

**NETWORK MIXTURE MODELS: AN INTRODUCTION AND  
APPLICATION TO PROTEIN INTERACTOMES**

By

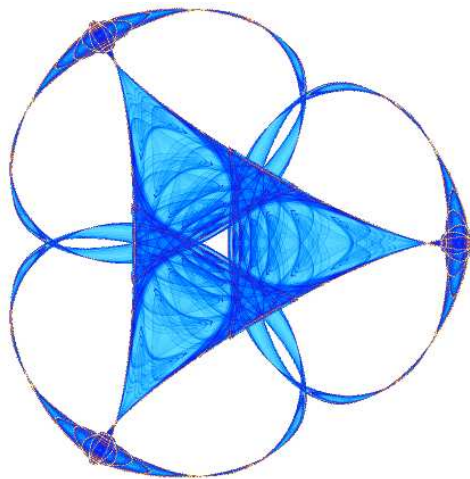
**Elisabetta Marras**

and

**Enrico Capobianco**

**IMA Preprint Series # 2230**

(January 2009)



**INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS**

UNIVERSITY OF MINNESOTA  
400 Lind Hall  
207 Church Street S.E.  
Minneapolis, Minnesota 55455-0436

Phone: 612-624-6066 Fax: 612-626-7370

URL: <http://www.ima.umn.edu>

# Network Mixture Models: an Introduction and Application to Protein Interactomes

*Elisabetta Marras and Enrico Capobianco*

CRS4 Bioinformatics Laboratory

Science & Technology Park of Sardinia

09010 Pula (Cagliari), Italy

*lisa@crs4.it, ecapob@crs4.it*

January 8, 2009

## Abstract

Human Protein Interactomics represents a crucial research area in systems biology, and several new data sources and methods have been recently proposed. In our work, we refer to data integration and network inference domains, and with focus on group selection and modularity refinement. Our reference dataset is Ulitsky, Karp and Shamir (<http://acgt.cs.tau.ac.il/clean>), which integrates various sources, *BIND*, *HPRD*, *BioGRID*, and includes interactions detected from small-scale experiments, thus listing 7,385 nodes and 23,462 links. Concerning network inference, we explore graph mining techniques from both a deterministic and a probabilistic standpoint. It is widely adopted the application of the former methods to protein interactomes, of which we provide two examples, but it is definitely harder to see the application of probabilistic methods. Comparative examples are provided here. We look first at coreness- and community-based grouping techniques, but then also at statistical mixture modelling with the goal of detecting human protein sub-interactomes. We suggest a novel way to select and prioritize information otherwise usually extracted under connectivity-based principles. A graph can be represented by an adjacency matrix (AM) embedding the interaction and non-interaction structure of the observed graph. The result for the case of undirected graphs is a binary AM. We show that the AM can be efficiently sampled to deal better with the inherent sparsity property of biological networks, and then proceed by adopting mixture models under flexible statistical assumptions.

*Keywords:* Protein Interactome Networks; Adjacency Matrix; Sub-sampling; Mixture Models.

## 1 Introduction

The current Protein Protein Interaction Networks (PPIN) interest growth (Hart et al, 2006; Vidal, 2005) is linked to the flow of data from modern high-throughput techniques targeted

to cover more and more organisms with increased accuracy. Human Protein Interactomics (HPI), in particular, represents a crucial topic in systems biology, particularly for disease-targeted application studies.

Some mainstream research domains are: **1.** Build-up of interactome warehouses from several biological sources, which call for data mining and integration (Chaurasia et al, 2007; Deane et al, 2002; Ewing et al, 2007; Ito et al, 2001; Lu et al, 2005; Rual et al, 2005; Uetz et al, 2000; von Mering et al, 2002); **2.** Development of network inference models (Bader et al, 2003; Edwards et al, 2002; Gavin et al, 2006; Jansen et al, 2003; Krogan et al, 2006; Sharan et al, 2007; Troyanskaya et al, 2003), which lead to new learning paradigms and method fusion techniques; **3.** Prioritization of protein identification and classification approaches, which drive disease-specific applications (Jonsson et al, 2006; Ulitsky et al, 2008; Wachi et al, 2005; Xu and Li, 2006).

In particular, many bioinformatic applications are implementing *in silico* classification and prediction of putative interactions through methods designed to extract protein complexes or functional modules. Apart from the limitations inherent to a protein interactome covered only up to a certain extent, and measurable with just sub-optimal accuracy according to the available sources (yeast two-hybrid, co-IP, literature mining, orthology, etc.), other aspects are worth mentioning.

Overall, a protein interaction map consists of real interactions and correspondingly non interactions, together with a complementary part of missing information (false negatives) and errors (false positives), which can of course change with time. Thus, a mix of static and dynamic associations appears, depending on the degree of uncertainty of the map. Ideally, both transient and persistent protein connectivity are characterizing the underlying structure across time scales. As knowledge of such variability is not in general available, or experimentally hard to obtain, then data integration, dimensionality reduction, and network learning become extremely relevant.

Data integration leads to the fusion of heterogeneous "omic" sources to improve the accuracy of the global protein interaction map. This at the price of adding further complexity, for instance by increasing the problem dimensionality. Proper scaling and normalization become necessities in order to handle heterogeneous data and platforms, and a natural shift to a more probabilistic treatment of the data can be consequent because of the confidence with which feature measurements might be characterized.

The latent dimensionality in PPINs is high, and as a consequence data are sparse and remain so even after feature reduction. Depending on this "curse of dimensionality" problem, machine learning algorithms are prone to give sub-optimal results. For each organism there is only a finite number of validated interacting protein pairs, and this coverage is very limited in cases like HPI. A much bigger set of non-interacting protein pairs is complementarily found, which is useful to build null models according to various possible schemes, and validate possible structural features.

Network learning leverages on graph mining and inference. The former topic aims to extract subgraphs from a large graph, and in PPINs there are specific kinds of biological sub-structures one would like to retrieve, such as protein complexes and functional modules. Examples of other sub-structures with an algorithmic nature are instead provided when cliques (or complete graphs), communities, and hubs are retrieved. The idea of cliquishness is very practical, as it delivers a dense set of interacting nodes which might represent a

functional unit.

Two popular approach for deterministic partitioning have specialized the cliquishness concept and are here studied and applied: MCODE (Bader et al, 2003), which is centered on the idea of coreness, and CFinder (Palla et al, 2005), centered on the idea of community. We leave details for later, and just stress at this point the fact that both methods depend on node connectivity through cliques<sup>1</sup> or k-cliques<sup>2</sup>, which roughly speaking are nested sub-structures. However, it holds that connectivity is just one of several classes of topological features. We thus explore other approaches.

The paper has the following structure: Section 2 introduces a probabilistic view of protein networks; Section 3 describes the gold standard concept; Section 4 is devoted to inference with some aspects such as dimensionality, model choice and network mixtures which are emphasized, while particular attention is given to sub-sampling strategies and a novel approach. Section 5 elucidates deterministic versus probabilistic methods for graph mining, and we implement the idea of mixture networks. Then, Section 6 is for the conclusions.

## 2 A probabilistic view of PPIN

A probabilistic approach in dealing with PPIN can be pursued in various ways, starting from basic ideas. An average network connectivity can be obtained by considering a fluctuation range according to some probability associated to an unobserved variable distribution. This idea is behind many latent variable models (*LVM*) popular in statistics (Bartholomew, 1987; Everitt and Hand, 1981; Goodman, 1974; McLachlan and Peel, 2000; Titterton et al, 1985) and leading to semi- or non-parametric joint estimation of a set of observed and hidden variables.

By drawing a joint distribution over both these variables, the observed one then be obtained by marginalization; complex distributions can now be more tractable despite a large variable space. A well-known example of *LVM* is the mixture distribution model in which the hidden variable is the discrete component label (but can also represent factor analyzers in case of continuous latent variables).

The *LVM* framework is also suitable for dealing with linear and nonlinear path analysis or graphical model. A network too, when characterized by an unknown degree distribution, may refer to the same statistical setting. In particular, it might be originated by a mix of networks with known degree distribution. Furthermore, the *LVM* associated with a scale-free network should decay as a power law. The problem has an inverse nature, as it addresses the recovery of the unobserved component from the observed degree distribution.

In particular, the rationale for using mixture models in PPINs refers to existing relationships between the observed protein-protein interactions (our data) and the latent intrinsic coordinates of the underlying manifold (complexes, functional modules, etc.). Given locally linear and smoothly varying interaction dynamics (such relaxed assumptions compensate for the "omic" information we lack in our approach, such as expressions, etc.), we attempt to link the observed and latent information layers embedded in PPINs.

---

<sup>1</sup>A clique is defined as a maximally connected graph or an induced sub-graph which is a complete graph, i.e. a simple graph in which every pair of distinct vertices is connected by an edge.

