Reproducing and updating results from *Uncovering disease-disease relationships through the incomplete interactome*

Robin Petit[1] and Tom Lenaerts[1]

[1]Université Libre de Bruxelles

Abstract

This paper intends to reproduce the results exposed of the article *Uncovering disease-disease relationships through the incomplete interactome* (Menche et al., 2015) and check the robustness of the procedure by confronting the results with the results of updated datasets, which intends to be a systematic analysis, still stands for a more recent version of the interactome, and if the results exposed have shifted towards a more or less significant level. We find that besides being reproducible, some of the results exposed turn out to be more significant, and others turn out to be less significant.

1 Introduction

An interactome is a graph containing all the molecular interactions within a cell. This notion appeared in 1999 (for the drosophila) (Sanchez et al., 1999) and the need to thoroughly study this structure has been expressed less than 15 years ago (Barabasi and Oltvai, 2004).

The interactome is one of the many biological networks, covering gene networks also called genome, protein networks also called proteome (Rolland et al., 2014), disease networks (Goh et al., 2007) among others, and providing the ability to deeply study biology (Barabasi and Oltvai, 2004) as well as medicine (Barabási et al., 2011) or more recently pharmacology (Hopkins, 2008).

As the need to use the interactome as a tool to analyze diseases genetic behaviour had already been expressed (Vidal et al., 2011), the authors or the original paper (Menche et al., 2015) showed that the human interactome has now reached sufficient completion to systematically study diseases. Yet, the interactome provides the availability to study many more genetic relations, such as drug-disease correlation (Yu et al., 2016), or digenic diseases (Gazzo et al., 2015).

In the original paper, authors applied disease genes association databases (in particular OMIM, the Online Mandelian Inheritance in Man (Amberger et al., 2008), and GWAS, the Genome-Wide Association Studies, compiled by PhenGenI (Ramos et al., 2014)) on the human interactome in order to determine the properties of their distribution in the graph.

The diseases studied are selected such that they possess at least 20 genes associated to them, which yields 29,775 disease-gene associations on 3,173 distinct genes, with 2,436 contained in the interactome.

Major results were that: firstly diseases tend to *cluster* in denser subgraphs than the interactome itself (shown by bigger largest connected component than expected in random interactome subgraphs), secondly that phenotypically close diseases tend to overlap on a significant amount of genes.

Authors of the original paper also remind that the interactome is incomplete and estimated around 20%-complete for the interactions by the estimation that the interactome contains roughly 6.5×10^5 interactions, and around 54%-complete for the proteins involved by the estimations that the interactome contains roughly 2.5×10^4 nodes (Amaral, 2008; Stumpf et al., 2008)

The first part of this paper focuses on the reproduction of exposed results in the original paper, namely the disease modules propensity to cluster into highly connected components and the significantly lower separation indicator values for highly gene related disease pairs.

The interactome used in the original paper contains 13,460 genes and 141,296 physical links constructed on several interactions databases including BioGRID (Chatr-aryamontri et al., 2017), IntAct (Kerrien et al., 2011), TRANSFAC (Matys et al., 2003), MINT (Licata et al., 2011), HPRD (Keshava Prasad et al., 2008), KEGG and BIGG (Lee et al., 2008), CORUM (Ruepp et al., 2009) and PhosphitePlus (Hornbeck et al., 2011). Authors chose to rely only on physical protein-protein interactions (PPI) and to exclude functional interactions (Caldera et al., 2017). This interactome is available to download as supplementary material.

2 Results

2.1 Reproducibility

Clustering of disease modules The original paper discusses the trend of diseases to cluster into dense subgraphs. In order to measure this hypothesis, the largest connected component (LCC) of the diseases in the interactome is used.

When analyzing this, we observe that 241 out of the 299 diseases (more than 80%) have a significantly bigger LCC than expected by chance.

