Using Topology Information for Protein-Protein Interaction Prediction

Adriana Birlutiu[(1),(2)] and Tom Heskes[(1)]

[(1)] Institute for Computing and Information Sciences,
Radboud University Nijmegen, The Netherlands
[(2)] Faculty of Science,"1 Decembrie 1918" University of Alba-Iulia, Romania

Introduction and motivation for our work

PPI networks

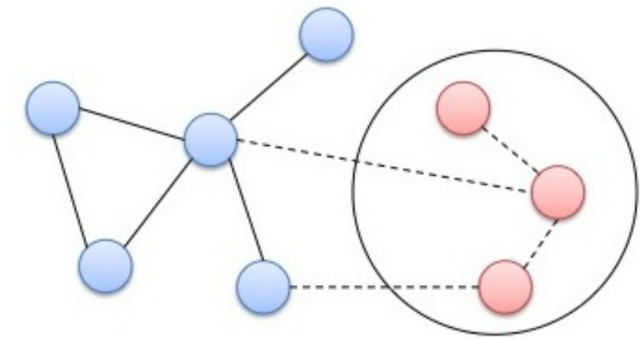
- Knowledge about protein-protein interactions (PPIs) is an active research area with important applications in biology and network medicine
- PPIs can only be established by tedious and costly laboratory experiments
- Computational methods for PPI prediction complement the laboratory experiments (correct and complete biological networks, guide future laboratory experiments)

Introduction and motivation for our work

Computational methods for PPI prediction

- Computational methods take roots in machine learning: frame the problem of PPI prediction in a supervised learning setting
- Infer missing edges in a graph (*dotted edges*) from the edges are already known (*solid edges*)
- Information about genes or proteins (sequence, structure, expression level) can give hints about presence/absence of interactions
- o,o': two proteins, $x(o)$ and $x(o')$: input feature vectors encoding some properties of o and o'

Learn a function $f : (x(o), x(o')) \rightarrow \{0, 1\}$ from training data



Introduction and motivation for our work

Topological properties of PPI networks

- Node degree distribution

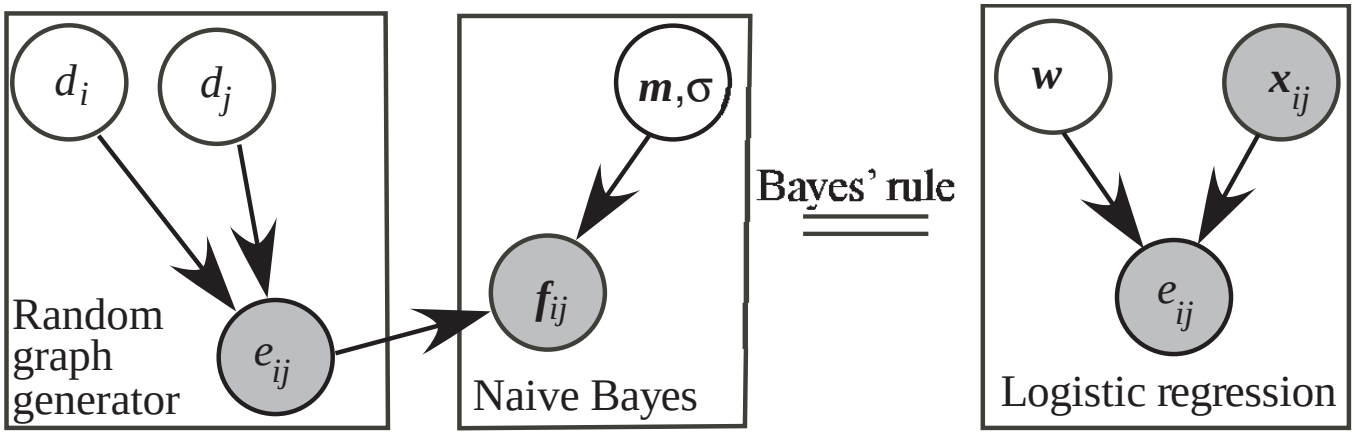
$$P(k) = \frac{N_k}{N} ,$$

  where N is the total number of nodes in the network and N_k is the number of nodes with degree k.
- Node degree distribution is right-skewed: most nodes have low degrees while a small number of nodes have high degrees (hubs)
- Clustering coefficient, network diameter, average shortest path
- Network motifs: small subgraphs which appear in the network significantly more frequently than in a random network

How can this topological information be used and integrated with other types of information?