<sup>2</sup>Fully connected subgraphs that share exactly  $k$  nodes in a network are called k-cliques.

As a result, it is through mixtures that we can build a probabilistic model valid away from the training set, and thus in principle able to classify proteins to groups, generalize or infer protein functions, predict or assign scores to protein-cluster associations. We show that probabilistic graph partitioning via mixture modelling offers different perspectives relatively to clustering from deterministic partitioning, and allows the approximation of densities from high dimensional data manifolds to lower ones.

### 3 Gold Standard

The uncertainty which naturally appears when extracting biological information depends on the protein interaction map and on the method detection's power. This problem presents a difficult quantification aspect, and can be faced by assigning probabilities as scores or weights to the network edges.

Interaction data are of two types: binary interactions (for protein pairs), and multiple interactions (for groups of interacting partners, such as protein complexes). The interaction network can be weighted or not, depending on the presence of weights on the links measuring the confidence we might have about the existence of the interactions.

The quality of interaction data is measured by coverage and accuracy; the former refers to false negatives, i.e. the missing information, while the latter refers to false positives, i.e. bad measurements. Currently, coverage represents a strong limitation since reference datasets with embedded new experimental evidence are often incomplete, and bias can be present (for instance, towards proteins of high abundance or cellular localization).

With regard to the methodology, the highest accuracy is achieved when more than one detection method supports the interaction, and for this reason literature-curated datasets have become precious sources. In order to correct the experimental data for the heavy presence of several spurious effects, features such as gene expression, knockout phenotype, subcellular localization, genetic interaction and phylogenetic profiles need to be integrated to the original raw interaction data for improving the precision level.

Due to substantial differences between methods, where some detect physical binding between proteins, and others genetic interactions, shared pathways, etc., a definition of a true positive reference dataset is needed, and based on the degree to which interacting proteins are annotated with the same functional category, the known protein complexes, 3D structures, etc. This result should naturally lead to high-confidence interaction maps.

The design of a so-called probabilistic network assumes that edges have weights, which can also be simply 0 – 1, thus characterizing binary interactions. Edge weights would add information, i.e. whether candidate proteins can be ranked according to membership in a known protein complex. Thus, protein classification would be easier. Usually, estimating a fraction of sampled networks that contains a certain connecting path between proteins would also allow error quantification, function generalization and score prediction.

A general approach assigns a reliability measure to the observed interactions. *Confidence scores* are usually formulated depending on the type of experiments performed, and on the integration among the available information sources. Thus, the approach is data-driven. The second approach predicts the interactions from a combination of features such as functional similarity, expression correlation, co-essentiality etc, and is centered on a so-called (*log-*

)*likelihood score*. This approach is model-driven.

Overall, data integration is behind both scores, but while confidence scores refer more to qualitative aspects of data sources, likelihood scores refer to probability ratios, i.e. the probability of observing a feature according to a chosen reference model compared to a background or null model. This analysis can then be extended to a comparison of different interaction probability assignment schemes, or to log-likelihood ratios relatively to reference and null models.

Useful definitions (see Jansen and Gerstein, 2004), among others) functional to the above results are *gold standards* (GS), i.e. set of positive ( $GS^+$ ) and negative ( $GS^-$ ) interaction sets, depending on what features characterize (or correlate with) the truly interacting or non-interacting proteins. For instance,  $GS^+/GS^-$  may refer to proteins inside/outside a cellular compartment, to interactions determined by different experimental methods, to the expression profiles of interacting partners, to small-scale experiments (see also Deane et al, (2002) for concrete examples).

Thus, GS are not unique, and can be refined to be more or less rich or conservative. Furthermore, it is not trivial to establish boundaries between the positive and negative classes, which is a necessary step to achieve the best possible separation between cases, and for protein function predictive algorithms to work properly.

## 4 Inference

### 4.1 Dimensionality

At present, the limited coverage extent of many organisms suggests that truly connected protein pairs are much less than the observed high dimensional map, which in turn implies that PPIN's intrinsic dimensionality and degrees of freedoms are driven by only a small number of variables. Common dimensionality reduction methods can be used with the aim to encapsulate information within a few salient dimensions. This sort of embedded dimensionality emphasizes the role of a limited number of relevant features or inner structures.

Such structures provide a natural representation of possible physical interactions among groups of proteins. We expect these clusters to have a higher density of points than their surrounding regions. However, measuring densities depends on distance measures assigned to interacting proteins, and this step would involve determining how to do so, i.e. what weight to each link and with what confidence.

Statistical inference methods elucidate PPIN's distributional aspects, starting by features such as degree distribution and clustering coefficient<sup>3</sup>. Figure 1 (with reference to yeast) is thus obtained along the coordinates of the mentioned features; an highly connected region can be identified as the main cluster, which might be interpreted as a well compartmentalized structure with well-defined functionality and coordinated activity.

Achieving suitable interactome dimensions is necessary to obtain good algorithmic per-

---

<sup>3</sup>The degree of a node is the number of connections it has to other nodes and the degree distribution is the probability distribution of the degrees over the whole network. The clustering coefficient is given by the proportion of links between the surrounding vertices, divided by the number of links that could possibly exist between them.

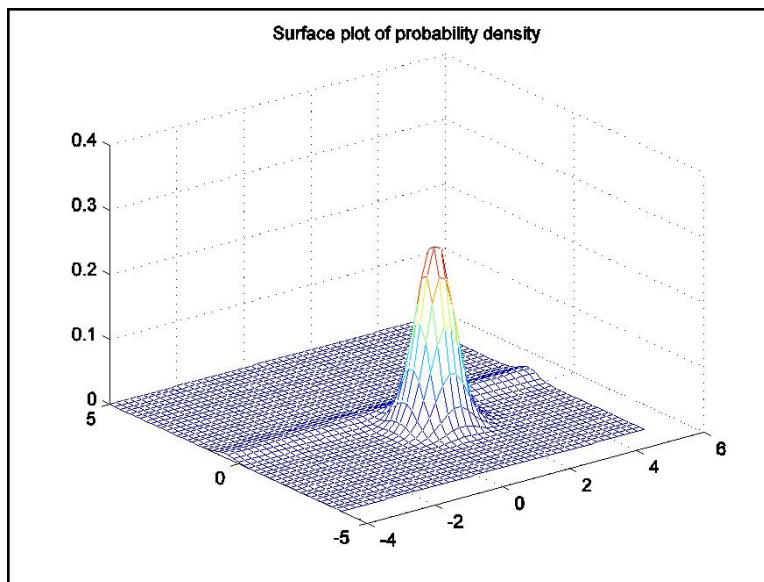


Figure 1: Connectivity map identified through selected features, i.e. degree and clustering coefficient. Yeast dataset.

formance. An interactome can be suitably represented by the AM, which in case of an undirected graph is binary and symmetric. By acting directly on the AM one can enforce dimensionality reduction, thus exploiting its inherently sparse structure.

Sparsity can be naturally justified in a biological framework; among the possible reasons, only a relatively small fraction of interacting protein pairs appears over all possible protein links, including many false positive interactions (typical of yeast two-hybrid - Y2H).

## 4.2 What statistical model?

Several data types can be assumed to lie on  $k$ -dimensional manifolds, say  $M$  with dimension  $k$ , where the latter indicates the number of independent parameters that explain the data. A smooth embedding  $m : M \rightarrow R^d$  maps the data into a feature space of  $d$  dimensions, where fixing a distance induces a metric on  $M$ .

When the latter is also endowed with a probability measure  $P$ , a suitable data representation is possible with  $x = m(\Psi) + e$ , where  $\Psi \sim P$  and  $e \sim N(0, \sigma)$ . Consequently, a finite set of data  $x_j$ ,  $j = 1, \dots, n$  is observed, and we would like to retrieve a corresponding data generating sequence  $z_j = m(\Psi_j)$  of the same length on some sub-manifold of  $M$ .

Noise will prevent this from happening with accuracy, and only an approximate sequence can be found, i.e.  $y_j \approx x_j$ . Then, inference on  $P$  depends on a sample, from which the assignment of probabilities, scores, likelihoods, or more refined parameterizations for possible sub-manifolds might lead to the design of a model framework with a certain accuracy, depending on the knowledge we have about the parameters.

Inferring the global protein network structure in terms of separated components calls for local manifold learning algorithms, where *LVM* can perform quite efficiently. Then, through the approximation of their covariance structure by principal modes or eigenvectors, the goal of approaching the intrinsic manifold dimensionality can also be achieved.

Within each PPIN’s sub-manifold, we usually find hierarchic structures, such as clusters, which might be marginally interconnected. Thus, threshold distances should be defined so to capture within- and between-clusters dependencies. However, clusters can only be instrumental for determining the ability to perform identification and reconstruction tasks; due to their heterogeneity, no universal way to define and extract them is yet available.

