Entropy for Sparse Random Graphs With Vertex-Names

David Aldous

11 February 2013

Research strategy (for old guys like me):

if a problem seems . . .

do-able $\Longrightarrow$ give to Ph.D. student
maybe do-able $\Longrightarrow$ give to post-doc
clearly not do-able $\Longrightarrow$ think about it myself.

I'm thinking about a topic in a very classical way (Shannon):

- lossless data compression
- ignoring computational complexity of coding

What is a network?

- A **graph** is a well-defined mathematical object – vertices and edges etc
- A **network** is a graph with context-dependent extra structure.

I want to consider a certain simple-but-interesting notion of “extra structure”, which is the notion

vertices have names.

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 [**key point of this set-up**] this is the "interesting as theory" setting for data compression of sparse graphs, because the "entropy" of both the graph structure and of the names has the same order, $N \log N$. And we want to exploit the association.

Formal setup.

Define a finite set $S = S(N, A, \beta, \alpha)$; an element of S is a network with

- N vertices
- ave degree $\leq \alpha$ (at most $\alpha N/2$ edges)
- finite alphabet $\mathbf{A}$ of size A
- vertices have distinct names – strings from $\mathbf{A}$ of length $\leq \beta \log_A N$.

Here $A \geq 2$, $0 < \alpha < \infty$, $1 < \beta < \infty$ are fixed and we study as $N \to \infty$.

Easy to see how big S is:

$$\log |S(N, A, \beta, \alpha)| \sim (\beta - 1 + \tfrac{\alpha}{2}) N \log N.$$

For some given model of random network $\mathcal{G}_N$ we expect

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

for some model-dependent entropy rate $0 \leq c \leq (\beta - 1 + \frac{\alpha}{2})$.

[cute observation: value of c doesn't depend on base of log.]

In this setting there are two "extreme" lines of research you might try.

Clearly do-able: Invent probability models and calculate their entropy rate.

With Nathan Ross we have a little paper doing this (search on "arxiv aldous entropy"). Will show 3 slides about this, soon.

Clearly not do-able: Design an algorithm which, given a realization from a probability model, compresses optimally (according to the entropy of the probability model) without knowing what the model is ("universal", like Lempel-Ziv for sequences).

So my goal is to find a project laying between these extremes.

[Pedagogical aside: this "$N \log N$" world could provide projects for a first course in IT.]

What is in the existing literature?

A: “More mathematical”. Topic called “graph entropy” studies the number of automorphisms of a N-vertex unlabelled graph. See the 2012 survey by Szpankowski - Choi *Compression of graphical structures: Fundamental limits, algorithms, and experiments*.

B: “More applied”. Seeking to exploit the specific structure of particular types of network.

Boldi - Vigna (2003). *The webgraph framework I: Compression techniques*.

Chierichetti - Kumar - Lattanzi - Mitzenmacher - Panconesi - Raghavan (2009). *On compressing social networks*.

Clearly do-able project: Invent probability models and calculate their entropy rate.

Many hundreds of papers in “complex networks” study probability models for random N-vertex graphs, implicitly with vertices labeled $1, \ldots, N$. Two simplest ways to adapt to our setting:
(i) write integer labels in binary.
(ii) assign completely random distinct names.

For the best-known models – Erdős-Rényi, small worlds, preferential attachment, random subject to prescribed degree distribution, . . . – it is almost trivial to calculate the entropy rate.

But none of these models has a very “interesting” association between graph structure and names. Here is our best attempt at a model that does, and that requires a non-trivial calculation for entropy rate.

Model: Grow sparse Erdős-Rényi $\mathcal{G}(N, \alpha/N)$ sequentially; an arriving vertex is linked to a previous vertex with chance α/N.

Vertex 1 is given a uniform random length-L_N A-ary name, where (as in other models) $L_N \sim \beta \log_A N$.

A subsequent vertex arrives with a tentative name $\mathbf{a}^0$; gets linked to $Q \geq 0$ previous vertices with names $\mathbf{a}^1, \ldots, \mathbf{a}^{Q_n}$; assign the name obtained by picking the letter in each coordinate $1 \leq u \leq L_N$ uniformly from the $1 + Q$ letters $a_u^0, a_u^1, \ldots, a_u^{Q_n}$.

