (not just *limsup*) and for $c_{global} = c_{local}$.

This should just be some very weak "no long range dependence" condition, without any assumption of models built from independent pieces. Part of the condition will be LWC of the unlabeled graphs.

2. Give some stronger condition on a model and a pseudo-universal algorithm which compresses realizations from such models to $(c + o(1))N \log N$.

To me these are "serious" problem that we don't know how to do.

What we do know how to do is to study particular models, where the arguments for the upper bound in $\mathrm{ent}(\mathcal{G}_N) \sim cN \log N$ are often themselves algorithmic. Here are 2 slides.

Simpler setting 1: Graphs with labels $1, \ldots, N$ as binary strings.

1. There is a simple universal algorithm that compresses to length $\frac{1}{2}\bar{d} N \log N + O(N)$, where $\bar{d}$ is average degree.

2. For sparse Erdös-Rényi random graph $\mathcal{G}(N, \alpha/N)$ we have entropy $\sim \frac{1}{2}\alpha N \log N$.

3. Intuitively, any probability model where there is no strong association of adjacent labels will have entropy rate $c = \frac{1}{2} \times$(ave degree)., e.g. configuration model.
Next is an example which does have strong association.

4. Construct a random tree $\mathcal{T}_N$ as follows. Take $V_3, V_4, \ldots, V_N$ independent uniform on $\{1, \ldots, N\}$. Link vertex 2 to vertex 1. For $k = 3, 4, \ldots, N$ link vertex k to vertex $\min(k-1, V_k)$.
It is known (by an indirect argument – not obvious) that if one first constructs $\mathcal{T}_N$, then applies a uniform random permutation to the vertex-labels, the resulting random tree $\mathcal{T}_N^*$ is uniform on the set of all N^{N-2} labelled trees. Here $(\mathcal{T}_N^*)$ has entropy rate $c = 1$ whereas $(\mathcal{T}_N)$ has entropy rate $c = 1/2$.

For a more geometric setting, start with the $N^{1/2} \times N^{1/2}$ discrete torus graph with its usual coordinate labels. Make a "small worlds" model in which extra edges (v, w) are present with chance proportional to a negative power of the distance $||w - v||_2$. We can parametrize so that mean degree $= \alpha$ and the length distribution L of these extra edges has $P(L = \ell) \asymp \ell^{-\gamma},\ 1 < \ell \leq O(L^{1/2})$.

If $\gamma > 2$ then $EL = O(1)$ and entropy grows as $O(N)$.

If $\gamma < 2$ then $\log L \sim \log N^{1/2}$ and entropy grows as $\frac{\alpha}{2} N \log N$ as in the non-geometric setting.

In the critical case $\gamma = 2$ we have $\log L$ uniform on $[0, \frac{1}{2}] \times \log N$ and the entropy grows as $\frac{\alpha}{4} N \log N$.

(no more slides – improvise on blackboard!)

1. First non-trivial model.
2. Lempel-Ziv style algorithms.