As more refined methodological solutions are needed, we consider a novel approach that we name *Network Mixture Models* (NMM). It is an integrative methodology only marginally explored in PPIN, but very flexible and capable to progress with regard to network dimensionality reduction and modularity discovery.

### 4.3 What Statistical Mixtures?

A probabilistic model  $M$  can be defined by an appropriate finite mixture of unimodal densities, in the spirit of unsupervised learning, and also non-Bayesian probabilistic clustering (to be compared with the classical *K-means* algorithm, Hartigan and Wong, 1979).

However, mixtures of distributions are more comprehensive than clustering, even in a network setting. In fact, they approximate densities of high dimensional data that lie near or on a low dimensional manifold, and to various approximation degree. An example is offered by mixture model which grow depending on the number of mixture components  $k$ . The latter is a naturally built-in sieve parameter (Lindsay, 1995) through which a control of both approximation depth and estimation accuracy can be established.

We start with parametric cases, in particular Gaussian and Bernoulli mixtures (*GMM* and *BMM*, respectively). Then, the distributional assumptions can be relaxed by looking first at Probabilistic PCA (*pPCA*) and then at Mixtures of Factor Analyzers (*MixFA*) (see Tipping and Bishop, 1999; Ghahramani and Beal (2000), and McLachlan et al., (2003), respectively). It turns out that for all these methods, the optimization strategies are quite similar and of relatively comparable performances.

GMM (Everitt and Hand, 1981; McLachlan and Basford, 1988; Redner and Walker, 1984; Titterton et al, 1985) are classically employed to perform clustering and density estimation (see Figure 2 for a general idea). Maximum likelihood (ML) is used to estimate the parameters of the different latent sub-populations. A cluster is associated with each of the component distributions, but GMM may not be appropriate for binary or integer data. For this reason, we explore *BMM*.

For a general dataset  $x = [x_1, \dots, x_n]$  and a parameter set  $\Theta$ , i.e. a  $k$ -component finite mixture distribution is described by:

$$p(x | \Theta) = \sum_{i=1}^k \alpha_i p(x | \theta_i) \tag{1}$$

where  $\Theta = [\theta_1, \dots, \theta_k, \alpha_1, \dots, \alpha_k]$ , with  $\alpha_i \geq 0$ ,  $i = 1, \dots, k$ ,  $\sum_{i=1}^k \alpha_i = 1$ . Under the assumption of independently and identically distributed samples  $X$ , the log-likelihood function is:

$$\log p(x | \Theta) = \log \prod_{j=1}^n p[x^{(j)} | \Theta] = \sum_{j=1}^n \log \sum_{i=1}^k \alpha_i p[x^{(j)} | \theta_i] \tag{2}$$



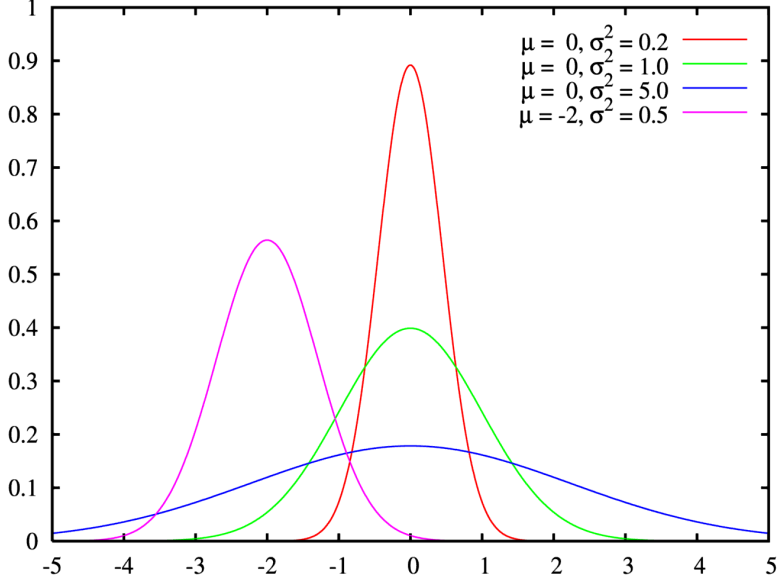


Figure 2: Gaussian Mixture.

By assuming Gaussian densities, the following specification for mixture density and likelihood, respectively, are achieved:

$$p(x | \Theta) = \frac{1}{k} (2\pi\sigma^2)^{-\frac{d}{2}} \sum_{i=1}^k \exp\left[-\frac{1}{2\sigma^2} \|x - \mu_i\|^2\right] \quad (3)$$

$$L(x, \Theta, \sigma^2) = \sum_{j=1}^n \log \sum_{i=1}^k \exp\left[-\frac{1}{2\sigma^2} \|x - \mu_i\|^2\right] \quad (4)$$

It is well-known that neither the ML estimate  $\hat{\Theta}_{ML} = \arg \max_{\theta} [\log p(x | \Theta)]$  nor the maximum a posteriori (MAP) estimate  $\hat{\Theta}_{MAP} = \arg \max_{\theta} [\log p(x | \Theta) + \log p(\Theta)]$  are analytically given.

The classical solution calls for treating a vector associated to the sample, i.e.  $Z = [z^{(1)}, \dots, z^{(n)}]$  as missing data or latent indicator variable that specifies which mixture component a certain data point comes from. Thus, its elements form a binary vector,  $z^{(l)} = [z_1^{(l)}, \dots, z_k^{(l)}]$ , which identifies whether a certain component produces a given sample.

The key point is that if we knew the values of  $Z$ , then we could straightforwardly optimize the complete data log-likelihood with respect to  $\Theta$ , while we need to estimate it when  $Z$  is unknown. Consequently, an expected data log-likelihood is optimized, such as  $Q(\Theta, \Theta^-) = E \sum_{i=1, n} \log p(x_n, z_n | \theta)$ , where the expectation is taken with respect to the previous estimates  $\Theta^-$ , but the function is now optimized with respect to the new parameters  $\Theta$ . The estimate of  $z$  has a recursive nature, due to the dependence of  $z$  on  $\theta$ .

In general, the *Expectation Maximization* (EM) algorithm (Dempster et al, 1977) proceeds by alternating till convergence E-step and M-step. The E-step computes an objective function, say  $Q$ , as the expectation of a complete data log-likelihood over the joint distribution of the unobservable data given the observed data by using the current parameter estimates.

The M-step updates the parameter estimates based on the optimization of  $Q$  at the previous step.

The usual choice is running the EM algorithm iteratively till convergence according to simple steps:

- Set initial  $\theta$ ;
- Compute log-likelihood  $L(\theta)$  till convergence as from:
  1. *E-step*: calculate  $p(z_n | x_n, \theta^-)$  for each  $n$ ;
  2. *M-step*: calculate  $\theta^+ = \arg \max_{\theta} Q(\theta, \theta^-)$ ,  
for  $Q(\Theta, \Theta^-) = E \sum_{i=1, n} \log p(x_n, z_n | \theta)$ ;
  3. Calculate the log-likelihood  $L(\theta) = \log \sum_n \sum_{z_n} p(z_n, x_n | \theta)$

In summary, the EM is usually employed to find the MLE for models with latent variables and make a soft assignment based on posterior probabilities, while *K-means* estimates only means and not covariances.

## 4.4 Network Mixture Models

Networks offer a natural view of inference with latent structure; thus, a new class of network mixture models (*NMM*) can be described. Consider a set of hidden graphs  $G_h = [G_{h_1}, \dots, G_{h_n}]$  mixed or combined to generate the observed noisy graph  $G_o$ .

In a sampled model, where the PPIN has been obtained by extracting nodes and links from a certain annotated database, then each sample  $G_s$  represents up to some degree an approximation to the true graph  $G_t$ , i.e.  $G_s \approx G_t$ . How far  $G_o$  is from  $G_t$  depends on both noise (measurement precision) and approximation accuracy (coverage level).

Instead, in a generative model one could aggregate networks constructed from different sources, or otherwise integrate various specifically designed datasets. This ensemble idea is having practical impact in data integration applications, when merging informatively different data warehouses. In theoretical terms, this matter offers reason for analysis through the lens of superstatistics, i.e. the study of non-linear and non-equilibrium systems (Beck and Cohen, 2003).

An ensemble view of random networks dynamics with the average connectivity assumed to be fluctuating according to an unknown distribution, is usually linked to the identification of an hidden variable distribution (*HVD*). As a result, a network with unspecified degree distribution can be a superposition of networks with assigned degree distribution, for instance random networks (Poisson type for large  $N$ ) whose degree distribution is:

$$p(k) = \int_0^{\infty} \Pi(\lambda) \frac{\lambda^k e^{-\lambda}}{k!} d\lambda \quad (5)$$

References (Abe and Thurner, 2005; Thurner and Biely, 2006; Thurner et al, 2007) have reported that in such case a power law for  $\Pi(\lambda)$  leads to a  $q$ -exponential degree distribution. The  $q$ -exponential distributions are a special case of the type-II generalized Pareto distribution, and are applied to systems with long-range interactions in order to model the distribution of many heavy-tailed phenomena in complex systems.

