Studying the effect of lethality on protein-protein interactions in yeast

Michael DARMOUTOMO, Guillaume FRANCHI and Bassel MASRI

1/1/2021

Contents

Introduction

Protein–protein interactions (PPI) are important to understand, since they give us an insight in how protein interactions affect larger organisms or cells. Modeling these interactions as a network gives us the opportunity to apply statistical graph models and discover relationships in a statistical manner. We can take advantage of graphical features from the network, such as number of edges, structure, centrality, etc. In this paper we analyse how the deletion of a protein affects the lethality of a cell. We focus on the yeast Saccharomyces cerevisiae. This paper is an expansion of earlier work by Jeong et al. (2001), but expands this literature by applying statistical models to find the relationship between lethality and its interactions. This paper is split up into two parts of analysis, which firstly consists of a descriptive analysis of the network's structure, and secondly it uses an exponential random graph model (henceforth, ERGM) to model the correlation of ties being formed between vertices and attributes such as lethality. These two parts allow us to answer the research question whether there is a statistically significant correlation between the formation of links between proteins and lethality?

This paper continues as follows: Section 2 elaborates the methodology used in this paper, Section 3 consists of the Exploratory Data Analysis, Section 4 contains the Statistical Graph Modeling, and Section 5 contains the Conclusion.

Methodology

This section discusses the methodology used in this paper. It goes through the data that we use, our analysis methodology and preprocessing steps.

Data

The S. cerevisiae PPI network was created by Jeong et al. (2001) and contains a network of 2114 proteins with 2203 interactions. The network contains data of the yeast proteins, the interactions between proteins and whether the deletion of a protein results in a death of the yeast (lethality). The data was directly downloaded using the package `networkdata` in R.

Before analyzing the data, we perform some transformations to the network. The attribute lethality contains the following classes: lethal, non-lethal, unknown and 69 proteins with their own name. Since our main interest for this research is just to answer whether we can predict the lethality, we group this last category into *Other*.

Analysis steps

Investigating protein–protein interactions network relate to examine the correlation between forming links and classifying the lethality of a protein. While typical machine learning algorithms such as random forests, neural networks, ensemble classifiers, and Naive Bayes classifiers are often proposed for classification problems and, therefore, detecting lethality of protein when analyzing PPIs datasets, network modeling techniques can offer a significant insight on the interaction between said proteins. Therefore, the steps for analyzing the PPI network start with a typical exploratory data analysis to capture as much information as we can about the network. We then implement numerical simulations to capture the structure of the network in the aim of quantifying its properties (e.g. random and small world properties) and finally, we fit different models on the network to predict the formation of links based on the network's attributes which would help answer our main research question -is there a statistically significant correlation between the formation of links between proteins and lethality?

Exponential Random Graph Model (ERGM)

The ERGM is a statistical model that allows us to analyse a network. We model the dependent variable Y as a member of the exponential family. Mathematically, we have the following[1]:

$$P(Y=y|\theta) = \Big(\frac{1}{\kappa(\theta)}\Big)\exp\Big\{\sum_H \theta_H g_H(y)\Big\},$$

where each H is a configuration, or in other words, a set of possible edges in G, $g_H(y)$ is the network statistic corresponding to configuration H, $g_H(y) = \Pi_{y_{ij\in H}y_{ij}=I_{Hoccursiny}}$, and finally $\kappa(\theta)$ is a normalization constant.

Exploratory Data Analysis

Visual exploration of the network

As with all statistical modeling, data exploration is advised prior to network model development. In the case of network data, visualization and descriptive statistics can give some insight into the structure of a network that can be helpful during the development and evaluation of a statistical model. The first thing we will do is a visualization of the network with node color showing the lethality attribute which can aid in identifying patterns of ties among proteins with different characteristics.

Table **??** shows the frequency distribution of the different levels of lethality in the network.

Table 1: Frequency table of the attribute lethality

Level	Count
Lethal	401
Non-Lethal	1335
Other	69
Unknown	309

Preliminary visual analysis of the network graphic through 1 and **??** shows that most proteins in the network fall under the category of non-lethal and many of them are *not* clustered towards the center (i.e. on the outskirts of the network with no edges) with no apparent edges (or links). More clustering formations are noticeable towards the center of the graph but we cannot easily distinguish the groups that are clustered together. We can also note that some lethal proteins, code colored red, have significant interactions with other proteins.

Having a large number of nodes in the network can sometimes obscure important patterns in a network graph. Displaying the largest component (i.e. largest connected group of nodes) in a network can aid in clarifying patterns visually. The largest component which contains most of the nodes in the network can be isolated and graphed.

Vertex size is another ways to visually discern patterns in our network structure. Therefore, the last visual exploratory analysis that we will perform is a plot where the network is colored according to lethality of the proteins and their corresponding vertices would be sized according to a specific measure like the degree. Degree is the number of links a network member has. In this case, degree would represent the number of ties for each pair of proteins in the protein network.

