Protein Interaction Networks, Essentiality and Sensitivty

Florian Knorn* and Oliver Mason†

* Otto-von-Guericke-Universität Magdeburg, Germany
† Hamilton Institute, NUI Maynooth, Ireland

1. Motivation

- Protein-Protein-Interactions (PPI) describe key biological processes
- New measurement techniques
  - ➠ unprecedented amounts of data
- Data representation by **graphs (networks)**:
  - **Nodes**: Proteins
  - **Edges**: Interactions between proteins

Key question: "Which proteins are important / essential?"

➠ Use **ranking algorithms** and **centrality measures** to identify important nodes.

Problem: Real data usually noisy / inaccurate / incomplete !

In this poster, based on the work in [2], we evaluate

- the **ability** of different ranking algorithms to **identify** essential proteins
- the **impact** of inaccurate data on these algorithms using
  - simulated networks (Barabási-Albert random graphs)
  - real data (PPI network of *Saccharomyces Cerevisiae*, data set by Uetz *et al.*)

2. Scale-free random graphs

- Seminal graph model by A.-L. Barabási and R. Albert, [3], that is based on **growth**
- Generated by an **iterative process**:
  - Start with small, connected graph
  - At each step: add in **one** node and **m** edges using **preferential attachement**
- Characteristics:
  - **Many** nodes **with few** connections, but a **few highly** connected "hubs"
  - **Scale free**: No characteristic node degree (exponential degree distribution)
  - **Ultra small world**: very short average path lengths
- Used to model many **real world networks**: WWW, phone networks, PPIN, … [4]

3. Ranking schemes and essentiality

Idea: identify the "important" nodes by establishing some form of **node ranking**, for instance by attributing some sort of **"score"** to each node and then sorting by it.

➠ Notion of "importance" depends on the **interpretation** and also the **application**

- × Node degrees (ND)
- □ PageRank (PR)
- ○ HITS (HI)
- ◇ Status (ST)
- △ Excentricity (EX)
- • Centroid value (CV)
- ✳ Damage (DA)

Question: Taking the top 1%, 5%, 10% and 25% of proteins from the **top** of the rankings, what is the fraction of **actually essential** proteins in that set (using truth data)?

Top % \ Scheme	ND	HI	PR	EX	ST	CV	DA
1 %	**83.3**	50.0	**83.3**	39.1	33.3	33.3	66.7
5 %	**48.4**	35.7	46.4	24.6	~~17.2~~	~~17.2~~	44.0
10 %	34.4	28.6	**39.3**	24.6	26.8	26.8	33.9
25 %	31.5	27.2	**32.6**	24.8	26.8	26.8	30.6

***Tab. 1**: Comparison of the algorithm's abilities to identify essential nodes in the Uetz et al. data set when considering the top 1%, 5%, 10% or 25% of the rankings as "essential". **Bold**: best value in row; ~~cancelled~~: value even below overall fraction of essential proteins*

4. Network perturbations

Among others, we evaluated these types of perturbations on the network:

- Edge removal
- Edge rewiring
- Edge addition
- Node removal

5. Deviation measures

Several notions of deviation have been explored. Here, we will display the following:

① How **important really are** highly ranked from the perturbed network
② The chance of **seemingly** important nodes to be, in fact, **not important**
③ The chance of important nodes **not to be identified** as such

- The first measure is calculated using the actual **ranks** of the top 5% of nodes
- The other two only compare intersections of the **sets** of the top 5% of nodes from both rankings (that is of the perturbed and unperturbed graphs)