The key point is thus the existence of an *HVD* linking a general complex network to a random graph, and corresponding to a varying probability of node connectivity. And for scale-free networks, there must be an associated *HVD* which decays as a power law. A marginalization is obtained as:

$$p(k) = \int_0^\infty \Pi(\lambda)p(k | \lambda)d\lambda \quad (6)$$

Actually, a wide class of complex networks can be derived from a fluctuating random graph, with variability depending on the form of the *HVD*. Conversely, given the one-to-one relation between the linking probability behind the interaction dynamics and the resulting degree distribution, an inverse problem has to be faced, i.e. how to recover  $\Pi(\lambda)$  as a function of the observed degree distribution<sup>4</sup>.

It has been noticed that the degree distribution in random networks could also be associated to a Gaussian  $P_G(k)$ , which can also be obtained from a Binomial law in the limit, and with  $\Pi(\mu, \sigma)$  to be found. Statistical mixtures have many possible characterizations, and we explore only a few of them. Another possibility involves the application of multiscale analysis, when it holds that  $P_G = \psi(\frac{k-\mu}{\sigma})$ , i.e. a wavelet-like transform becomes interesting for investigation (Hasegawa, 2006; recent advances in multiscale networks have been proposed also by Marras and Capobianco, 2009).

## 5 Results

### 5.1 Datasets

Our reference HPI dataset (Ulitsky et al, 2008) (<http://acgt.cs.tau.ac.il/clean>) is a fusion of various source data (*BIND*, *HPRD*, *BioGRID*) with interactions detected from small-scale experiments. The total number of 7,385 nodes and 23,462 links are thus available for analysis and modelling.

Some preliminary tests have been also carried out on a model organism such as yeast, and by using the dataset built by (Bader et al, 2003), a combination of protein networks constructed from curated Y2H and Co-Immunoprecipitation data including a set of 5787 high-confidence high-throughput interactions.

### 5.2 PPIN Mixtures

A class of *NMM* is based on (see Newman and Leicht, 2007) the assignment of directed network vertices to groups  $r$  which are unknown a priori, i.e. missing data  $z$  that need to be inferred from the observed data, and a generalization to the case of undirected networks has also been derived. The stochastic characterization of the models allows to specify parameters which define the probability to find groups, and then maximize the likelihood to fit the data.

Thus, model parameters can be set such that  $\pi_r$  is the probability that a vertex is in group  $r$ , and  $\theta_{r,i}$  is the probability that a link from any vertex in group  $r$  connects to vertex  $i$ , given the observed data  $a_{i,j}$  (either 1 or 0, as from the AM) and the missing data  $z_i$ . The

---

<sup>4</sup>Series expansions (such as Laguerre) can be provided to solve the problem in a general way.

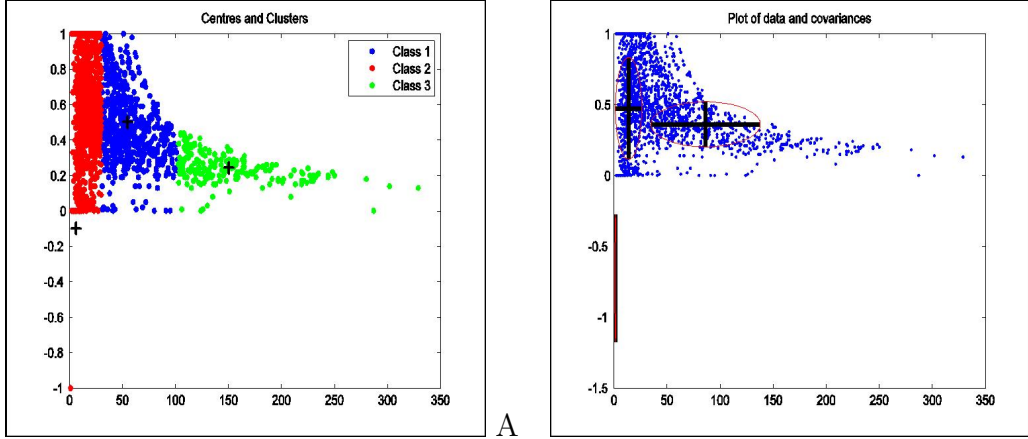


Figure 3: Extracted groups (A) and their covariances. Features used are degree and clustering coefficient.

log-likelihood  $L = \ln p(a, z | \pi, \theta) = \sum_i [\ln \pi_{z_i} + \sum_j a_{i,j} \ln \theta_{z_i,j}]$  with respect to  $\pi$  and  $\theta$  can then be maximized.

From the previous example, a semiparametric likelihood is obtained, as  $z$  is unknown but can be averaged out. This because  $d_{i,r} = p(z_i = r | a, \pi, \theta)$  inserted into the likelihood function leads to  $\hat{L} = \sum_{i,j} d_{i,r} [\ln \pi_r + \sum_j a_{i,j} \ln \theta_{r,j}]$ . By iterative estimation via EM, a solution is achieved for the best possible model parameter estimates.

An application of the above scheme in terms of *GMM*, which is reported for the PPIN yeast data set (constructed by Bader et al, 2003), which consists of 3632 nodes and 22500 interactions. Figure 3 emphasizes three groups depending on the features which have been selected, i.e. degree and clustering coefficient, together with the associated covariances. Such solutions can be only in part satisfactory, as neither modules nor complexes information is delivered.

Nevertheless, *GMM* are just one of the possible choices to consider, and other parametric distributions can be chosen (see below) or extra covariate information can be incorporated to allow for more flexibility in the model selection with regard to the number of effectively needed components (sparsity versus redundancy aspects).

### 5.3 Sampling sub-interactomes

PPIN are undirected networks with a connectivity structure represented by either a binary or a weighted adjacency matrix, usually sparse. Working with the former type of AM is particularly hard. As a result, extracting sub-interactomes becomes an even more difficult task that remains to be done up to a certain acceptable approximation degree to guarantee a certain homogeneity and density level within each group.

Two useful pre-conditions for a good AM decomposition method should be 1. assigning a score function to all possible sub-matrices, and 2. restricting the search space. We have condensed the two aspects in a bi-clustering algorithm, *BicBin* (van Uiter et al, 2008), which we found efficient with binary data, and effective because converging in a few iterations. *BicBin* returns for a binary matrix the top scoring bi-cluster built from sequential associations of

row and column elements in sets (when the score exceed certain thresholds, then it converges quite quickly).

Through a localized pattern detection approach, a bi-cluster is composed by objects with similarity over only a subset of attributes. In general, given a data matrix A, one identifies a set of bi-clusters B such that each one satisfies characteristics of homogeneity. Some accuracy is lost via this form of sub-sampling, relatively to the global interactome, but we make some corrections as explained below.

The general idea of extracting informative samples from high-dimensional datasets refers to the Johnson-Lindenstrauss Lemma (Johnson and Lindenstrauss, 1984), stating that an embedding can be found for a dataset of  $n$  points in a subspace of dimension  $O(\log n)$  with quite little distortion on the pair-wise distances, and this embedding can be simple just like picking at random a certain subspace where to project all points.

Similarly, the pre-processing step that we implement works through bi-clustering applied to the AM for extracting lower-dimensional data, or identifying a sub-space (sub-interactome) of lower dimensions. Then, mixture modelling is applied to the extracted sub-interactome, and compared to the same application over a random set.

Other sampling methods could be used to select and extract AM blocks so to get sub-interactomes for then testing inference methods. But the strength of our approach remains that combining sub-sampling with EM-based clustering or mixture modelling finds structure in high-dimensional data.

Against the typical instability of sampling, we have two possible choices: A. build an ensemble framework through replicates or multiple runs of the algorithm (resampling path); B. adopt scoring rules (optimization path). In summary, while we are currently designing techniques for the first scenario, we show below the latter approach and adopt the following bi-clustering steps:

- **Search Phase** - *Search the complete binary AM for sub-matrices*
- **Scoring Phase** - *Assign a score (computed coordinate-wise, i.e. from row and column elements) to all possible sub-matrices*
- **Selection Phase** - *Choose the best (top scoring, in terms of algorithmic not biological aspects) one.*

In protein maps we look at the denser regions (or in AM, where more 1s are) to find higher interaction activity. This densification strategy is thus induced by our approach.

We have extended from the *BicBin*'s method, due to the need to adapt the final bi-cluster to the real meaning of AM. Therefore, we have changed the output from the extracted top scoring sub-interactome when it is not squared. In such case we square it, and we call this step its closure because we keep the union of each element uniquely detected along the two coordinates, thus  $[\forall c_i \in C^I] \cup [\forall c_j \in C^J]$ , where  $c_i = [a_{ij}]_i$ , and  $c_j = [a_{ij}]_j$ , and  $c_i, i = 1, \dots, I$  and  $c_j, j = 1, \dots, J$  are elements of the extracted bi-cluster, and  $C^I, C^J$  are respectively the selected row and column bi-cluster blocks.