Figure 2 below helps distinguish some more apparent clustering among the proteins by plotting only the main components and sizing the vertices by their degrees. Indeed, we can clearly discern the biggest nodes

[1] Notation from the slides of Eftychia Solea (UE-MSD01), Topic 4

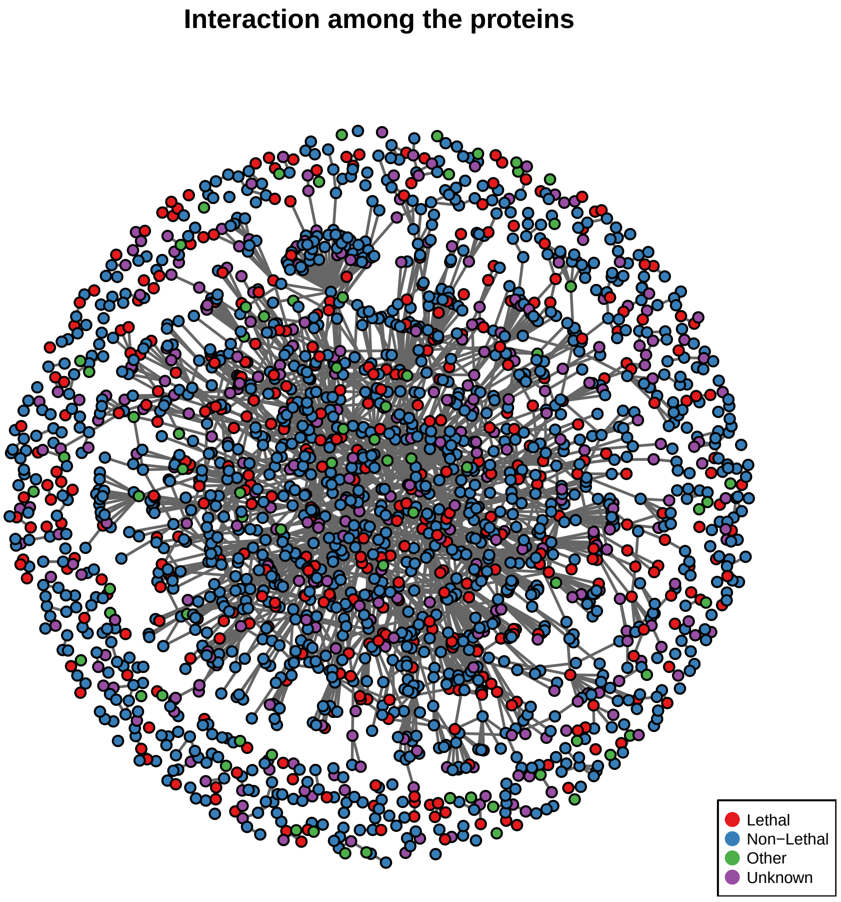

Figure 1: Network visualization

that correspond to the non-lethal and lethal proteins indicating higher degrees and, by interpretation, higher probability of forming a tie with other proteins. In other words, lethal and non-lethal proteins have high degrees. To what class of proteins have a tendency to form ties with remains, at this stage, unexplored. Later data exploration techniques will help bring these patterns to light.

It is also worth mentioning that clusters of protein classified *Unknown* and *Other* are less significant.

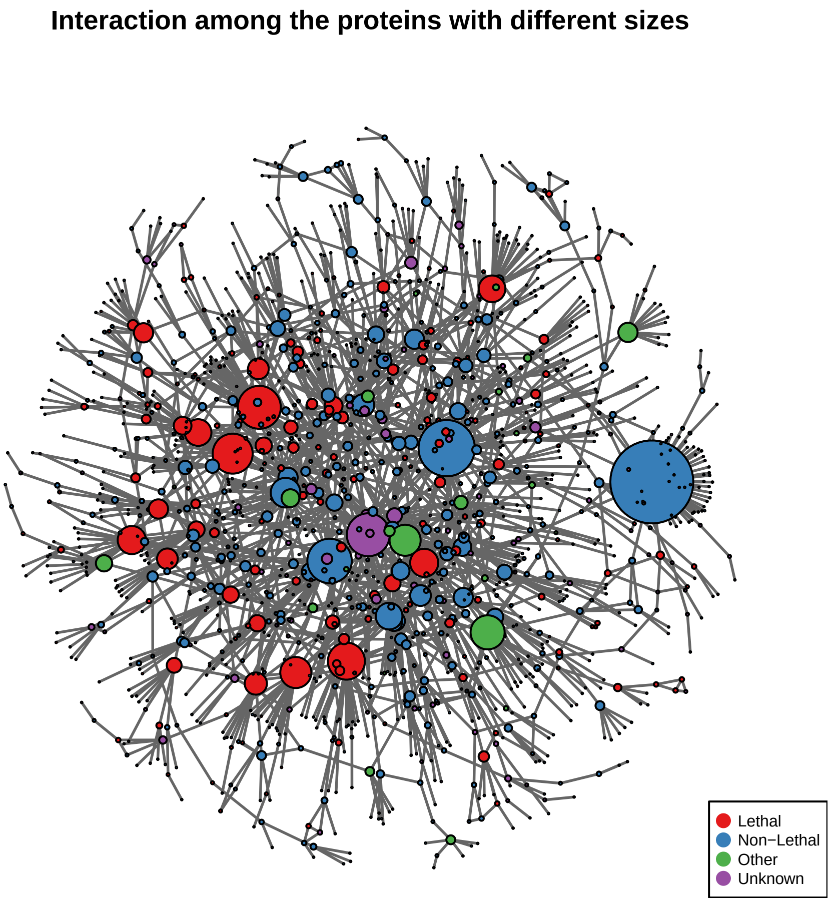

Figure 2: Network showing main components and size according to their degrees

Descriptive analysis of the network

Degree distribution

In addition to visualization, examining network and node characteristics can provide some insight into network structures and possible modeling strategies. Network size and density have been discussed earlier in the introduction. In this section, we will focus on the average number of links per node (mean degree), through a histogram describing the distribution along with a summary statistics table for a better quantitative description.