The z-score of the LCC size of a disease is strongly related to its relative module size, which is defined by $r = S/N_d$, with S the size of the LCC of the disease, and N_d, the number of genes associated with the given disease (Figure 1).

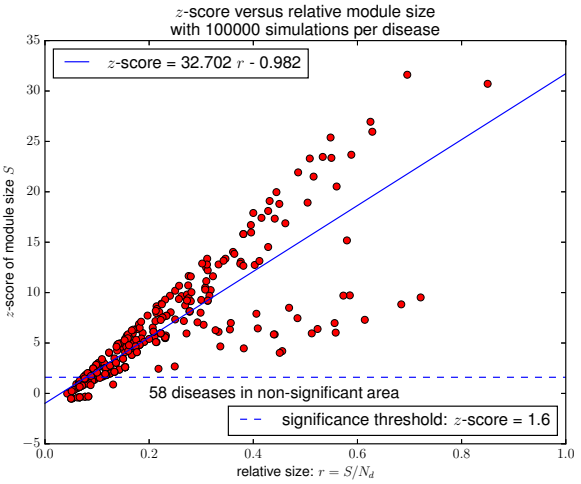

Figure 1: z-score of largest connected component size vs relative module size. 100,000 simulations have been performed per disease in order to determine the size of the largest connected component of the disease subgraph expected by chance. Diseases being highly connected (thus highly covered by the interactome) present a higher z-score and then a higher confidence about the significance of the size difference. Assuming that the distribution of largest connected component is normal for random samples (Barraez et al., 2000), each z-score can be associated to a p-value, in particular, z-score $\geq$ 1.6 corresponds to p-value $\leq$ 0.05, representing significance threshold.

This allows to observe that several diseases do not present a significantly larger largest connected component than expected by chance, but that these diseases have a thin relative size (< 0.2, so that less than 20% of their related genes are connected in the current interactome). Also, diseases with bigger relative size have a higher z-score, leading to think that a more complete interactome, with higher coverage of the diseases would increase the significance of the result.

Separation distribution The original paper describes the separation of diseases in the interactome through two different measures: the overlapping score (C-score and J-score defined respectively as $|A \cap B| / \min(|A|, |B|)$ and $|A \cap B| / |A \cup B|$ for A and B two disease genes sets) and the separation score s_{AB} defined as follows. For two diseases A and B, let $\langle d_A \rangle$ and $\langle d_B \rangle$ be the mean distance between two proteins in the disease subgraphs of A and B respectively, and let $\langle d_{AB} \rangle$ be the mean distance between two proteins of each disease subgraph. Then define s_{AB} as:

$$s_{AB} = \langle d_{AB} \rangle - \frac{\langle d_A \rangle + \langle d_B \rangle}{2}. \quad (1)$$

$J = 0$ $C = 0$	$0 < J < 1$ $0 < C < 1$	$J < 1$ $C = 1$	$J = 1$ $C = 1$
No common gene	Partial overlap	A is complete subset of B	A and B are identical

Table 1: Meaning of the different J-score/C-score combinations.

Table 1 details the meaning of J/C-scores combinations.

When analyzing the separation of the interactome, we find out the same results than those presented in the original article. The only difference being that for non-overlapping disease pairs, the amount of pairs having a separation value below 0 is 710 versus 717 in the original paper, being totally non-significant since 7 pairs represent less than 0.03% of all the non-overlapping pairs set (Figure 3).

2.2 Robustness

Databases update In order to update the interactome, the tool *inter-tool* (Catabia et al., 2017) has been used. Latest datasets from BioGRID, IntACT, and the Database of Interacting Proteins (DIP) (Salwinski et al., 2004) (on July 25th) and of MINT (on July 27th) have been downloaded and merged via inter-tools.

This newer version merged with the original interactome yields a new graph having 17,786 nodes and 370,326 edges (so a bit more than 1.3 times the initial amount of nodes, and more than 2.6 times the initial amount of edges). From the 4326 genes added in this newer version, 361 are associated with one of the 299 diseases, leading to 2,797 disease-associated genes in the interactome.