Bayesian framework for integrating topology and feature information

Bayesian model for integrating network topology information with protein features information



Random graph generator for a give topology

Latent variable, d_i, related to the degree of node i.

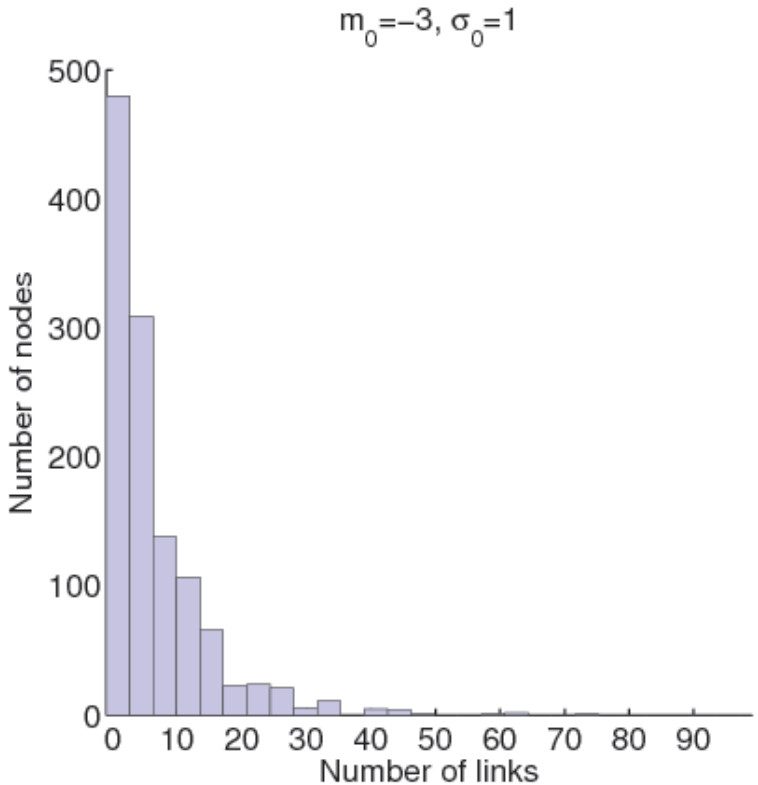
$$p(e_{ij}|d_i, d_j) \propto \exp\left[e_{ij}\frac{1}{2}(\log d_i + \log d_j)\right]$$

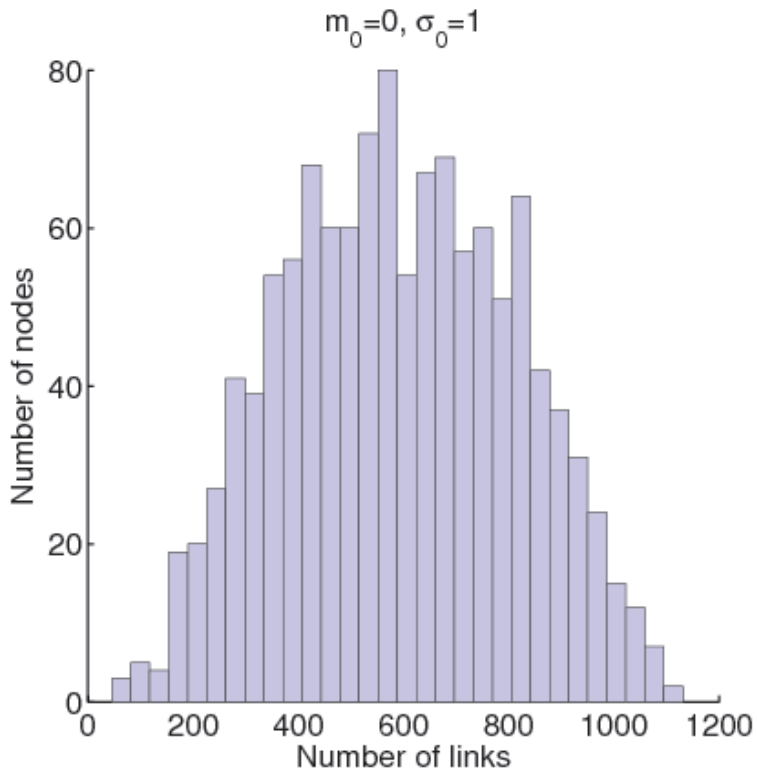
We consider a log-normal distribution for d_i.