Table 2: Summary statistics of the degree distribution of PPI network

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	1	1	2.08	2	56

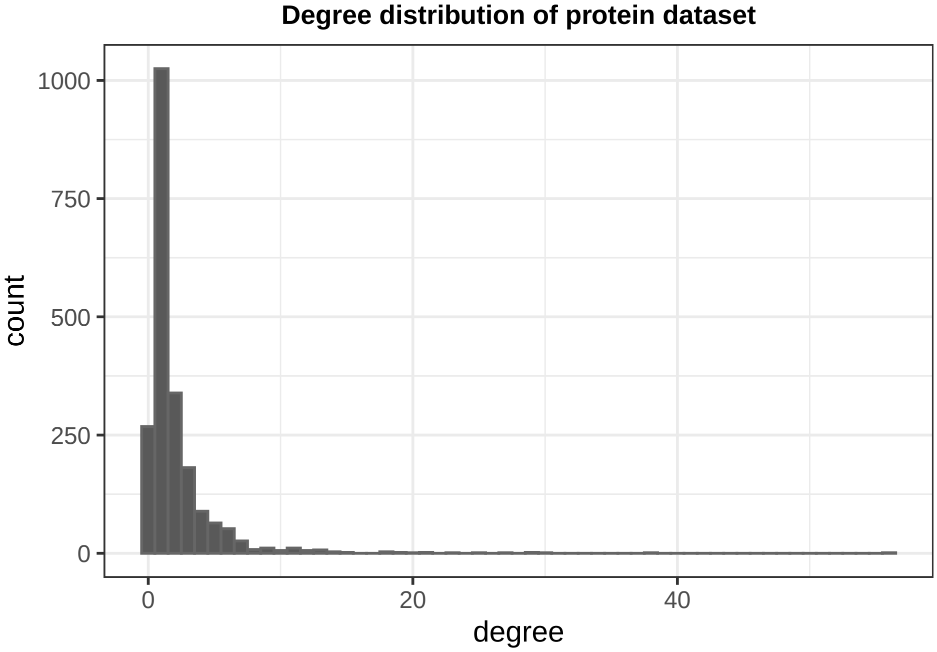

Figure 3: Degree distribution of the proteins

Indeed, like any real world dataset, we see through figure 3 and table **??** that most nodes do not interact worth other nodes indicating that proteins mostly do not interact with each other. The degree distribution histogram also shows that some proteins have a high degree indicating the bigger nodes seen in graph 2 in the PPI network.

The average mean degree is roughly 2.1 indicating that proteins are connected to, and therefore linked with, an average of about 2 other proteins.

Centrality measures

In biological networks, classical centrality measures can offer some rich insight to identify central and, by extension influential nodes (**???**). Typical descriptive centrality measures such for networks such as the closeness, betweenness and eigenvalue centralities will be examined herein in order to identify important nodes in our PPI network.

Table 3: Average closeness centrality measure

	Lethal	Non-Lethal	Other	Unknown
Betweenness	0.006	0.003	0.012	0.003
Eigenvalue	0.004	0.010	0.004	0.011
Closeness	0.154	0.151	0.156	0.147

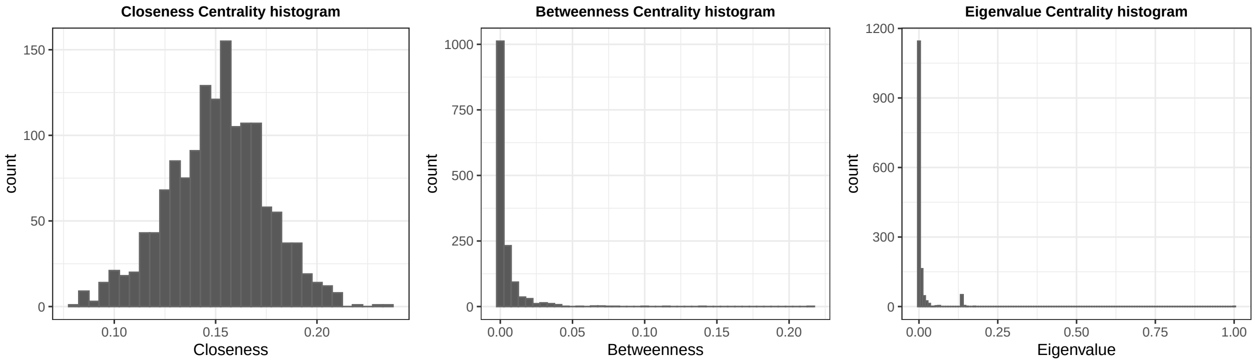

Figure 4: Centrality measures

When examining the plots in figure 4, we notice that the closeness centrality gives Gaussian distribution with minimal changes in scores. This may be an indication of a highly-connected network. The betweenness centrality shows that most nodes (i.e. proteins) are not central, hinting that not many proteins are on the periphery of multiple clusters. The visual exploration of the data in figure 2 also does not show a high number of "central" proteins.