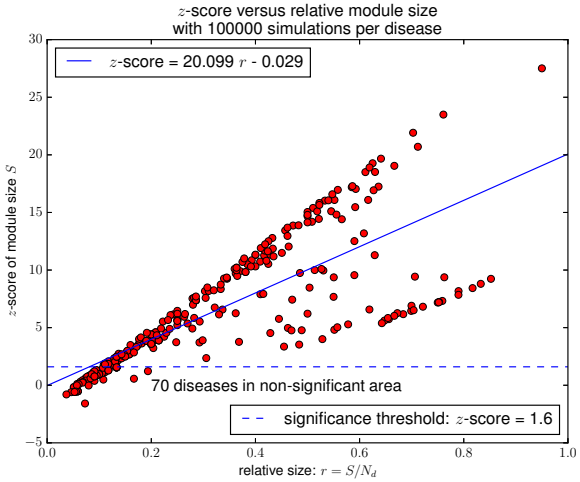

Figure 2: z-score of the largest connected component size vs relative module size of the newer interactome. Adaptation of Figure 1 on the newer version of the interactome.

This newer version is around 71%-complete for the proteins involved and 57%-complete for the interactions

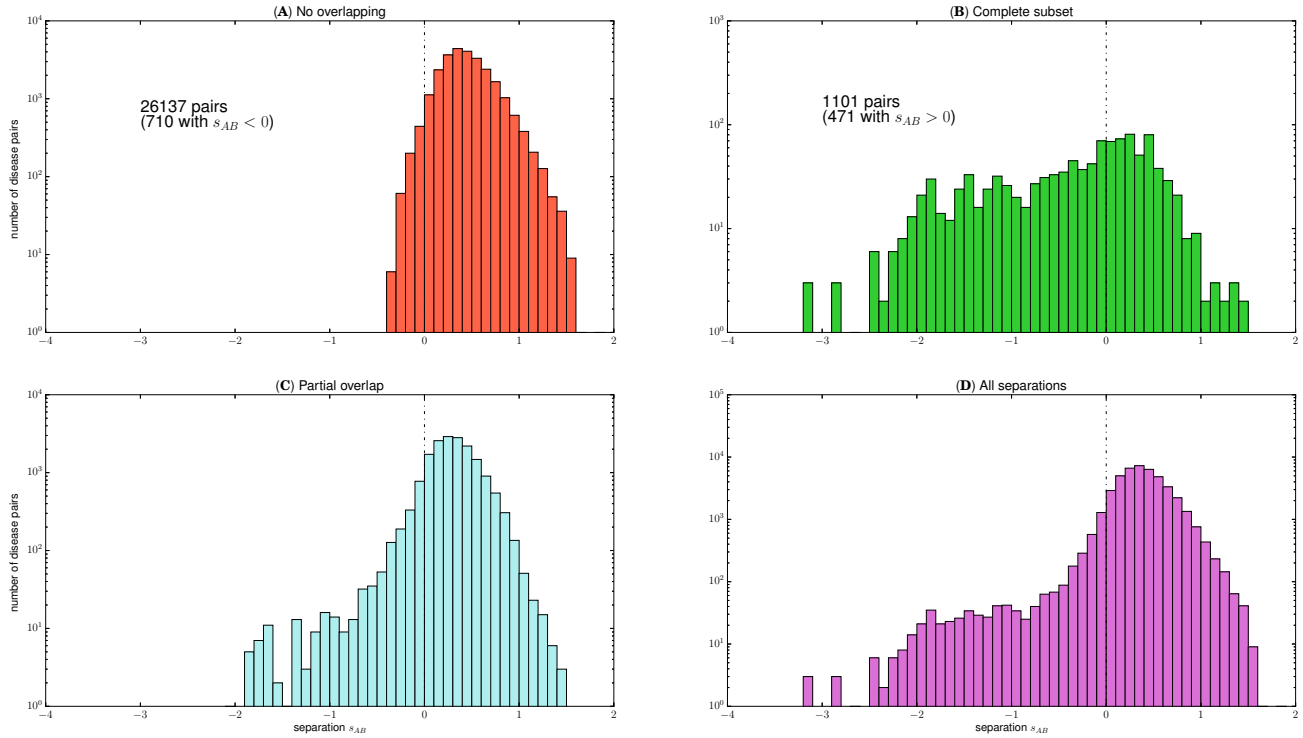

Figure 3: Disease pairs separation. (**A**) The s_{AB} distribution of disease pairs with no common gene (C-score = J-score $= 0$). We observe that even though no gene is shared, 710 of the 26,137 pairs (less than 3%) have a negative separation score (between 0 and 0.5). (**B**) The s_{AB} distribution of disease pairs with complete overlap (J-score $< C$-score $= 1$). We observe that despite the inclusion of one disease genes set in the other, 471 of the 1,101 pairs (more than 42%) have a positive separation score. Yet, separation goes to very low values (below 2). (**C**) The s_{AB} distribution of disease pairs partially overlapping ($0 < J$-score $\leq J$-score < 1). These disease pairs show the same spike of frequency right to $s_{AB} = 0$ as in (A) and the same tail of frequency left to $s_{AB} = 0$. (**D**) The s_{AB} distribution of all the disease pairs.

(Amaral, 2008; Stumpf et al., 2008).

Disease genes considered in this update are the same 299 diseases studied in the original paper. An update of these is yet to be done in further investigations.

2.3 Discussion