1. Choose m_0 and σ_0 the parameters of the log-normal distribution for d_i.
2. Draw from this distribution a random sample $(d_1, \ldots, d_N)$ of size N the number of nodes in the network.
3. Based on this sample construct the network by inserting edges with probability given above.

Node degree distributions

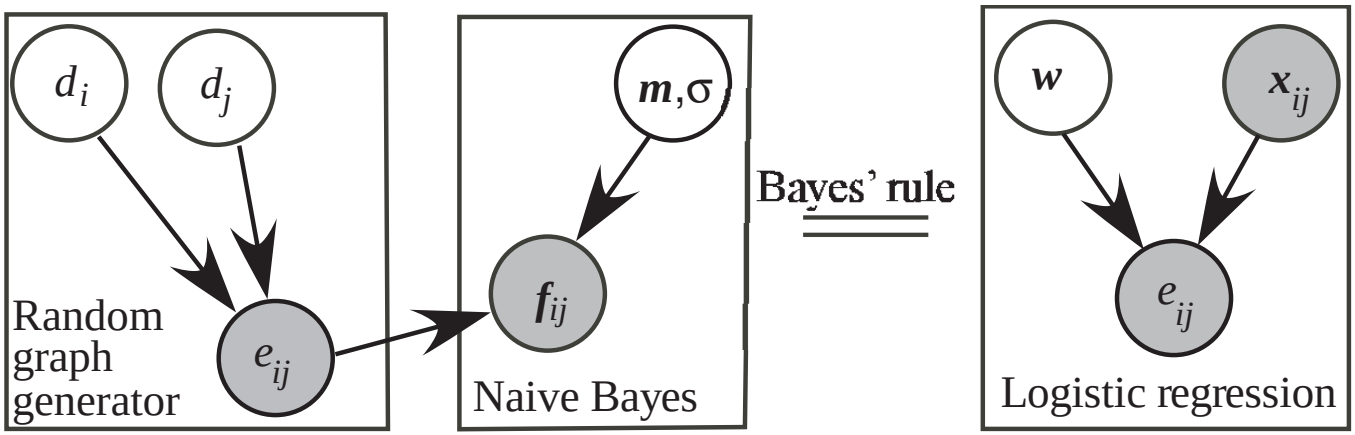
Networks randomly generated with different parameter settings





Bayesian framework for integrating topology and feature information

Bayesian model for integrating network topology information with protein features information



Likelihood model

We use a naive Bayes model to express the likelihood of a protein-pair features given the absence/presence of an interaction:

$$P(\mathbf{f}_{ij}|e_{ij},\mathbf{m},\sigma) = \prod_{k=1}^{D} \mathcal{N}(f_{ij}^{k}; m_k e_{ij}, \sigma) \propto \prod_{k=1}^{D} \exp\left(-\frac{(f_{ij}^{k} - e_{ij} m_k)^2}{2\sigma^2}\right).$$

Combining topology and feature information

The posterior distribution for e_{ij} which combines topology and feature information is computed using Bayes rule:

$$p(e_{ij}|\mathbf{f}_{ij}, d_i, d_j) \propto p(e_{ij}|d_i, d_j)p(\mathbf{f}_{ij}|e_{ij}, d_i, d_j)$$

Adjoin the unknown quantities in a single random variable:

$$\mathbf{w} = [\frac{m_1}{\sigma^2}, \ldots, \frac{m_D}{\sigma^2}, \frac{1}{2}\log d_1, \ldots, \frac{1}{2}\log d_N] \,,$$