Finally, the Eigenvalue centrality which identifies nodes with influence over the whole network, shows insignificant scores for most of the nodes in the network. Only one node holds a higher value of 1 which, given the total number of nodes in the network, is still insignificant. Therefore, we conclude that centrality measures do not offer much insight at this stage regarding the subclass of the proteins.

Mixing matrix

The network visualizations above showed some potential clustering; another option for identifying clustering is to examine mixing matrices and correlation coefficients. Mixing matrices can be used to examine the number of connected dyads (pairs of proteins in our case) for each possible combination of levels for a categorical node attribute.

For example, how many connected dyads are forming when both proteins are classified as non-lethal or is there some clustering between lethal proteins and non-lethal ones? We already found some evidence of clustering by different levels of lethality and a mixing matrix can help confirm these patterns.

Table 4: Mixing matrix of the attribute lethality

	Lethal	Non-Lethal	Other	Unknown
Lethal	218	532	24	112
Non-Lethal	532	805	127	316
Other	24	127	0	27
Unknown	112	316	27	42

Indeed, the mixing matrix shown in table **??** indicates a bias towards forming ties between proteins when they hold different classes (i.e. non-lethal with lethal). This may be captured later on we take into account homophily in the network in the following section.

More in detail, of the 886 connected dyads including lethal proteins, about 25% (218 proteins) are linked with ones that are also classified as lethal. However, a much higher number of links are formed with non-lethal proteins.

One way to further explore the likelihood of forming a tie based on the class of proteins is through a visual representation of the average number of links for each class. The latter can help further examine the relationship between lethality and network structure.

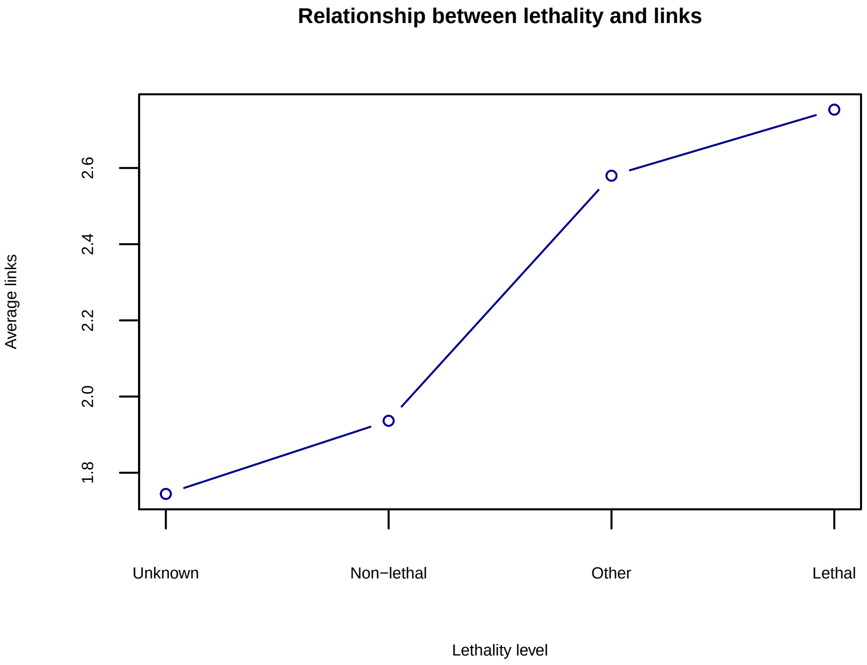

Figure 5: Average link per protein category

Table 5: Average links per class of protein

	Average link
Unknown	1.7
Non-Lethal	1.9
Other	2.6
Lethal	2.8

Figure 5 and table **??** confirm that lethal proteins have a higher number of links thus indicating that proteins of high degree are indeed more likely to be lethal than those of lower degree.

Small World simulation

As with many forms of model building, statistical network modeling should begin with a null model. The null model is a simple random graph model, there are no complex dependence assumptions to account for.

Although the null model is a good representation of the observed density of the PPI network, it is unlikely to be a good representation of other observed network characteristics. Simulating the network's measures based on the small world property can aid in understanding how well the latter captures the structures comprising the observed network.

To gain more knowledge about the PPI network, we explore its small-world property. A small-world network is a graph where most nodes are not neighbors of one other, but any path between two nodes can be reached in a few steps. In order to do this, we compare the properties of a an Erdős–Rényi model with the same number of edges and vertices as the PPI dataset. We specifically look at the main graph and its properties.

```r
df = data.frame(
  diameter = c(19, 16.524),
  avg_sp = c(6.812366, 7.091927),
  clustering_coeff = c(0.0518, 0.001916)
)

row.names(df) = c("PPI Dataset", "ER model")
knitr::kable(
 df, col.names = c("Diameter", "Average shortest path", "Transitivity (Clustering Coefficient)")
)
```

	Diameter	Average shortest path	Transitivity (Clustering Coefficient)
PPI Dataset	19.000	6.812366	0.051800
ER model	16.524	7.091927	0.001916

In order to quantify the property small-worldness, we compare the shortest average path of the random model L_r with the one of our data, L, and also the clustering coefficient of the random model C_r compared to the data C. We expect that $\frac{L}{L_r} \approx 1$ and $\frac{C}{C_r} \gg 1$. We find these features in the tables above, since $\frac{L}{L_r} \approx 0.96$ and $\frac{C}{C_r} \approx 27 \gg 1$. These properties give us a fair bit of evidence that the data has the small-world property.