By defining the density of a graph $\Gamma = (V, E)$ as being $d(\Gamma) := |E| / \binom{|V|}{2}$, we find that the newer interactome is more than 1.5 times denser than the original one (0.156% versus 0.234% for the newer one).

When performing the same localization analysis of the disease modules in the interactome on the newer version, we observe that 12 more diseases have a z-score below the threshold of 1.6, which makes results less significant (Figure 2).

Despite the decrease in z-scores, we observe that general relative size has shifted towards right (Figure 4). This is explained by the increase in the gene number, leading to a higher coverage of the disease genes, as well as the increase in density leading to a more connected graph, having therefore more connected subgraphs. Although, z-score maximum has dropped from 31.6 to 27.5, the mean z-score has increased from 6.2 to 6.4 due in part to the density increase in the region z-score $\in [10, 20]$.

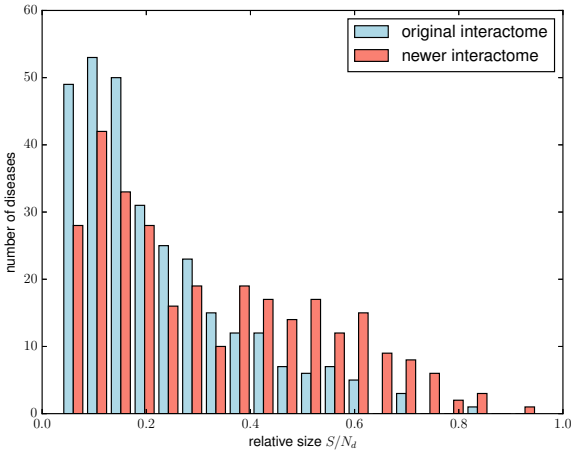

Figure 4: Comparison of relative size distribution between original and newer interactomes. We observe that the number of diseases having a relative size below $\simeq 0.35$ has lowered, whereas the number of diseases having a relative size above $\simeq 0.35$ has increased. This is explained by the bigger density of the newer interactome, leading to larger LCC in the disease subgraphs.

As well, average relative size for the given diseases has increased from 22% to 32% in the newer interactome, showing

the better coverage allowed by the more recent version.