We are conservative in our selection criterion, because we keep all the detected nodes and their interactions. Otherwise, by choosing the intersection of the above sets we would

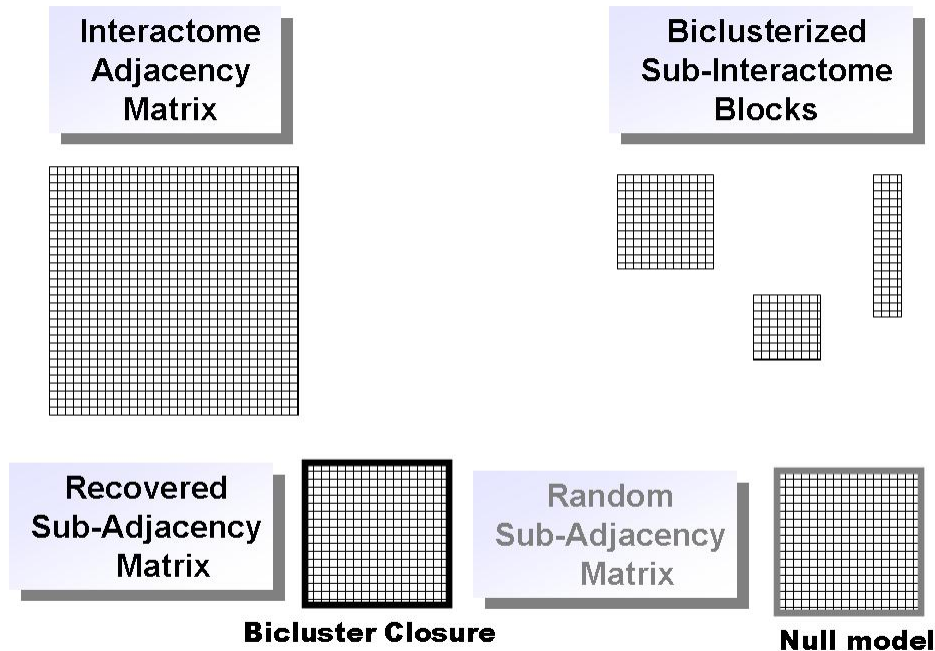


Figure 4: Sub-interactome extraction: step-by-step procedure.

limit the sub-interactome to a restricted region at the price of also being missing network structure compared to the case of keeping the bi-cluster as it is.

In other words, the AM has on both rows and columns the labels which refer to the same elements, i.e. the node lists. We add labels which have been excluded in one dimension but included in the other so to restore an AM squared block (see Figure 4). Instead, the row-column blocks intersection would allow to consider less interactions, likewise a non-squared bi-cluster.

The second adaptation is the construction of a corresponding random sub-interactome of the same size of the closed bi-cluster, which represents a null model through which mixture network models performance can be measured.

## 5.4 Deterministic vs Probabilistic Graph Mining Methods

Cluster or community detection methods are aimed to reveal network modularity features and usually employ partitioning techniques condensing strongly connected nodes and separating sparsely connected ones. However, finding the optimal partition in networks is impossible with large network dimensions.

Approximate solutions are obtained by decomposition methods which shrink the network by assigning links to a group and removing them from successive search, synthesis methods which allow recursively sub-network growth, and spectral methods centered on Laplacian matrix building.

We describe next two popular algorithms which offer a connectivity-based modularity view of PPIN and work by measuring the cliquishness of the node neighborhood. They

adopt different stopping criteria to avoid the risk of myopic greedy generation of oversized groups. More importantly, they do not have any probabilistic property.

### 5.4.1 MCODE

The precision of matching extracted clusters to well established protein complexes is generally biased by two main factors: incompleteness of PPINs, and the fact that protein complexes may not be complete subgraphs.

MCODE exploits local graph density to match to protein complexes explores some locally dense regions of a graph computed from a clustering coefficient, i.e.  $C_i = \frac{2n}{k_i(k_i-1)}$ , where  $k_i$  is the node size of the neighborhood of node  $i$ , and  $n$  is the number of edges in the neighborhood.

The  $k$ -core is the structure that one finds with MCODE in a graph; it is a network of minimal degree  $k$  defined as the remaining sub-graph after that all the nodes with degrees  $1 - k$  have been removed successively<sup>5</sup>. In other terms, given  $G = (V, E)$ , the  $k$ -core is computed by pruning all the  $V$  (with their  $E$ ) with degree less than  $k$  until all nodes in the remaining network have at least degree  $k$

Then, if a node  $\in k$ -core but  $\notin (1 + k)$ -core of the graph, it has coreness degree  $k$ . The highest  $k$ -core of a network is the central most densely connected sub-network. Thus, after vertex weighting, complex prediction is conducted where the relevance of each cluster is validated against known complexes or functional modules, and final statistics are computed about clusters size, density and functional homogeneity.

For the HPI we examine, the main groups we identified are reported in Table 1.

<b>MCODE</b>					
<i>2-core:</i>	129	<i>3-core:</i>	35	<i>4-core:</i>	4
		<i>5-core:</i>	4	<i>6-core:</i>	5
		<i>7-core:</i>	4		
		<i>8-core:</i>	1	<i>9-core:</i>	1
		<i>10-core:</i>	1		
<b>CFinder</b>					
<i>3-com:</i>	380	<i>4-com:</i>	171	<i>5-com:</i>	71
		<i>6-com:</i>	22	<i>7-com:</i>	10
		<i>8-com:</i>	4		
		<i>9-com:</i>	3		

Table 1: Extracted Groups: k-cores and communities.

### 5.4.2 CFinder

A community in a network is a group of nodes more densely connected to each other than to nodes outside the group. In real networks communities often overlap. The Clique Percolation Method (CPM) is the built-in algorithm in CFinder designed to locate the  $k$ -clique communities.

Typically, a member in a community is not necessarily linked to all other nodes, but only to some. The groups thus formed generate a community which is a union of complete (fully

<sup>5</sup>The procedure is as follows: (A) when a node is removed, all its adjacent edges will also be removed; (B) after a node of degree  $\leq 1 - k$  is removed, in the remaining graph all the remaining nodes with a new degree  $\leq 1 - k$  also need to be removed.

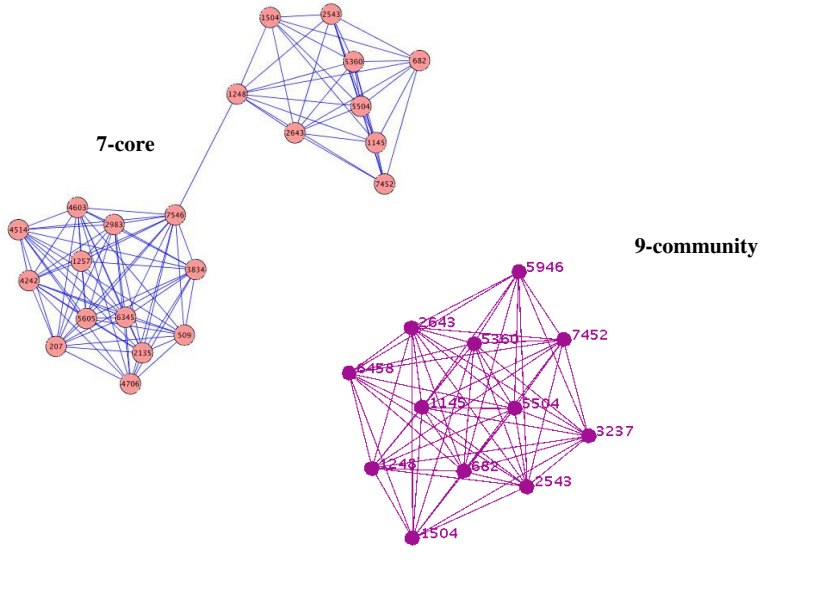


Figure 5: Example of 7-core from MCODE, and 9-community from CFinder.

connected) subgraphs, and a  $k$ -clique-community is the union of all  $k$ -cliques that can be reached from each other through a series of adjacent  $k$ -cliques.

The HPI community-based main groups we have identified are also reported in Table 1. Figure 5 offer examples of graphical representation of a core and a community.