Modeling and predicting probabilities

Exponential random graph modeling

In the aim of building a model that captures the hidden structure the network as well as answering the question whether node attributes influence the likelihood of a link, we will fit an exponential random graph model. Descriptive statistics throughout the study indicated that the class of proteins may influence the number of interactions among them. To examine the influence of these node attributes on the likelihood of a tie, we will build three different models:

- ERGM where we consider the lethality as the main effect (i.e. using nodefactor)
- ERGM where we consider the interaction terms for the nodal attribute lethality (i.e. nodematch)
- ERGM where both interaction terms as well as main effects are considered.

As means of model comparison and model goodness of fit study, we will look at the Bayesian Information Criterion (BIC) and the diagnostic plots produced by the function *gof* of the *ergm* package in *R*.

Hypotheses testing the main effects and the interaction terms can be worded as follows :

H_0 = *The category lethal does not affect the probability of forming ties between proteins*
H_1 = The category lethal affects the probability of forming ties between proteins

By definition, nodefactor adds multiple statistics to the model, each one equal to the number of times a node with the specified attribute is at one end of an edge. To interpret the results in summary statistics of the model, it is necessary to know the reference group for any categorical variables. In the statnet package, the reference group in an ERGM defaults to the first group in the list shown in the summary of the network.

In our case, the class "lethal" is the reference group for the attribute "lethality" by default. Since the "non-lethal" group contains the most nodes and, by extension, the highest degree, we have set it as the reference group.

The summary statistics of the first exponential random graph model produces the following output:

```
## Call:
## ergm(formula = protein_net ~ nodefactor("lethality", base = 2) + 
##     edges)
##
## Iterations:  9 out of 20 
##
## Monte Carlo MLE Results:
##                              Estimate Std. Error MCMC %  z value Pr(>|z|)    
## nodefactor.lethality.Lethal   0.35252    0.03599      0    9.796  < 1e-04 ***
## nodefactor.lethality.Other    0.28734    0.07756      0    3.705 0.000212 ***
## nodefactor.lethality.Unknown -0.10456    0.04738      0   -2.207 0.027334 *  
## edges                        -7.06790    0.03309      0 -213.612  < 1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 3096207  on 2233441  degrees of freedom
##  Residual Deviance:   34782  on 2233437  degrees of freedom
##  
## AIC: 34790    BIC: 34840    (Smaller is better.)
```

The estimates 'lethal' is significant and positive which indicates an increased likelihood to form ties for proteins when they have been classed as 'lethal' compared to those classified as 'unknown' and 'other'. These results are consistent with the mixing matrix in table **??** and figure 5.

We can then proceed by computing the probability of forming a link between proteins based on the assumption that the main effect is "lethality" using the *plogis()* function. The result is 54.7% which is significantly high based on the fact that the density of network is just below 0.001.

In order to unveil the homophily in the network, we will fit the second model considering interaction terms. The summary statistics of the model is the following :

```
## Call:
## ergm(formula = protein_net ~ edges + nodematch("lethality", diff = T))
##
## Iterations:  8 out of 20 
##
## Monte Carlo MLE Results:
##                                Estimate Std. Error MCMC %  z value Pr(>|z|)    
## edges                          -6.97053    0.02965      0 -235.069   <1e-04 ***
## nodematch.lethality.Lethal      1.06547    0.07402      0   14.394   <1e-04 ***
## nodematch.lethality.Non-Lethal -0.03720    0.04607      0   -0.808    0.419    
## nodematch.lethality.Other          -Inf    0.00000      0     -Inf   <1e-04 ***
## nodematch.lethality.Unknown    -0.06121    0.15718      0   -0.389    0.697    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 3096207  on 2233441  degrees of freedom
##  Residual Deviance:   37973  on 2233436  degrees of freedom
##  
## AIC: 37981    BIC: 38032    (Smaller is better.) 
##
##  Warning: The following terms have infinite coefficient estimates:
##   nodematch.lethality.Other
```

The first thing to notice is that the estimate that corresponds to nodematch.lethality.Other is -Inf. Looking again at the mixing matrix in table **??**, we conclude that this is due to the absence of interactions of this class with other classes. Therefore, we will continue focusing on the two main classes of the network "Lethal" and "Non-lethal".

The resulting model shows significant positive coefficient for lethal homophily in the network. That is, two proteins in the same category "lethal" are more likely to be connected further confirming our findings. The model also shows non significant coefficients for the class non-lethal. In this case, predicting the probability of forming a tie goes up to 74.3% which means that the homophily is better captured in this model compared to the one where we only consider lethality as the main effect.

However, both models contained estimates in which the p-value is small enough to reject the null hypothesis.