The interpretation of Subsection 2.1 still stand: diseases having a low z-score still have a low relative size, and diseases with a higher z-score have a higher relative size. So either the interactome is still too incomplete for these diseases, or they lie in very sparse regions of the interactome.

Due to the graph density increase, the degree distribution has changed as well (Figure 5). Both the original interactome and the new one are highly alike and comparable on their degree distribution. The most visible change is that the newer interactome contains more highly connected nodes which implies a mean twice as big.

Yet, the network is still scale-free, and its degree distribution still follows a power law, which is inherent to biological networks (coefficient 1.6 for the newer one versus 1.53 for the original one).[1] γ is bigger than 1 and smaller than 3, which is considered relevant (Barabasi and Oltvai, 2004; Vidal et al., 2011).

The 12 diseases which have a decrease of z-score below 1.6 is due to the increase of the interactome density implying that a subgraph taken at random tends to have a wider LCC at equal size.

When applying the separation analysis on the newer interactome, we observe that non-overlapping diseases have a higher separation score in the newer interactome: from 710 disease pairs to 324 pairs having a $s_{AB} > 0$ score, i.e. so 54% of the non-overlapping disease pairs increased their separation score above 0 (Figure 6).

We also observe that 6% of the complete subset disease pairs decreased their separation score below 0, which is way less significant.

More generally, separation scores have tightened around 0: s_{AB} is in $[-3.2, 1.6]$ in the original interactome, and in the newer one, s_{AB} is in $[-2.5, 1.1]$.

[1] Determined by taking coefficient γ in the relation $\log(P(k)) \sim -\gamma \log(k)$ found by a linear regression.

3 Extension

3.1 Subgraph largest connected component distribution

Figures 1 and 2 require a null hypothesis in order to define a z-score, being the random one. Those are computed as follows: if S_D is the disease module associated with a given disease D, then its z-score is given by:

$$z\text{-score} = \frac{|S_D| - \mu(S^{\text{rand}})}{\sigma(S^{\text{rand}})}, \tag{2}$$

with $\mu(S^{\text{rand}})$ and $\sigma(S^{\text{rand}})$ being respectively the mean and the standard deviation of the largest connected component size of a random subgraph of size $|D|$ in the interactome.

These values are obtained by simulations: taking subgraphs at random of given size in the interactome yields a

