Defining and computing entropy for networks

Henrique Ferrolho - s1683857, Team: Alex Hoppen, Charles Desmonty

***Abstract*—We will define two different measures to calculate the entropy of a graph network, i.e. its level of disorder. Both these measurements will use very basic properties of the graph's structure itself.**

We will run test trials on these measurements with *Erdős-Rényi* and *Watts-Strogatz* random graphs. Furthermore, we will run more trials using real-world datasets of very large road networks.

Finally, we will introduce the concept of *Kolmogorov Complexity*, and suggest an approach that can be used as a different entropy measurement for future works.

I. Introduction

Entropy is the lack of *order* or *predictability*, it is a measure of *randomness* or *disorder* in a system - the greater the disorder the higher the entropy.

We are going to define a measure in order to calculate the entropy of a network graph. But before we do that, we need to agree on what makes a network more or less organized. How does the human notion of order apply to graphs, i.e. having two graphs A and B, what properties make graph A more or less organized and structured than graph B?

Entropy must be carefully defined both as a measure and as a concept.

In this paper we will briefly go through related works on the subject from the past.

Furthermore, we will define two different definitions for entropy of a graph.

We will then proceed to test them not only on two types of random graphs: *Erdős-Rényi* graphs [3], and *Watts-Strogatz* graphs [6], but also on two real-world road network datasets [4].

II. Related work

Trying to come up with a robust definition for the entropy of a graph has been an extensively explored topic during the past: Dehmer [1] and [2].

Zenil and Kiani [7] have explained how information-theoretic measures are not independent of the way in which a graph can be observed. They explore Shannon Entropy as a computable measure, and introduce recursive and non-recursive uncomputable graphs to demonstrate the weaknesses of measures of complexity.

III. Graph entropy

A. Notation

Let $\mathcal{G} = (V, E)$ be a graph where V is the set of *nodes* and E the set of *edges*. Furthermore, let *deg(v)* denote the *degree* of a node $v \in V$, and *avg_deg(V)* the average degree of the graph's node set.

$$avg\_deg(V) = \frac{1}{|V|} \sum_{v \in V} deg(v) \tag{1}$$

Let $C(V)$ be the subset of nodes in V that have a degree greater than 0. Then, let $distinct(C(V))$ denote the number of distinct values in $C(V)$.

We are now going to define two different ways to calculate the entropy of a graph: $\mathcal{H}1(\mathcal{G})$ and $\mathcal{H}2(\mathcal{G})$.

The entropy $\mathcal{H}1$ of a graph $\mathcal{G}$ is given by the quotient of the distinct number of node degrees in the set minus one, and the total number of nodes in the graph.

$$\mathcal{H}1(\mathcal{G}) = \frac{distinct(C(V)) - 1}{|V|} \tag{2}$$

The entropy $\mathcal{H}2$ of a graph $\mathcal{G}$ is the quotient of the square root of the summation of the a node's degree variance, and the total number of nodes in the graph. $\mathcal{H}2$ is given by the following equation:

$$\mathcal{H}2(\mathcal{G}) = \frac{1}{|V|} \sqrt{\sum_{v \in V} [deg(v) - avg\_deg(V)]^2} \tag{3}$$

B. Applying H1 to a complete graph

Let $\mathcal{G}1$ be a complete graph with 6 nodes ($|V| = 6$). Then, `nx.degree_histogram(`$\mathcal{G}1$`)` = [0, 0, 0, 0, 0, 6], $distinct(C(V)) = 1$, and finally $\mathcal{H}1(\mathcal{G}1) = (1 - 1)/6 = 0$

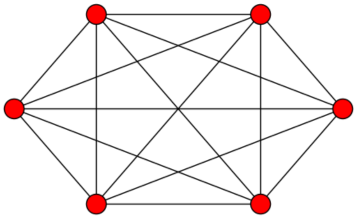

Fig. 1. A complete graph with 6 nodes. Each node is connected to every other node in the graph. $\mathcal{H}1$ classifies it as having entropy 0.

IV. Results

In this section we present results for $\mathcal{H}1$ and $\mathcal{H}2$ on random graphs with $n = 500$ and $n = 1000$. These were obtained by running Jupyter Notebooks that can be found at https://github.com/ferrolho/uoe-stn.