The last model that we will test is one where we include both nodematch and nodefactors for the attribute lethality which gives the following summary :

```
## Call:
## ergm(formula = protein_net ~ edges + nodematch("lethality", diff = T) +
##     nodefactor("lethality", base = 2))
##
## Iterations:  8 out of 20
##
## Monte Carlo MLE Results:
##                                Estimate Std. Error MCMC % z value Pr(>|z|)
## edges                          -6.96224    0.10230      0 -68.055  < 1e-04 ***
## nodematch.lethality.Lethal      0.97373    0.12539      0   7.766  < 1e-04 ***
## nodematch.lethality.Non-Lethal -0.04549    0.10821      0  -0.420  0.67418
## nodematch.lethality.Other          -Inf    0.00000      0    -Inf  < 1e-04 ***
## nodematch.lethality.Unknown     0.27262    0.18999      0   1.435  0.15129
## nodefactor.lethality.Lethal     0.04172    0.09467      0   0.441  0.65943
## nodefactor.lethality.Other      0.30355    0.10593      0   2.866  0.00416 **
## nodefactor.lethality.Unknown   -0.17106    0.09242      0  -1.851  0.06417 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 3096207  on 2233441  degrees of freedom
##  Residual Deviance:   37946  on 2233433  degrees of freedom
##
## AIC: 37960    BIC: 38049    (Smaller is better.)
##
##  Warning: The following terms have infinite coefficient estimates:
##   nodematch.lethality.Other
```

We notice an overall worse fit when we include the above features in our model. Estimates seem to be slightly less significant than the first two models where only main effect and similarity effect are considered independently. In the next subsection, we will further explore the goodness of fit diagnostics.

Goodness of fit test

Finally, we will summarize all three BIC and AIC scores for all three models for comparison in the following table:

Table 7: BIC & AIC score for all three models

	With main effect	With nodematch	nodematch & main effect
BIC	34840	38032	38049
AIC	34790	37981	37960

We can conclude that the model where lethality is considered the main effect yields the lowest BIC and AIC score indicating a better fit than the other two models seen above. To further explore the goodness of fit of the statistical model, we will compute the plots produced using the *gof* function.

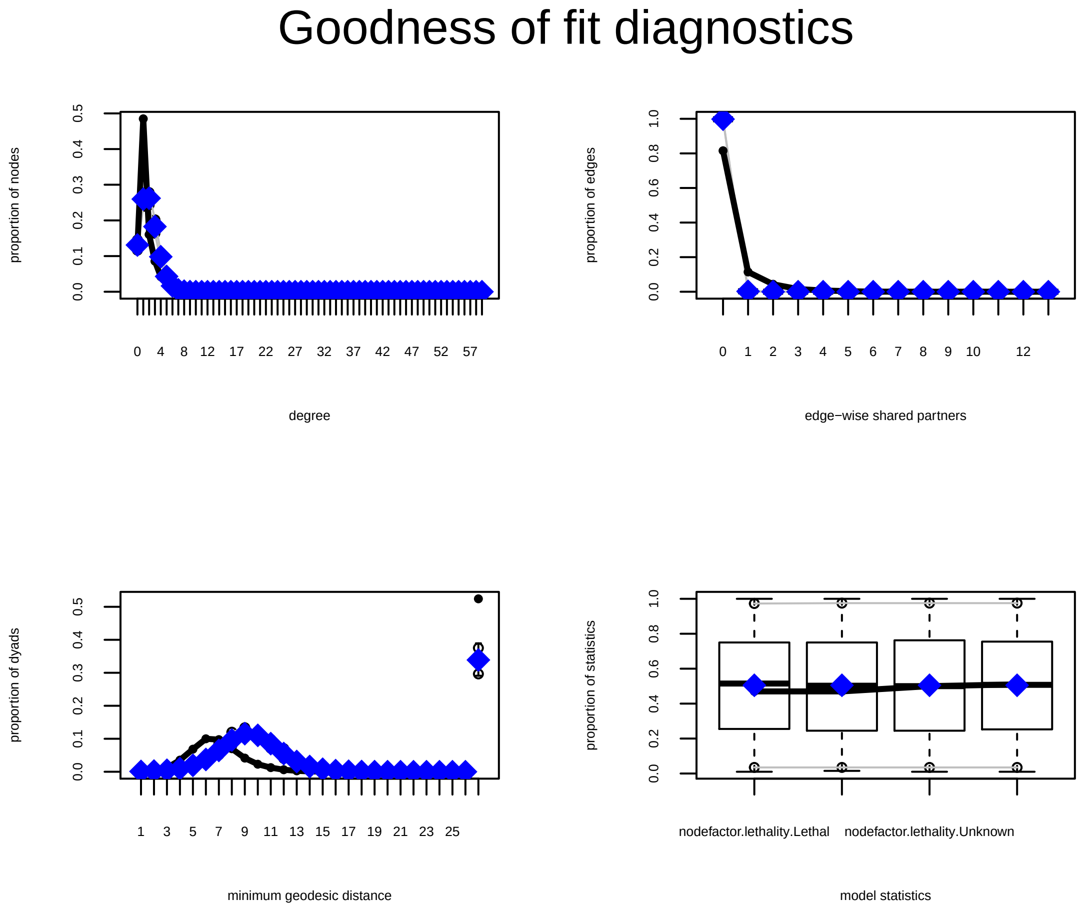

Figure 6: Diagnostic plots of the goodness of fit

Visual analysis of the above goodness of fit plots indicate that the model is able to reconstruct the proportion of edges with respect to the edgewise shared partner terms. This is an indication that our model was able to capture the underlying transitivity in the PPI network. The overall statistics of the levels of the attribute "lethality" are also well represented with a slight underestimation for the nodefactor "unknown".

