This document contains reasoning and ideas about the final report: ideas of structure, and ideas of methods to try.

Use probability theory instead of simulations to compute z-score of LCC

04/25

The report in `Graph LCC size/` shows how to determine analytically the probability distribution of the largest component size of a random graph having n vertices and m edges, i.e. $\mathbb{P}\left[\left|\mathrm{LCC}(\Gamma)\right| = k\right]$ for $k \in \mathbb{N}$.

For a graph $\Gamma = \Gamma(V, E)$, let's define its *density* by:

$$d(\Gamma) := \frac{|E|}{X(|V|)}.$$

If $I = I(G, E_I)$ is the interactome, and D is a genetic disease, let's define G_D the set of all disease genes related to D (found with OMIM and GWAS databases). Authors of the article look for $\left|\mathrm{LCC}(\Delta_{G_D}(\Gamma))\right|$, the size of the largest connected component of the disease subgraph, and compare it with the average largest component size of random subgraphs having $|G_D|$ vertices (genes) in the interactome.

The goal is to use analytical probabilities to be more accurate. For this, we use a random variable $\mathcal{G}(G_D, \cdot)$, a graph chosen randomly in the set of all graphs having G_D as vertices set. Furthermore, we need to simulate the density of interactome d_I. Thus, for the rounding function:

$$r : \mathbb{R} \to \mathbb{Z} : x \mapsto \left\lfloor \frac{2x+1}{2} \right\rfloor,$$

we want the random subgraph to have $r\left(d_I \cdot X(|G_D|)\right)$ edges. We can then deduce the average LCC size, taking the density into account.

Let's thus plot figure S4b from supplementary materials but changing the way of computing the z-score.