Project Plan:
Define and compute entropy for networks

Henrique Ferrolho (s1683857), Charles Desmonty (s1685948), Alex Hoppen (s1678543)

November 22, 2016

Problem formulation

We want to define a measurement $H : \mathcal{G} \to \mathbb{R}_{\geq 0}$ that maps a graph $\mathcal{G} = (V, E)$ to its entropy, i.e. how much information is contained in the structure of the graph. In particular we want $H(\mathcal{G}) \approx 0$ if $\mathcal{G}$ is a complete graph or a graph without any edges.

Furthermore, we will answer to the following questions concerning H:

- How does $H(\mathcal{G})$ evolves in p for randomly generated parametric graphs?
- What is the value of $H(\mathcal{G})$ for well-structured graphs such that balanced trees, grid, n-hypercube, ...?
- (If we have time) Can we extend our definition of entropy to $\aleph_0$-graph (discretely infinite graph)?

Importance

A good entropy measure would allow us to automatically differentiate between graphs that contain a lot of information (often graphs created by humans) and randomly created ones (often created by computers). Analysing parts of social networks using this entropy measurement may be used to detect manipulation of social networks, e.g. the automatic creation of users on StackOverflow that automatically upvote a specific user's answers.

The Datasets

Firstly, we will use randomly generated network with the aim of studying the property of the entropy functions. We will then use a few different real-world networks to verify our entropy measurement. It may be interesting to considering an evolving network to see how entropy evolves with it.

Related work

[1] Matthias Dehmer. Information processing in complex networks: Graph entropy and information functionals. *Applied Mathematics and Computation*, 201(1):82–94, 2008.

[2] Matthias Dehmer and Abbe Mowshowitz. A history of graph entropy measures. *Information Sciences*, 181(1):57–78, 2011.

[3] Edward Kenley and Young-Rae Cho. Entropy-based graph clustering: application to biological and social networks. *IEEE International Conference on Data Mining (ICDM 2011)*, December 11-14 2011.

Preliminary Ideas

One idea aims to detect if some edges in the network are more *important* than others by increasing the score of each edge for every use in a shortest path between two nodes. The entropy of the edge's scores has proven to be a rough estimate of the graph's entropy.

We will also study the standard deviation of the degrees of every nodes with a local definition of the mean degree, relative to each nodes. This methods gives good results but seems to *local* and is not very precise.

Another method that we'll study is the *overlap matrix* of the graph which is $\mathcal{O} = (\text{Neighborhood overlap of } ij)_{i,j}$. This aims to capture the information about communities in the graph and the relation between them.

Schedule

01/11: Read existing papers, 14/11: Develop a measurement for entropy, 21/11: Finalize our report