By analysing figures 2 to 5, we can see that both measurements $\mathcal{H}1$ and $\mathcal{H}2$ are robust, and all the plots look very similar, whether or not the number of nodes n of the graph changed.

On all the experiments it can be observed that the entropy of *Erdős-Rényi* graphs tend to zero whether p tends to 0 or 1. Furthermore, the entropy measurement $\mathcal{H}1$ seems to differ only from $\mathcal{H}2$ by a constant.

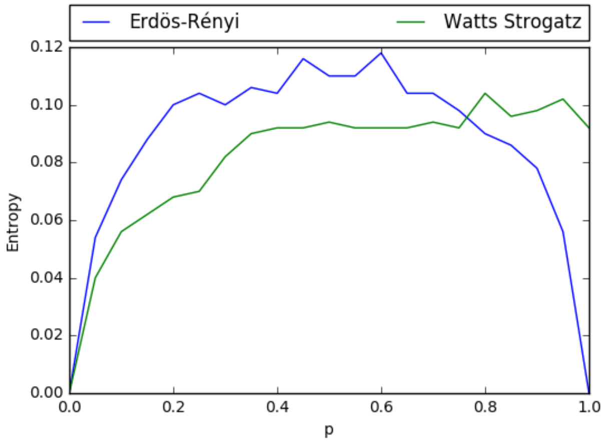

Fig. 2. $\mathcal{H}1$ results on random graphs with $n = 500$.

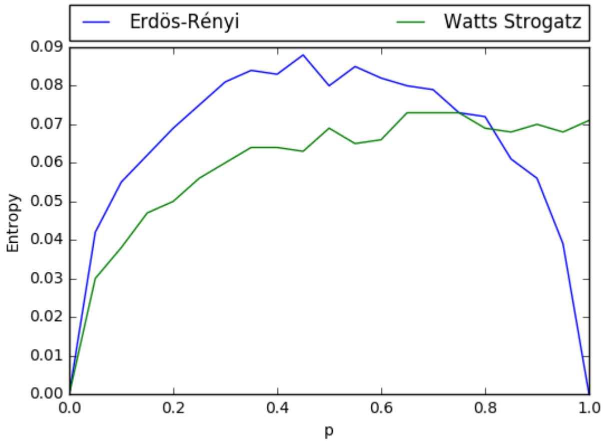

Fig. 3. $\mathcal{H}1$ results on random graphs with $n = 1000$.

Regarding *Watts Strogatz* graphs, it behaves differently from *Erdős-Rényi* graphs as the probability p increases. But it is consistent as to being 0 when p tends to 0. And just like *Erdős-Rényi* graphs, measurement $\mathcal{H}1$ and $\mathcal{H}2$ seem to differ only by a constant.

To sum up the experimental results on random graphs: we ran multiple trials for the two measurements we defined, and we varied the number of nodes of the random graphs on the test trials. The results seem to suggest that no matter what the number of nodes in the graph is, both entropy measurements stand robust and output very similar results, differing only by a product constant.

Therefore, the results seem to imply that both $\mathcal{H}1$ and $\mathcal{H}2$ are measuring the same properties of graphs, even though they have very different equations. If this is correct, and $\mathcal{H}1$ is indeed an approximation of $\mathcal{H}2$, then the first should be preferred over the latter because its computational complexity is remarkably simpler.

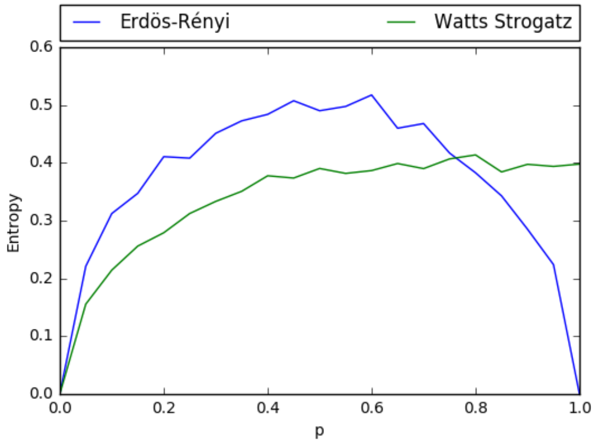

Fig. 4. $\mathcal{H}2$ results on random graphs with $n = 500$.