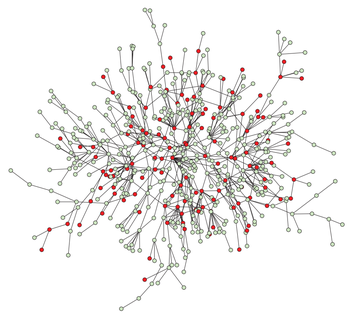

***Fig. 1**: PPI Network of S. Cerevisiae. Data set by Uetz et al., [1]*

6. Results from simulated graphs

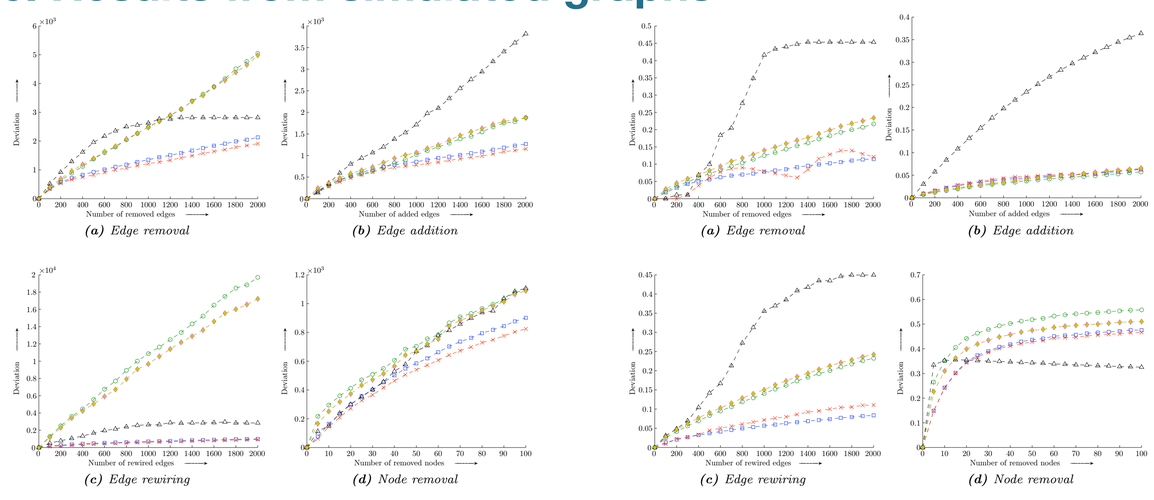

***Fig. 2**: Scale-free graph, deviation measure ①*

***Fig. 3**: Scale-free graph, deviation measure ②*

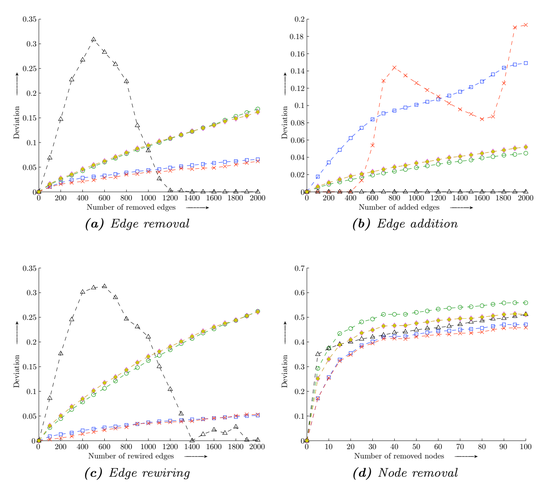

***Fig. 4:** Scale-free graph, deviation measure ③*

- Results from 250 repetitions, using scale-free random graphs with 3000 nodes and ≈9000 edges
- Odd behaviours of EX (△) result from disconnection of the graphs
- ST (◇) and CV (•) are almost always identical (this is no surprise, it follows from their definition)
- PR (□) and ND (×) seem to be generally **most robust**
- However: in Fig. 4 (b) **HITS** and the centrality based measures perform clearly better

7. Results from real data

- Used the largest connected component of the Uetz *et al.* data set. It contains 558 proteins and 646 interactions; **22.6% of proteins are known to be essential**
- Introduced **increasing** amounts of perturbation; results are averages from 50 runs
- Took top 5% off each ranking and calculated **fraction of essential proteins**

Perturb. % \ Scheme	ND	HI	PR	EX	ST	CV	DA
0 %	**48.4**	35.7	46.4	24.6	~~17.2~~	~~17.2~~	44.0
5 %	**48.6**	35.2	44.6	~~20.6~~	~~19.1~~	~~19.1~~	41.4
10 %	**45.3**	35.9	44.6	~~18.8~~	~~18.9~~	~~18.9~~	42.1
15 %	**46.3**	36.8	44.4	~~19.1~~	~~19.1~~	~~19.1~~	41.4
20 %	**47.1**	37.2	45.2	~~16.8~~	~~16.8~~	~~16.8~~	40.6

***Tab. 2**: Perturbation: Edge removal. Comparison of the fraction of correctly identified proteins with increasing amounts of perturbation.*

Perturb. % \ Scheme	ND	HI	PR	EX	ST	CV	DA
0 %	**48.4**	35.7	46.4	24.6	~~17.2~~	~~17.2~~	44.0
5 %	**48.3**	37.2	45.6	26.3	23.2	23.2	37.0
10 %	**47.9**	38.9	46.7	22.7	~~21.7~~	~~21.7~~	37.0
15 %	**47.1**	40.1	46.6	~~20.4~~	~~20.4~~	~~20.4~~	37.2
20 %	46.3	39.7	**47.0**	~~19.6~~	~~19.4~~	~~19.4~~	36.2

***Tab. 3**: Perturbation: Edge rewiring.*

Perturb. % \ Scheme	ND	HI	PR	EX	ST	CV	DA
0 %	**48.4**	35.7	46.4	24.6	~~17.2~~	~~17.2~~	44.0
5 %	**47.6**	37.1	45.4	23.6	~~19.7~~	~~19.7~~	37.6
10 %	**46.5**	38.9	45.2	29.8	26.1	26.1	35.7
15 %	**46.1**	40.2	45.1	27.9	28.3	28.3	35.0
20 %	45.7	41.9	**46.4**	27.4	29.7	29.7	34.5

***Tab. 4**: Perturbation: Edge addition.*

- **ND most successful**, consistent *and* fairly robust in detecting essential proteins
- **PR** second best for detection of essentiality, but **most robust**
- HITS second most robust; centrality based measures rather sensible and only give low detection rates, often performing worse than would purely random picks!

8. Future directions

- Evaluate **more structured** perturbations, or **combinations** of perturbations
- Find more **theoretical results** on the robustness of the different algorithms
- Investigate damage on **larger data sets**, as it showed some promising results

More extensive results and analysis on other data sets can be found in [2].

[1] Peter Uetz and Loic Giot. A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature*, 403:623–627, February 2000.
[2] Florian Knorn. Ranking and importance in complex networks. *Studienarbeit*, October 2005.
[3] Albert-László Barabási and Réka Z. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.
[4] Réka Z. Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–96, January 2002.

Work funded by *Science Foundation Ireland* grants 04/IN1/I478 and 03/RP1/I382 as well as the *German National Academic Foundation*.