and the protein features and topological information

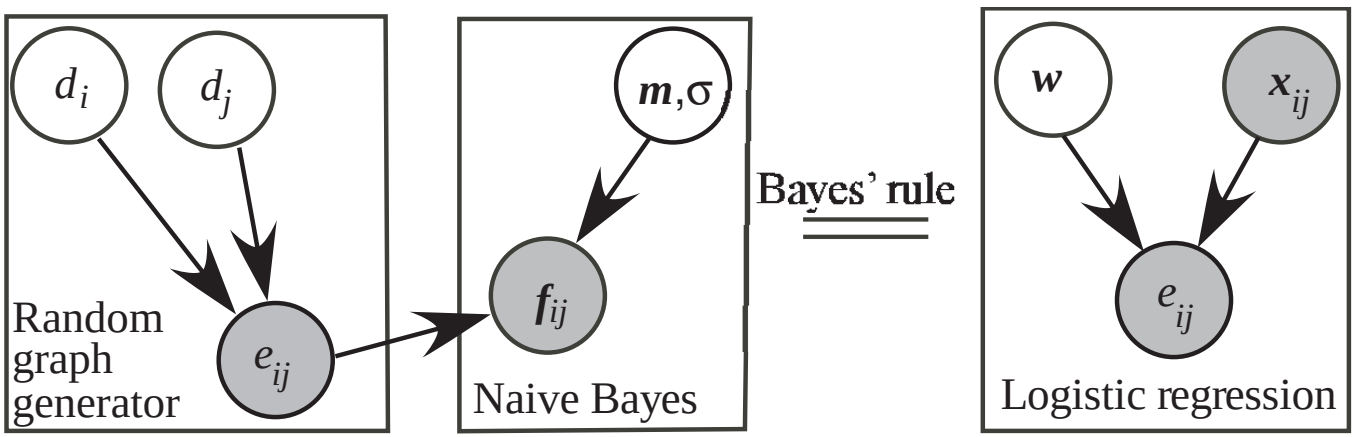
$$\mathbf{x}_{ij} = [\mathbf{f}_{ij}, \mathbf{t}_{ij}] \,,$$

The unknown parameter $\mathbf{w}$ is learned in a Bayesian framework

$$P(\mathbf{w}|\text{observations}) \propto \prod_{o=1}^{n_{\text{obs}}} P(e_{ij}^o|\mathbf{x}_{ij}^o, \mathbf{w})P(\mathbf{w}) \,.$$

Bayesian framework for integrating topology and feature information

Bayesian model for integrating network topology information with protein features information



Experimental evaluation

We compare four models:

- Model 1 (Features+Topology)
- Model 2 (Features only)
- Model 3 (Topology only)
- Model 4 (Topology-enriched features)

Experimental evaluation

Data sets

1. Yeast data set: PPI network has 984 nodes (proteins) connected by 2438 links (interactions). Each protein has associated a vector of dimension 157 representing gene expression values in various experiments
2. Human data set: the PPI graph consists of $24,380$ nodes connected by $14,608$ edges. Each pair of proteins is characterized by a 27-dimensional feature vector

Experimental evaluation

Experimental protocol

- We randomly sampled a training set containing 1%, 5%, 10% and 20% protein pairs and their labels as interacting or not from the yeast and human data set. These data samples were used to train the classification models
- The remaining protein pairs were used for testing the performance. AUC scores were computed on the test set
- We report average results (mean $\pm$ standard deviation) over 10 random runs

Experimental evaluation

Results

% Train data	Model 1 Features+ Topology	Model 2 Features only	Model 3 Topology only	Model 4 Topology features
1%	**0.639 ± 0.014**	**0.639 ± 0.018**	0.577 ± 0.016	0.582 ± 0.022
5%	**0.708 ± 0.006**	0.697 ± 0.009	0.688 ± 0.010	0.689 ± 0.009
10%	**0.731 ± 0.005**	0.712 ± 0.005	0.720 ± 0.006	0.717 ± 0.007
20%	**0.746 ± 0.009**	0.719 ± 0.006	0.742 ± 0.009	0.737 ± 0.010
1%	**0.863 ± 0.006**	0.851 ± 0.006	0.608 ± 0.014	0.822 ± 0.012
5%	**0.909 ± 0.002**	0.859 ± 0.001	0.793 ± 0.007	0.899 ± 0.003
10%	**0.931 ± 0.002**	0.861 ± 0.001	0.864 ± 0.005	**0.931 ± 0.002**
20%	0.952 ± 0.002	0.862 ± 0.001	0.917 ± 0.003	**0.954 ± 0.002**

Conclusions

- We introduced a framework for predicting PPI by considering the network topology information
- Bayesian framework consisting of a prior distribution over the network topology and likelihood terms for observations about links in the network
- Simplifying assumptions which reduce the computational complexity and at the same time yield a good performance

Conclusions

- We introduced a framework for predicting PPI by considering the network topology information
- Bayesian framework consisting of a prior distribution over the network topology and likelihood terms for observations about links in the network
- Simplifying assumptions which reduce the computational complexity and at the same time yield a good performance

Thank you for your attention!