## 5.5 Probabilistic PCA

A probabilistic formulation of PCA ( $pPCA$ ) was proposed by (Tipping and Bishop, 1999) by viewing it as an  $LVM$  in which the  $d$ -dimensional observed data vector  $x_n, n = 1, \dots, N$ , can be described in terms of an  $m$ -dimensional unobserved vector  $z_n$  and a noise term  $\epsilon$ ,  $x_n = Az_n + \epsilon$ , where  $A$  is a matrix ( $m < d$ ) and  $\epsilon$  is a multivariate Gaussian independently distributed with a diagonal covariance matrix  $\sigma^2 I$ . Then:

$$p(x_n | A, z_n) = \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \exp\left[-\frac{1}{2\sigma^2} (x_n - Ay_n)^T (x_n - Ay_n)\right] \quad (7)$$

The EM algorithm (with a prior over  $y_n$ ) can find ML estimates of both  $A$  and  $\sigma$ . For mixtures of  $pPCA$  the same algorithm finds an estimate for the means  $\mu_i$  of each component, and an estimate for  $p(x_n)$  again for each component jointly with the prior probability of each model itself. However, from the results of our application (Figure 6), the model selection step appears hard, due likely to an underlying Gaussianity assumption too weak to discriminate among clusters.

The log-likelihood convergence paths (Figure 6, plot C) show that by augmenting the



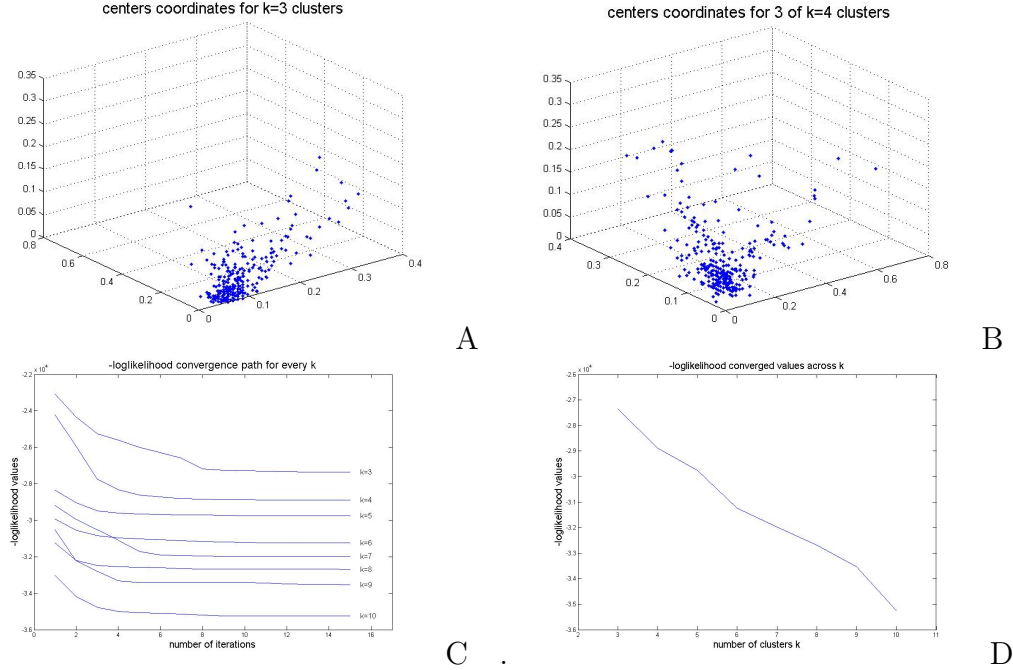


Figure 6: Scatter of three PCs (A) and of three out of four (B). Corresponding likelihood convergence paths.

number of clusters less iterations are needed for convergence. In Figure 6 (plot D.), the more clusters we have and the better the likelihood behaves, a clear overfitting signature. The problem is of course how do we choose the number of mixture components, as the ML tends to favor the largest possible value of  $k$  (i.e. a big number of parameters maximizes the fit to the data), unless some form of penalization is taken into account.

## 5.6 Bernoulli Mixtures

As said, *GMM* fit well real-valued data, but binary data may require different treatment. We consider mixtures of discrete binary variables described by Bernoulli distributions. A product of Bernoulli is given by:

$$p(x | z = k, \theta) = \prod_{i=1}^K B(x_i | \theta_{ki}) = \prod_{i=1}^K x_i^{\theta_{ki}} (1 - x_i)^{1 - \theta_{ki}} \quad (8)$$

By extending to the class of finite mixtures of multivariate Bernoulli distributions (Carreira-Perpinan and Renals, 2000), lack of identifiability is possible since different values of the mixture parameters can correspond to exactly the same probability distribution. However, in practical cases the estimation of this class of mixtures can still produce meaningful results, thus downsizing identifiability problem, and letting the EM algorithm converge to a proper maximum likelihood estimate.

Finite mixtures of multivariate Bernoulli have  $K$  groups of  $N$  variables, where  $k \in K$  is chosen with probability  $\pi_k$ . Thus, while the probabilities are  $p(X/\mu, p) \sum_{k=1, K} \pi_k p(X/\mu, k)$  and  $p(X/\mu, k) = \sum_{i=1, N} \mu k_i (1 - \mu k_i)$ , for  $\mu = [\mu_{k=1, K}]$ ,  $p = [\pi_{k=1, K}]$ ,  $a = x_{i=1, N}$ , the *BMM* log-likelihood is:

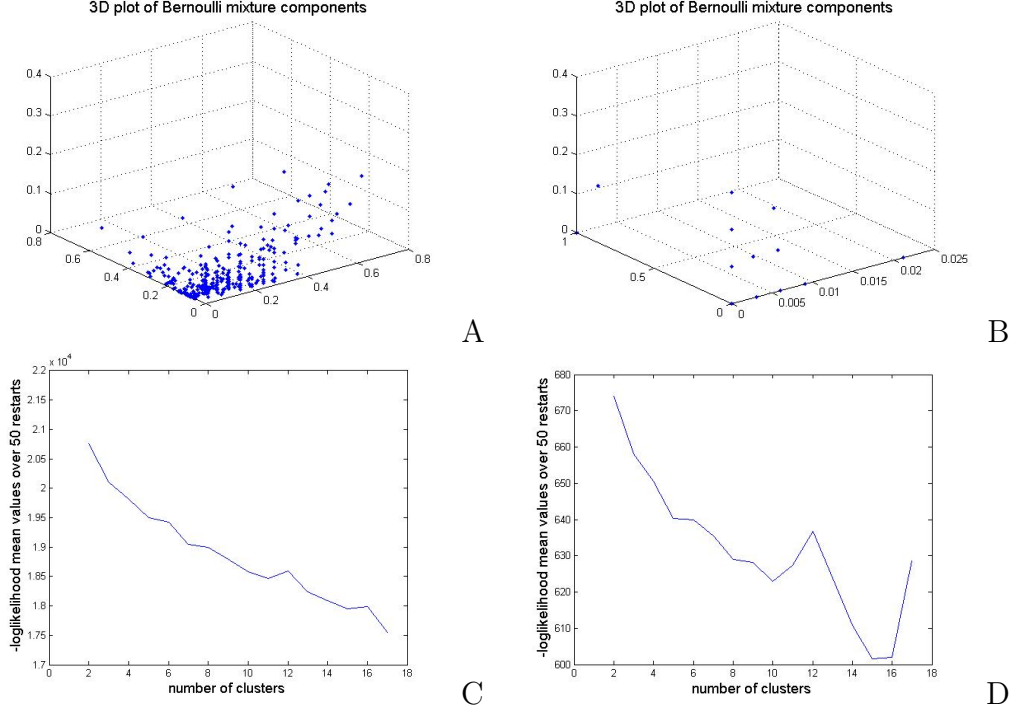


Figure 7: Scatter of three Bernoulli (A) and random case (B). Corresponding likelihood convergence paths

$$\ln p(X/\mu, p) = \sum_{i=1, N} \ln \left[ \sum_{k=1, K} \pi_k p(X_i/\mu_k) \right] \quad (9)$$

A more compact expression is given by  $\ln p(X/\theta) = \ln \sum_z p(X, Z/\theta)$ , where  $X$  are the observed data,  $Z$  the latent variables,  $\theta$  the parameters (log-likelihood in terms of sum rule of probability).

For *BMM* too we look at likelihood-based model selection. The number of clusters we find is still not small; for three clusters we estimate weights (referred to the plotted example) given by  $\pi = [0.0645, 0.3147, 0.6208]$ . Again we notice overfitting (Figure 7, plot C.), despite a clear differentiation from the patterns observed in the random case.