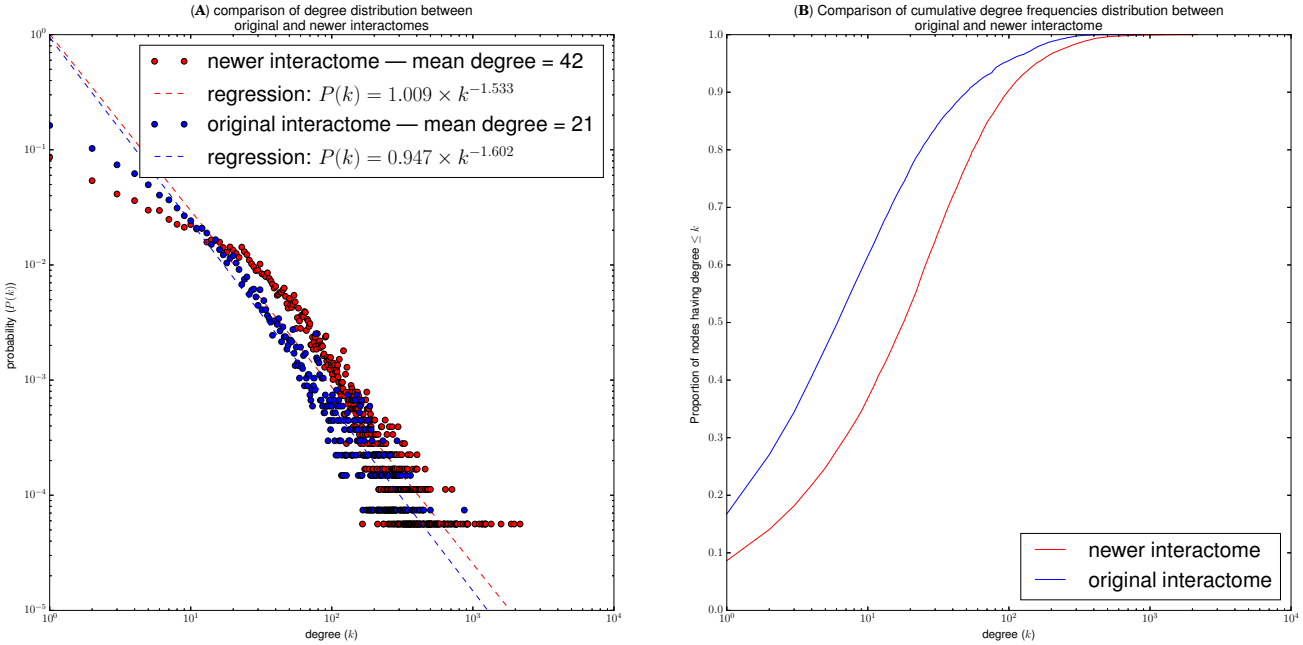

Figure 5: Degree distribution comparison. (**A**) The degree of a node A in a graph is the number of nodes B adjacent to A. Both the original interactome and the newer one are scale-free, i.e. their degree distribution follows a power law. This means that for a number k, $P(k)$, the probability that a node in the graph is of degree k is given by $\alpha k^{-\gamma}$ with γ the distribution parameter, and $\alpha = \zeta(\gamma)^{-1}$ the regularization parameter. A power law is characterized by a few nodes being highly connected to the other ones (right part of the graph, with big values of k), whereas most nodes are connected to only a few other ones (left part of the graph, with small values of k). The newer interactome has a smaller γ coefficient, meaning that a bigger proportion of nodes have a high degree, compared to the original one. (**B**) Cumulative degree distribution of both the original and the newer interactome. We observe easily the power law characteristic that even though degree can reach 2,000, almost all the interactome nodes have a degree ≤ 100.