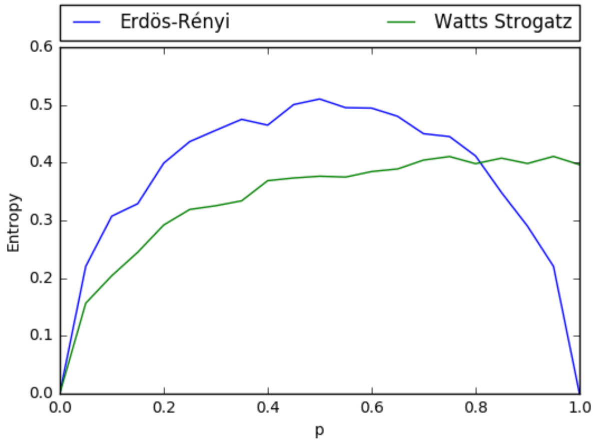

Fig. 5. $\mathcal{H}2$ results on random graphs with $n = 1000$.

A. Real-world datasets

We now turn into real-world dataset results from SNAP [4]. We used two datasets: a road network of California, and of Texas. These datasets contain intersections that are represented by nodes, and roads connecting these intersections which are represented by undirected edges.

We implemented a simple Python program to parse each dataset, load the information to NetworkX graph, and to calculate the entropy of each one using our two different measurements. This program can also be found at https://github.com/ferrolho/uoe-stn.

The following table shows the number of nodes and the number of edges on each of the road network datasets.

As we had previously suspected, the calculation of the graphs' entropy using $\mathcal{H}1$ was much faster and less costly compared to the entropy measurement $\mathcal{H}2$.

Dataset	Nodes	Edges
roadNet-CA	1965206	2766607
roadNet-TX	1379917	1921660

TABLE I
TABLE WITH THE NUMBER OF NODES AND THE NUMBER OF EDGES ON EACH OF THE ROAD NETWORK DATASETS.

Entropy 1	Entropy 2	Time 1	Time 2
2.54×10^{-6}	501.70×10^{-6}	5.75 s	57.68 s
2.90×10^-6	605.23×10^{-6}	3.42 s	40.05 s

TABLE II
TABLE WITH THE TRIALS' RESULTS: TOP LINE IS IN RESPECT TO CALIFORNIA, BOTTOM LINE IS IN RESPECT TO TEXAS.

The entropy values are also compatible with the results of the test trials using *Erdős-Rényi* and *Watts Strogatz* random graphs.

Finally, we can conclude that the Texas road network is slightly more disordered than California, since it has a greater entropy.

APPENDIX A
KOLMOGOROV COMPLEXITY

The *Kolmogorov complexity* [5] of an object is the length of the shortest program that outputs that same object.

One can think of *entropy* as an indicator of disorder, but it is also valid to think of it as the predictability of a system, or even more: the amount of information required to describe a system exactly as is.

Therefore, it might be interesting to try and come up with a measurement of the entropy of a graph with respect to the *Kolmogorov complexity*.

For example, if two distinct graphs - A and B - can be represented with the same *Kolmogorov complexity*, i.e. a minimal representation with the same length, than it is only reasonable to think of them as having the same level of entropy.

REFERENCES

[1] Matthias Dehmer. "Information processing in complex networks: Graph entropy and information functionals". In: *Applied Mathematics and Computation* 201.1 (2008), pp. 82–94.

[2] Matthias Dehmer and Abbe Mowshowitz. "A history of graph entropy measures". In: *Information Sciences* 181.1 (2011), pp. 57–78.

[3] Paul Erdős and Alfréd Rényi. "On random graphs, I". In: *Publicationes Mathematicae (Debrecen)* 6 (1959), pp. 290–297.

[4] Jure Leskovec and Andrej Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection*. http://snap.stanford.edu/data. June 2014.

[5] Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer Science & Business Media, 2009.

[6] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world'networks". In: *nature* 393.6684 (1998), pp. 440–442.

[7] Hector Zenil and Narsis Kiani. "Low Algorithmic Complexity Entropy-deceiving Graphs". In: *arXiv preprint arXiv:1608.05972* (2016).