## 5.7 Mixtures of Factor Analyzers

With *MixFA* models (Ghahramani and Beal, 2000; MLachlan et al, 2003), a link between observed high dimensional data  $x$  and their lower dimensional manifold  $z$  is obtained via some discrete hidden states of the manifold itself  $y$ . *MixFA* parameterize a joint distribution over both observed and latent variables:

$$p(x, y, z_y) = p(x | y, z_y) p(z_y | y) p(y) \quad (10)$$

The underlying assumption is that the samples belong with some prior probabilities  $p(y)$  to different manifold's regions where the data are Normally distributed. Then, the link between high and low dimensional coordinates is linearly established through:

$$p(x | y, z_y) = \kappa \exp[-\frac{1}{2}(x - \mu_y - \sigma_y z_y)^T \delta_s^{-1} (x - \mu_y - \sigma_y z_y)] \quad (11)$$

for  $\kappa = |2\pi\delta_s|^{-\frac{1}{2}}$ , and the marginal data distribution  $p(x)$  is obtained via a Gaussian mixture such as:

$$p(x) = \sum_y p_y |2\pi(\sigma_y\sigma_y^T + \delta_y)|^{-\frac{1}{2}} \exp[-\frac{1}{2}(x - \mu_y)^T (\sigma_y\sigma_y^T + \delta_y)^{-1} (x - \mu_y)] \quad (12)$$

In this unsupervised mixture model the parameter set is  $\Theta = [\mu_y, \sigma_y, \delta_y, p(y)]$ , and is estimated by EM. In discrete terms, *MixFA* are latent variable models where the data  $x$  are generated by a linear transformation of Normally distributed factor scores  $z_y$  plus some noise.

The model is thus  $x = \Omega z + e$  where  $z \sim N(0, \sigma)$ , and  $e \sim N(0, \delta)$ , with the  $\Omega$  matrix containing the factor loadings. The observed variables  $x$  are conditionally independent given  $z$ . Note that Factor Analyzer subspaces from  $\Omega$  do not correspond to PCA subspaces, unless isotropic errors are considered  $\sigma = sI$ .

PPINs can have heterogeneous degree distribution, and a likelihood available through an *NMM* assigns a significant measure to the network itself. By adding information on localization through the manifold's regions, we may better understand whether the resulting distributions are more or less separable, and what features characterize them.

In Figure 8, *MixFA* statistical learning suggests two main things: **1.** with three (plot A.) or four (plot B.) factors, a good contrast is obtained with respect to the null model (plot C.); **2.** a likely strong random effect occurs when more than 6-7 clusters are considered (see plots D. and E.).

Overall, some uncertainty remains about the best possible choice in terms of optimal number of detected components. However, the fact of splitting the data into more homogeneous groups may for sure help in terms of prediction accuracy and confidence measures based on localized interaction dynamics.

Ongoing work is looking at two possible methodological extensions: towards non-parametric characterization (i.e. kernel density estimators, penalized likelihoods, AIC, BIC, etc.) of the *NMM* estimation problem, and towards comparative performance of AM sub-sampling methods in order to assign robustness and stability to the extracted patterns. Once *NMM* are tuned, it will make sense to transfer their superior detection power in biological terms.

## 6 Conclusions and future directions

Graph mining through coreness- and community-based algorithms is very useful at a global interactomic scale, but present limitations when localization is sought. In such context, probabilistic algorithms are shown to be effective in extracting information from sub-interactomes, despite a strong dependence on the selected statistical model (and overfitting) when parametric assumptions are retained.

We believe that the two approaches can in principle work together, in complementary way. Deterministic-then-probabilistic graph mining could be useful for establishing a coarse decomposition first, followed by a more localized and finer-grid analysis within a probabilistic

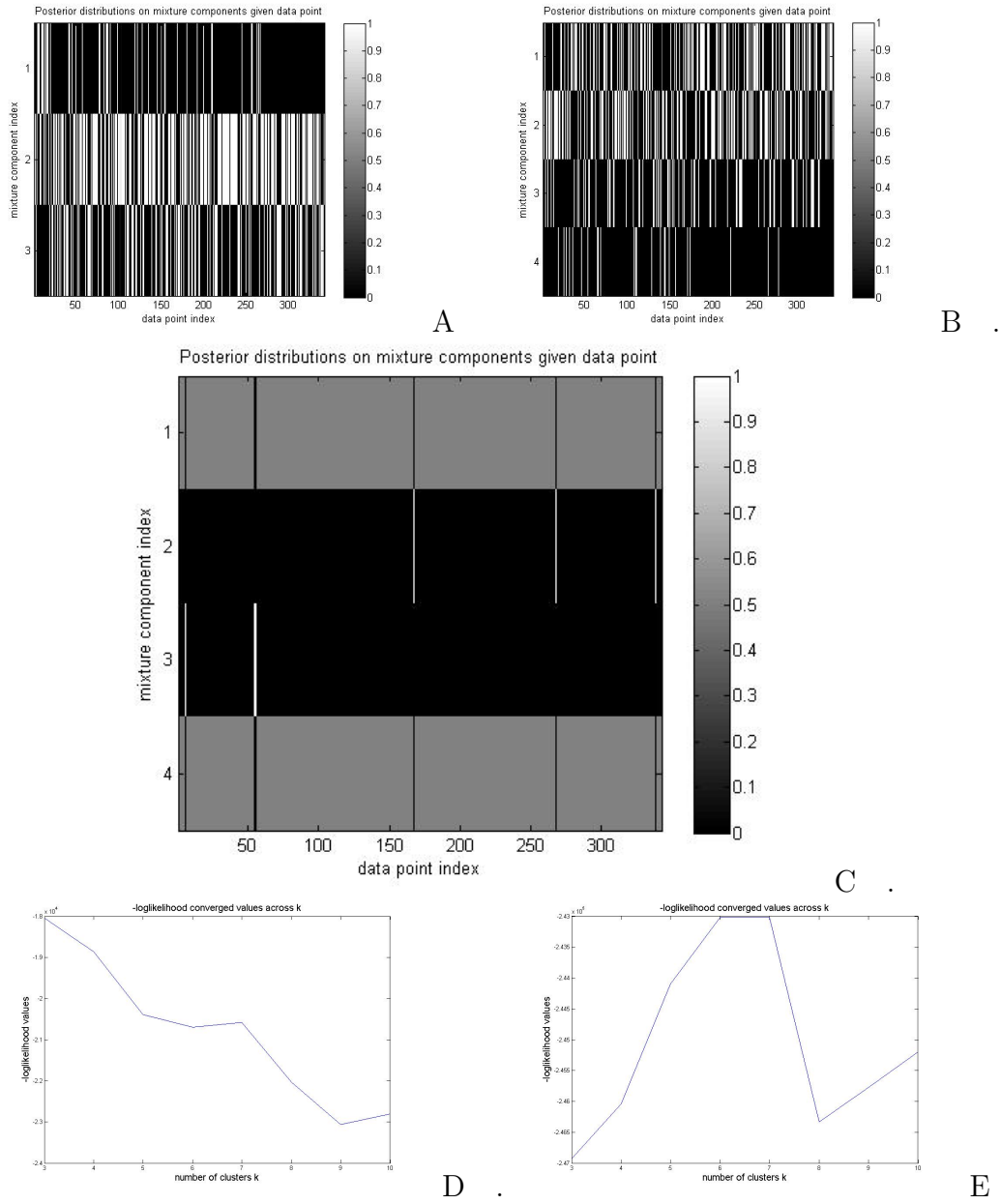


Figure 8: Three (A) or four (B) FA in biclusterized sub-interactomes. Random case (C) with four FA. Likelihood convergence paths at the bottom.

framework. We envisage a great utility of these hybrid approaches when for instance protein relationships within an extracted module need to be further characterized in statistical way, thus accounting for uncertainty degrees.

We aim to compare different sub-sampling techniques and calibrate their scorings or block selection criteria. An important milestone involves measuring the robustness of block extraction for validation scopes relatively to community/coreness-based subgraphs and against null models. For the latter, in particular, the methods we are looking at suggest ways to construct a null model which might be differentiated, possibly stratified. In turn, this result might help with disease-based interactomes.

## References

- [1] Abe S. & Thurner S. 2005 Complex networks emerging from fluctuating random graphs: analytic formula for the hidden variable distribution. *Phys. Rev. E* 72, 036102-4.
- [2] Bader J.S., Chaudhuri A., Rothberg J.M. & Chant J. 2003 Gaining Confidence in High-throughput protein interaction networks. *Nat. Biotech.* 22(1), 78-85.
- [3] Bartholomew D.J. 1987 Latent variable models and factor analysis. New York: Oxford University Press.
- [4] Beck C. & Cohen E.G.D. 2003 Superstatistics. *Phys. A* 322, 267-275.
- [5] Carreira-Perpinan M.A. & Renals S. 2000 Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Computat.* 12, 1411-52.
- [6] Chaurasia G., Iqbal Y., Hnig C., Herzel H., Wanker E.E., & Futschik M.E. 2007 UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res.* 35 Datab. issue D590-4.
- [7] Deane C.M., Salwinski L., Xenarios I., & Eisenberg D. 2002 Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteom.* 1, 3493-56.
- [8] Dempster A., Laird N., & Rubin D. 1977 Likelihood from incomplete data via the EM algorithm. *J Royal Statist. Soc. B* 39(1), 138.
- [9] Edwards A.M., Kus B., Jansen R., Greenbaum D., Greenblatt J., & Gerstein M. 2002 Bridging structural biology and genomics: assessing protein interaction data with known complexes *Trends Genet.* 18, 529-36.
- [10] Everitt B.S., & Hand D.J. 1981 Finite mixture distributions. London: Chapman and Hall.
- [11] Ewing R.M., Chu P., Elisma F., Li H., Taylor P., Climie S., McBroom-Cerajewski L., Robinson M.D., O'Connor L., Li M., Taylor R., Dharsee M., Ho Y., Heilbut A., Moore L., Zhang S., Ornatsky O., Bukhman Y.V., Ethier M., Sheng Y., Vasilescu J., Abu-Farha M., Lambert J.P., Duewel H.S., Stewart I.I., Kuehl B., Hogue K., Colwill K., Gladwish K., Muskat B., Kinach R., Adams S.L., Moran M.F., Morin G.B., Topaloglou T., & Figeys D. 2007 Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* 3(89), 1-17.