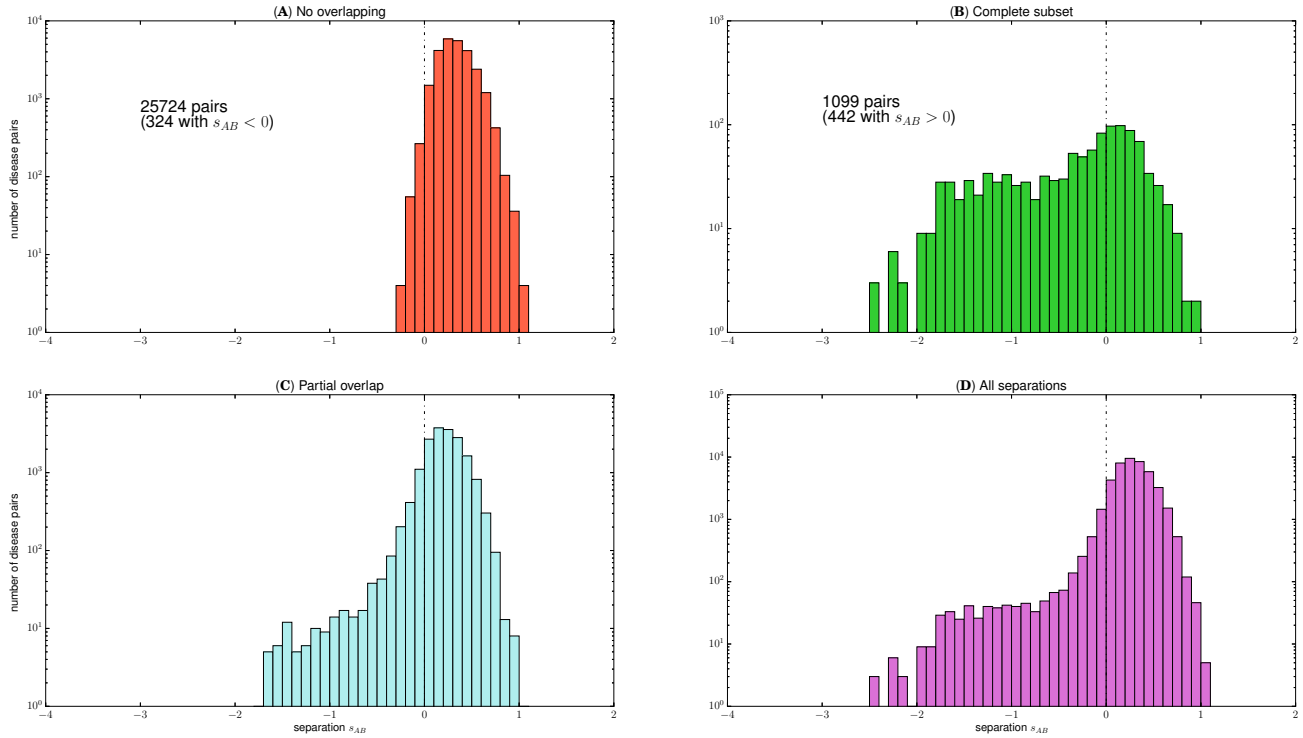

Figure 6: Disease pairs separation in the new interactome. Adaptation of Figure 3 on the newer interactome. (**A**) We observe that more than half of the disease pairs sharing no genes that had a negative s_{AB} score in the original interactome now have a positive score, due to a decrease in $\langle d_A \rangle$ and $\langle d_B \rangle$ because of the higher density of the newer interactome. (**B**) We also observe that 29 disease pairs related by inclusion had a score right shift towards positive values.

distribution $P(S^{\text{rand}})$ with a given mean $\mu(S^{\text{rand}})$ and standard deviation $\sigma(S^{\text{rand}})$ (Figure 7).

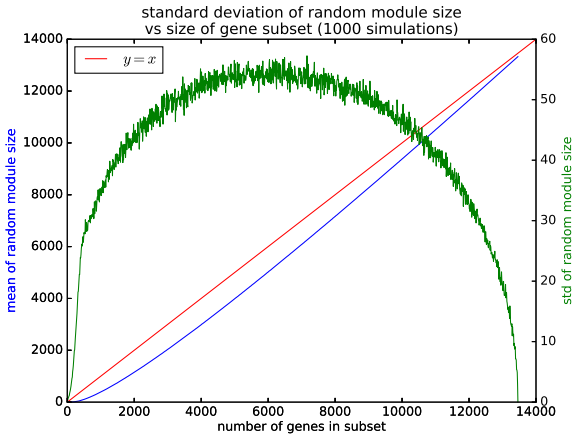

Figure 7: S^{rand} mean and standard deviation distribution. With 10^3 simulations per subgraph size, we obtain the distribution of the largest connected component size in the interactome. We observe that for subsets of small size k, the expected LCC size is significantly smaller than k whereas for subsets of big size K (giant components), the expected LCC size is much closer to K.

3.2 Analytically determined probability density

In order to avoid simulation computation time, a probability mass function has been determined. For a graph $\Gamma = (V, E)$ such that $|E| = m$ and $\Lambda_k^m(V, \cdot)$, the set of all graphs having V as vertex set, m edges, and a LCC of size k, we define p_k, the probability that Γ has LCC of size k as:

$$p_k = \left|\Lambda_k^m(V, \cdot)\right| / \binom{\binom{|V|}{2}}{m}, \tag{3}$$

which requires $\left|\Lambda_k^m(V, \cdot)\right|$ to be computed. Yet, this set cardinality is defined by a recurrence relation (see proof in supplementary materials), which makes computations several orders of magnitude slower (even with dynamic programming and caching).