This gives a family $(\mathcal{G}_N)$ parametrized by (A, β, α). One can calculate its **Entropy rate**:

$$-1 + \frac{\alpha}{2} + \beta \sum_{k \geq 0} \frac{\alpha^k J_k(\alpha) h_A(k)}{k! \log A}$$

where

$$J_k(\alpha) := \int_0^1 x^k e^{-\alpha x} dx$$

and the constants $h_A(k)$ are defined as follows:.

$$h_A(k) := A^{-k} \sum_{(a_1,\ldots,a_k)\in\mathbf{A}^k} \mathrm{ent}(\mathbf{p}^{[a_1,\ldots,a_k]}), \tag{1}$$

and where $\mathbf{p}^{[a_1,\ldots,a_k]}$ is the probability distribution $\mathbf{p}$ on $\mathbf{A}$ defined by

$$p^{[a_1,\ldots,a_k]}(a) = \frac{1 + A \times |\{i : a_i = a\}|}{(1+k)A}.$$

Question: Can you be more creative than us in inventing such models?

[repeat earlier slide]

In this setting there are two "extreme" lines of research you might try.

Clearly do-able: Invent probability models and calculate their entropy rate.

......

Clearly not do-able: Design an algorithm which, given a realization from a probability model, compresses optimally (according to the entropy of the probability model) without knowing what the model is ("universal", like Lempel-Ziv for sequences).

In the rest of the talk I will outline a "maybe do-able" project laying between these extremes.

Background: Shannon without stationarity

Finite alphabet $\mathbf{A}$. For each N we will be given an $\mathbf{A}$-valued sequence $\mathbf{X}^{(N)} = (X_i^{(N)}, 1 \leq i \leq N)$. No further assumptions.

Question: How well can a "universal" data compression algorithm do?
Answer: Can guarantee that

$$\limsup_N N^{-1}(\text{compressed length of } \mathbf{X}^{(N)}) \leq c^*$$

where c^* is the "local entropy rate" defined as follows.

First suppose we have "local weak convergence", meaning:

- take $U^{(N)}$ uniform on $1, \ldots, N$;
- there is a process $\mathbf{X}^*$ which is the limit in the sense: for each k,

$$(*) \quad (X^{(N)}_{U^{(N)}+i}, 1 \le i \le k) \xrightarrow{d} (X^*_i, 1 \le i \le k).$$

Then $\mathbf{X}^*$ is stationary and has some entropy rate c. Define $c^* = c$.

In general (*) may not hold but we have compactness; different stationary processes may arise as different subsequential limits; define c^* as [essentially] the *sup* of their entropy rates.

Not a particularly useful idea for sequences, because stationarity is a plausible assumption and gives stronger results. But I claim it is a natural way to start thinking about data compression in our setting of sparse graphs with vertex-names.

Given rooted unlabeled graphs g_N and g_∞ where g_∞ is locally finite, there is a natural notion of *local convergence*

$$g_N \to_{local} g_\infty$$

meaning convergence of restrictions to within fixed distance from root. This induces a notion of convergence in distribution for random such graphs

$$\mathcal{G}_N \to_{local} \mathcal{G}_\infty \text{ in distribution.} \tag{2}$$

Now given **un**rooted unlabeled **random** graphs $\mathcal{G}_N$ and a rooted unlabeled **random** graph $\mathcal{G}_\infty$, we have a notion called *local weak convergence* or *Benjamini-Schramm convergence*

$$\mathcal{G}_N \to_{LWC} \mathcal{G}_\infty$$

defined by:

- first assign a uniform random root to $\mathcal{G}_N$
- then require (2).

Some background for LWC

- Random N-vertex 3-regular graph converges to the infinite 3-regular tree.
- The rooted binary tree of height h converges to a different tree.
- The limit random rooted graph $\mathcal{G}_\infty$ always has a property (*unimodular*) analogous to stationarity for sequences.
- In contrast (to stationarity) it is not obvious that every unimodular $\mathcal{G}_\infty$ is a local weak limit of some sequence of finite random graphs (Aldous-Lyons conjecture).
- Presumably one could develop a theory of entropy/coding for sparse graphs where vertices have marks from a fixed finite alphabet, in terms of an entropy rate for the limit infinite marked graph.

Program: I want to make correct version of following conjecture about sparse random graphs-with-vertex-names $\mathcal{G}_N$.

Write $\mathcal{G}_{N,k}$ for the restriction of $\mathcal{G}_N$ to a k-vertex neighborhod of a uniform random vertex. For fixed k we expect

$$\mathrm{ent}(\mathcal{G}_{N,k}) \sim c_k \log N \text{ as } N \to \infty$$

and then we expect a limit

$$k^{-1} c_k \to c^* \text{ as } k \to \infty.$$

Conjecture. As with sequences, this "local entropy rate" is an upper bound for global entropy:

$$\lim_N \frac{\mathrm{ent}(\mathcal{G}_N)}{N \log N} \le c^*$$

with equality under a suitable "no long range dependence" condition analogous to "ergodic" for sequences.

Such a "theory" result **would give a target** for "somewhat universal" compression algorithms.