- [12] Gavin A.C., Aloy P., Grandi P., Krause R., Boesche M., Marzioch M., Rau C., Jensen L.J., Bastuck S., Dmpelfeld B., Edelmann A., Heurtier M.A., Hoffman V., Hoefert C., Klein K., Hudak M., Michon A.M., Schelder M., Schirle M., Remor M., Rudi T., Hooper S., Bauer A., Bouwmeester T., Casari G., Drewes G., Neubauer G., Rick J.M., Kuster B., Bork P., Russell R.B., & Superti-Furga G. 2006 Proteome survey reveals modularity of the yeast cell machinery. *Nat.* 440, 631-636.
- [13] Ghahramani Z., & Beal M.J. 2000 Variational Inference for Bayesian Mixtures of Factor Analyzers, in *Advances in Neural Information Processing Systems* (ed. S.A. Solla, T.K. Leen & K. Miller), 12, pp. 449-455. MIT Press.
- [14] Goodman L.A., 1974 Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61(2), 215-231.
- [15] Hart G.T., Ramani A.K., & Marcotte E.M. 2006 How complete are current yeast and human protein-interaction networks? *Gen. Biol.* 7(11), 120.1-120.9.
- [16] Hartigan J.A., & Wong M.A. 1979 A K-Means Clustering Algorithm. *Appl. Statist.* 28(1), 100-108.
- [17] Hasegawa H., 2006 Nonextensive aspects of small-world networks. *Phys. A* 365, 383-401.
- [18] Ito T., Chiba T., Ozawa R., Yoshida M., Hattori M., Sakaki Y. 2001 A comprehensive two-hybrid analysis to explore the yeast protein interactome. *PNAS* 98, 4569-4574.
- [19] Jansen R., & Gerstein M. 2004 Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr. Opin. in Microbiol.* 7, 535-545.
- [20] Jansen R., Yu H., Greenbaum D., Kluger Y., Krogan N.J., Chung S., Emili A., Snyder M., Greenblatt J.F., & Gerstein M. 2003 A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644), 449-453.
- [21] Johnson W., & Lindenstrauss J. 1984 Extensions of Lipschitz maps into a Hilbert space. *Contemp. Math.* 26, 189-206.
- [22] Jonsson P.F., & Bates P.A. 2006 Global topological features of cancer proteins in the human interactome. *Bioinform.* 22(18), 2291-2297.
- [23] Krogan N.J., Cagney G., Yu H., Zhong G., Guo X., Ignatchenko A., Li J., Pu S., Datta N., Tikuisis A.P., Punna T., Peregrn-Alvarez J.M., Shales M., Zhang X., Davey M., Robinson M.D., Paccanaro A., Bray J.E., Sheung A., Beattie B., Richards D.P., Canadien V., Lalev A., Mena F., Wong P., Starostine A., Canete M.M., Vlasblom J., Wu S., Orsi C., Collins S.R., Chandran S., Haw R., Rilstone J.J., Gandi K., Thompson N.J., Musso G., St Onge P., Ghanny S., Lam M.H.Y., Butland G., Altaf-Ul A.M., Kanaya S., Shilatifard A., O'Shea E., Weissman J.S., Ingles C.J., Hughes T.R., Parkinson J., Gerstein M., Wodak S.J., Emili A., & Greenblatt J.F. 2006 Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nat.* 440, 637-643.
- [24] Lindsay B. 1995 *Mixture Models: Theory, Geometry and Applications*. Hayward: Inst. Math. Stat.

- [25] Lu L.J., Xia Y., Paccanaro A., Yu H., & Gerstein M. 2005 Assessing the limits of genomic data integration for predicting protein networks. *Gen. Res.* 15, 945-53.
- [26] McLachlan G.J., & Basford K.E. 1988 *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker Inc.
- [27] McLachlan G.J., & Peel D. *Finite Mixture Models*. New York: Wiley & Sons.
- [28] McLachlan G.F., Peel D., & Bean R.W. 2003 Modelling high dimensional data by mixtures of factor analyzers. *Computat. Stat. Data. Anal.* 41, 379-388.
- [29] Marras E., & Capobianco E. 2008 Advances in human protein interactome inference, in *Functional and Operational Statistics*. (eds. S. Dabo-Niang & F. Ferraty) pp. 8994. Heidelberg: Physica-Verlag.
- [30] Marras E., & Capobianco E. 2008 Mining Protein-Protein Interaction Networks: Denoising Effects. *JSTAT*, forthcoming 2009.
- [31] Newman M.E.J. & Leicht E.A. 2007 Mixture models and exploratory analysis in networks. *PNAS* 104(23), 9564-9569.
- [32] Palla G., Dernyi I., Farkas I., & Vicsek T., 2005 Uncovering the overlapping community structure of complex networks in nature an society. *Nat.* 435, 814-8.
- [33] Redner R.A., & Walker H.F. 1984 Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* 26(2), 195-239.
- [34] Rual J.F., Venkatesan K., Hao T., Hirozane-Kishikawa T., Dricot A., Li N., Berriz G.F., Gibbons F.D., Dreze M., Ayivi-Guedehoussou N., Klitgord N., Simon C., Boxem M., Milstein S., Rosenberg J., Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala J., Lim J., Fraughton C., Llamosas E., Cevik S., Bex C., Lamesch P., Sikorski R.S., Vandenhaute J., Zoghbi H.Y., Smolyar A., Bosak S., Sequerra R., Doucette-Stamm L., Cusick M.E., Hill D.E., Roth F.P., & Vidal M. 2005 Towards a proteome-scale map of the human protein-protein interaction network. *Nat.* 437, 11731178.
- [35] Sharan R., Ulitsky I., & Shamir R., 2007 Network-based prediction of protein function. *Mol. Syst. Biol.* 3:88.
- [36] Thurner S., & Biely C., 2006 Two statistical mechanics aspects of complex networks. *Phys. A*, 372, 346-353.
- [37] Thurner S., Kyriakopoulos F., & Tsallis C., 2007 Unified model for network dynamics exhibiting nonextensive statistics. *Phys. Rev. E* 76(2), 3.
- [38] Tipping M.E., & Bishop C.M., 1999 Mixtures of probabilistic principal component analyzers. *Neur. Computat.* 11(2), 443-482.
- [39] Titterton D.M., Smith A.F.M., & Makov U.E., 1985 *Statistical analysis of finite mixture distributions* Chichester: John Wiley & Sons.
- [40] Troyanskaya O.G., Dolinski K., Owen A.B., Altman R.B., Botstein D. 2003 A Bayesian framework for combining heterogeneous data sources for gene function prediction in *Saccharomyces cerevisiae*. *PNAS* 100(14), 8348-53.

- [41] Uetz P., Giot L., Cagney G., Mansfield T., Judson R., Knight J., Lockshon D., Narayan V., Srinivasan M., & Pochart P., 2000 comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nat.* 403, 623-627.
- [42] Ulitsky I., Karp R.M., & Shamir R., 2008 Detecting disease-specific disregulated pathways via analysis of clinical expression profiles. *Proc. RECOMB, Lect. Not. Comp. Sci.*, 4955.
- [43] Vidal M., 2005 Interactome Modeling. *FEBS Let.* 579, 1834-1838.
- [44] van Uitert M., Meuleman W., & Wessels L., 2008 Biclustering sparse binary genomic data. *J. Comput. Biol.*, in press.
- [45] von Mering C., Krause R., Snel B., Cornell M., Oliver S.G., Fields S., & Bork P, 2002 Comparative assessment of large-scale data sets of protein-protein interactions. *Nat.* 417, 399-401.
- [46] Wachi S., Yoneda K., & Wu R., 2005 Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinform.* 21(23), 4205-4208.
- [47] Xu J., & Li Y., 2006 Discovering disease-genes by topological features in human proteinprotein interaction network. *Bioinform.* 22(22), 2800-2805.