A Python3 implementation is given in `source/lcc_size/`.

4 Conclusion

5 Materials and Methods

Inter-build (one of the programs from Inter-tools) requires datasets in the PSI-MITAB format, an extension of the PSI-MI format (Kerrien et al., 2007), and outputs a tsv file.

The outputted file by Inter-build and the interactome provided with the original paper both use Entrez gene IDs, they

can therefore easily be merged. In order to merge these, the script `merger.py` has been written. As these two tsv files have a different format (different columns), only the gene IDs are outputted by `merger.py`.

References

Amaral, L. A. N. (2008). A truer measure of our ignorance. *Proceedings of the National Academy of Sciences*, 105(19):6795–6796.

Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2008). Mckusick's online mendelian inheritance in man (omim®). *Nucleic acids research*, 37(suppl_1):D793–D796.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12(1):56.

Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5(2):101.

Barraez, D., Boucheron, S., and Fernandez De LaVega, W. (2000). On the fluctuations of the giant component. *Comb. Probab. Comput.*, 9(4):287–304.

Caldera, M., Buphamalai, P., Müller, F., and Menche, J. (2017). Interactome-based approaches to human disease. *Current Opinion in Systems Biology*.

Catabia, H., Smith, C., and Ordovás, J. (2017). Inter-tools: a toolkit for interactome research.

Chatr-aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., et al. (2017). The biogrid interaction database: 2017 update. *Nucleic acids research*, 45(D1):D369–D379.

Gazzo, A. M., Daneels, D., Cilia, E., Bonduelle, M., Abramowicz, M., Van Dooren, S., Smits, G., and Lenaerts, T. (2015). Dida: A curated and annotated digenic diseases database. *Nucleic acids research*, 44(D1):D900–D907.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–690.

Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2011). Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research*, 40(D1):D261–D270.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., et al. (2011). The intact molecular interaction database in 2012. *Nucleic acids research*, 40(D1):D841–D846.

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod, N., Bader, G. D., Xenarios, I., Wojcik, J., Sherman, D., et al. (2007). Broadening the horizon–level 2.5 of the hupo-psi format for molecular interactions. *BMC biology*, 5(1):44.

Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2008). Human protein reference database—2009 update. *Nucleic acids research*, 37(suppl_1):D767–D772.

Lee, D.-S., Park, J., Kay, K., Christakis, N., Oltvai, Z., and Barabási, A.-L. (2008). The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences*, 105(29):9880–9885.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., et al. (2011). Mint, the molecular interaction database: 2012 update. *Nucleic acids research*, 40(D1):D857–D861.

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., et al. (2003). Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, 31(1):374–378.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224).

Ramos, E. M., Hoffman, D., Junkins, H. A., Maglott, D., Phan, L., Sherry, S. T., Feolo, M., and Hindorff, L. A. (2014). Phenotype–genotype integrator (phegeni): synthesizing genome-wide association study (gwas) data with existing genomic resources. *European Journal of Human Genetics*, 22(1):144.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2009). Corum: the comprehensive resource of mammalian protein complexes—2009. *Nucleic acids research*, 38(suppl_1):D497–D501.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic acids research*, 32(suppl_1):D449–D451.

Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Röder, L., Euzenat, J., Rechenmann, F., and Jacq, B. (1999). Grasping at molecular interactions and genetic networks in drosophila melanogaster using flynets, an internet database. *Nucleic acids research*, 27(1):89–94.

Stumpf, M. P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964.

Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome networks and human disease. *Cell*, 144(6):986–998.

Yu, L., Wang, B., Ma, X., and Gao, L. (2016). The extraction of drug-disease correlations based on module distance in incomplete human interactome. *BMC systems biology*, 10(4):111.