However, the model overestimates the proportion of nodes that have lower degrees. For example, we notice a significant peak in the estimated line (code colored black) compared to the empirical estimations (code colored blue). This may be an indication that the degree distribution is not well represented by the model we built. One way to improve the fit of the model is to account for the degree distribution using the gwdegree term which stands for geometrically weighted degree.

Adding dependent terms to improve the model

First, we tune our parameter of decay for the *gwdegree()* function, between 0.1 and 0.9. We do not look for higher parameters, since the decrease of the degree distribution is fast. Therefore, it seems reasonable to search a grid of small values of the decay parameter. Next, we will extract the BIC score of each value in the sequence of the decay parameter and refit the model using the optimal one.

```
## Call:
```

```
## ergm(formula = protein_net ~ edges + nodefactor("lethality",
##     base = 2) + gwdegree(min.decay, fixed = TRUE))
##
## Iterations:  5 out of 20
##
## Monte Carlo MLE Results:
##                               Estimate Std. Error MCMC %  z value Pr(>|z|)
## edges                         -5.93349    0.04810      0 -123.356  < 1e-04 ***
## nodefactor.lethality.Lethal    0.19645    0.02886      0    6.807  < 1e-04 ***
## nodefactor.lethality.Other     0.16269    0.05292      0    3.074  0.00211 **
## nodefactor.lethality.Unknown -0.05670    0.03734      0   -1.519  0.12883
## gwdeg.fixed.0.8               -1.41134    0.05767      0  -24.473  < 1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 3096207  on 2233441  degrees of freedom
##  Residual Deviance:   34295  on 2233436  degrees of freedom
##
## AIC: 34305    BIC: 34369    (Smaller is better.)
```

The BIC criterion is slightly better, so it seems a better fit than the other models. Let's check the goodness of fit with the following plot :

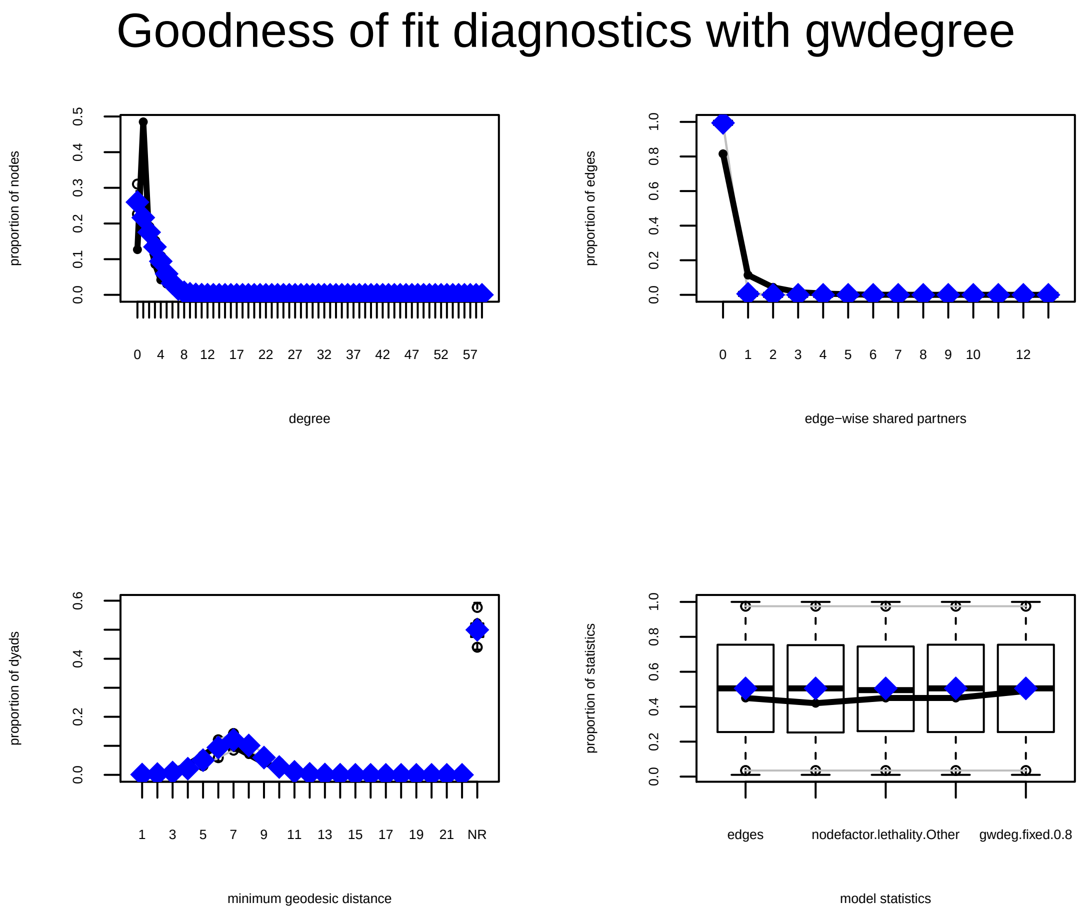

Figure 7: Diagnostic plots of the goodness of fit with gwdegree

Here, the minimum geodesic distance indicates a good fit of the model, and the transitivity is still well captured and the degree distribution estimation is improved over the first model. However, we still over estimate the proportion of nodes in the simulated network produced by the model. We will try to improve this fit by taking into account both the degree distribution and the transitivity and get its summary statistics and its diagnostic plots for comparison.

```
## Call:
## ergm(formula = protein_net ~ edges + nodefactor("lethality",
##     base = 2) + gwdegree(log(2), fixed = TRUE) + gwesp(log(3),
##     fixed = TRUE))
##
## Iterations:  19 out of 20
##
## Monte Carlo MLE Results:
##                                Estimate Std. Error MCMC %  z value Pr(>|z|)
## edges                          -6.66608    0.05736      0 -116.220   <1e-04 ***
## nodefactor.lethality.Lethal     0.18878    0.03234      0    5.837   <1e-04 ***
## nodefactor.lethality.Other      0.13158    0.06710      0    1.961   0.0499 *
## nodefactor.lethality.Unknown   -0.04213    0.04086      0   -1.031   0.3025
## gwdeg.fixed.0.693147180559945 -0.76208    0.06210      0  -12.272   <1e-04 ***
## gwesp.fixed.1.09861228866811   1.45651    0.03496      2   41.658   <1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 3096207  on 2233441  degrees of freedom
##  Residual Deviance:   33172  on 2233435  degrees of freedom
##
## AIC: 33184    BIC: 33259    (Smaller is better.)
```

Goodness of fit diagnostics with gwdegree and gwesp

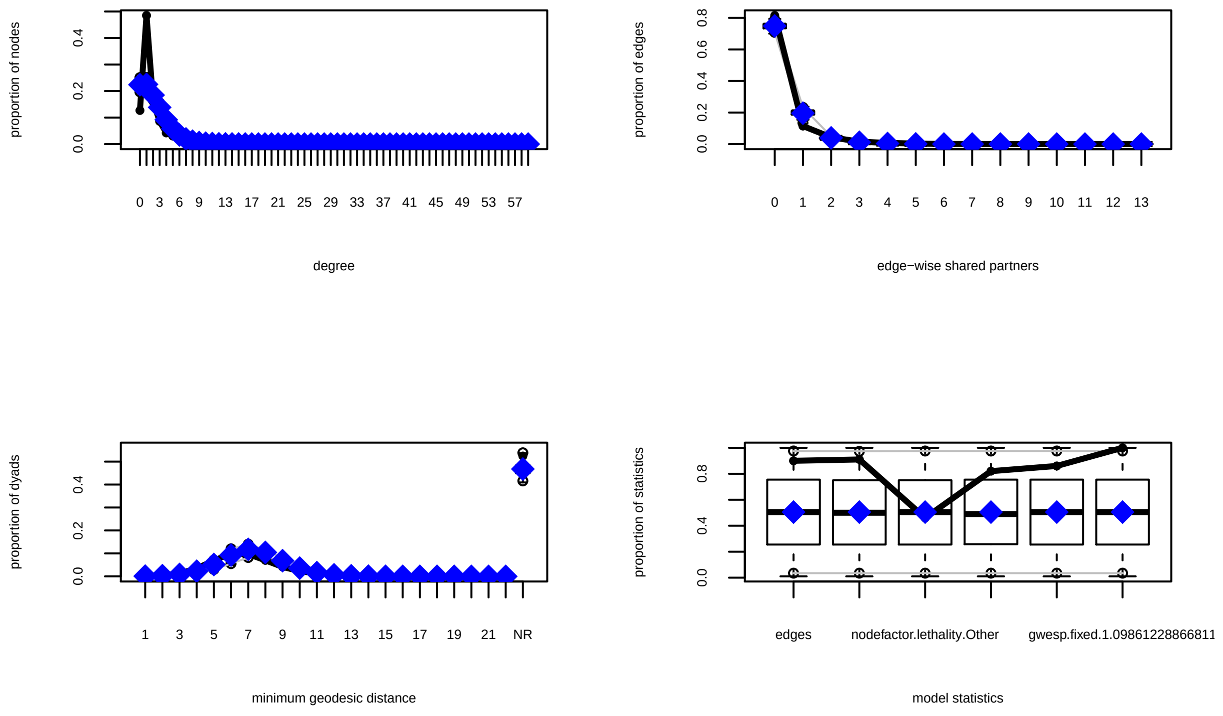

Table 8: Table summarizing all BIC scores for the models

	BIC	V2
gwdegree and gwesp	33259.34	33259.34
gwdegree	34368.59	34368.59
nodefactor	34839.99	34789.51
nodematch	38031.79	37981.32
nodefactor and nodematch	38048.67	37960.34

Summary statistics of the last model along with figure **??** show a good fit except the proportion of statistics across the lethality attribute and an over estimation of the proportion of nodes. In contrast, it yields the lowest BIC score out of all the models tested.

Figure 7, shows an overall better fit for our network when exploring the 4 plots produced by the gof function and its summary statistics. The increase in its BIC score is acceptable given the fact that it gives better estimations of the model's statistics. We conclude that the model where we only consider gwdegree is best one based on the produced results.

Conclusion

Knowledge about how they interact can lead to better understanding of a great many diseases and it can help in the design of appropriate therapies. Results showed that proteins of high degree were far more likely to be lethal than those of lower degree. Findings through the exploratory data analysis and the modeling of the network showed consistent results. We were able to extract rich insight from the network analysis of the PPI.

References

Jeong, Hawoong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. 2001. "Lethality and Centrality in Protein Networks." *Nature* 411 (6833). Nature Publishing Group: 41–42.