Bayesian Mixed-Membership Models of Complex and Evolving Networks

by

EDOARDO MARIA AIROLDI

A dissertation submitted in partial satisfaction

of the requirements for the degree of

Doctor of Philosophy

in

COMPUTER SCIENCE

with designated emphasis in

COMPUTATION, ORGANIZATION & SOCIETY

in the

SCHOOL OF COMPUTER SCIENCE

of

CARNEGIE MELLON UNIVERSITY

December 2006

To my wife Xue and my parents Carlo and Carla.

Acknowledgments

This piece of writing marks the end of a glorious journey and the beginning of a new one. None of this would have been possible without the help and support of many friends and colleagues.

First and foremost, I thank my advisors Stephen E. Fienberg and Kathleen M. Carley for their vision, support and friendship over the years at Carnegie Mellon. Steve and Kathleen have been sources of inspiration as well as exemplary teachers, mentors, and collaborators. I thank my dissertation committee David M. Blei, Christos Faloutsos, and Eric P. Xing for their insightful comments and suggestions. I also thank Joseph B. Kadane, William W. Cohen and Latanya Sweeney for their support and continued advice on many aspects of my research.

I am fortunate to have interacted with a number of superb colleagues, and I would like to recognize their contribution to this work. Annelise Anderson, Michael Ashworth, David Banks, Michael Benisch, David Blei, Deepay Chakrabarti, Taeryon Choi, Donato Michele Cifarelli, Francesco Corielli, James Cronin, George Davis, Robyn Dawes, Massimo Di Rienzo, Adrian Dobra, Elena Eneva, Jason Ernst, Elena Erosheva, Mauro Filippini, Zoubin Ghahramani, Anna Goldenberg, Joel Greenhouse, William “Spike” Gronim, Ralph Gross, Michele Guindani, Stephen Hanneke, Aleks Jakulin, Alan Karr, Robert Kass, Krishna Kumaraswamy, John Lafferty, Ann Lee, Jure Leskovec, Yiheng Li, Xiaodong Lin, Mauro Maggioni, Bradley Malin, Tom Mitchell, Rema Padman, Eugenio Regazzini, Alessandro Rinaldo, Ronald Rosenfeld, Aleksandra Slavkovic, Nicoleta Serban, Teddy Seidenfeld, Cosma Shalizi, Michael Shamos, Kiron Skinner, Ricardo Silva, Peter Spirtes, David

Tolliver, Ottavio Tucci, Leonid Teverovski, Isabella Verdinelli, Piero Veronese, Larry Wasserman, Eric Xing, and Alice Zheng, have all been influential to my research through collaboration, discussion, constructive criticism, and wine. I particularly thank David Blei and Eric Xing who effectively launched the line of work I develop in this thesis; a major portion of the work presented in Chapters 3 and 4 is joint with them, and more is on the way. I greatly enjoy their friendship and the many discussions about research, world domination, and life in general. Alan Karr, David Banks, the National Institute of Social Sciences, and the Statistical (NISS) and Applied Mathematical Sciences Institute (SAMSI) for sponsoring my visits to many of the outstanding scientific activities and meetings they organize that revolve around the statistical, mathematical, and computing sciences. Piero Veronese, Donato Michele Cifarelli, Eugenio Regazzini, and Francesco Corielli for nurturing my passion of mathematical statistics and probability.

I thank all my friends and family for their support and distraction. I especially thank my parents Carlo and Carla Airoldi and my sister Valeria Airoldi who have given me a lifetime of love and care. Finally, I thank my wife Xue Bai for sharing with me the good and the bad, and filling my every day with joy and happiness; her sweetness, humor, and friendship sustain me.

This work was supported in part by the National Institutes of Health under Grant No. R01 AG023141-01, by the Office of Naval Research under Contract No. N00014-06-1-0772, the Army Research Lab Grant No. ARL/CTA DAAD19-01-2-0009, by the National Science Foundation under Grant No. DMS-0240019, and Grant No. IIS-0218466, and the Department of Defense, all to Carnegie Mellon University. Additional support was provided by the center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. Travel grants were provided by the National Institute of Health, the National Science Foundation, the Office of Naval Research, the American Association for Artificial Intelligence, the International Biometric Society, the International Society for Bayesian Analysis, and Microsoft Corporation.

Foreword

Over the past three years I have been working on probabilistic models of multivariate attributes and relations. My work suggests a probabilistic framework and a general modeling approach to complex and evolving networks, which is based on the four concepts of mixed membership, motifs, dynamics and integration. In this thesis, I present such a framework and discuss its properties. In particular, the main goal of the research is to establish the essential elements of formal models of complexity that reconcile the global properties of a system with local phenomena of interest, in a generative fashion.

A solution to the global/local trade-off is to express complexity through hierarchical mixtures of simple patterns, i.e., motifs, that evolve over time. Complex global behavior emerges from the combination of local interaction patterns and their dynamics. I discuss the extent to which this novel framework incorporates, generalizes, and extends other probabilistic approaches present in the literature, and argue that it provides the foundations of a statistical theory of random graphs.

A major part of the effort is devoted to the analysis of modeling issues related to the four essential aspects listed above, in the context of applications to social and biological networks. I also investigate theoretical and computational issues such as the geometrical intuition of the latent allocation task—an important inference objective shared by the various models encompassed by this framework.

Contents

Chapter 1

Introduction

This thesis provides a methodological framework for the statistical analysis of complex graphs and dynamic networks.[1] In it, I develop probabilistic algorithms that generate, evolve and integrate a heterogeneous collection of graphs, I study the statistical models these algorithms implicitly specify, and I develop strategies for estimating the set of quantities on which they depend in the context of applications to social and biological networks.

1.1 Complex Data

My investigations concern a population of objects of study. Objects can be divided into few different categories, or types, e.g., gene, proteins, and stable protein complexes; or documents, words and references; or agents, tasks, and resources. Observations consist of measurements taken on individual objects, i.e., attributes, and on pairs of objects i.e., relations. Both attributes and relations are typically multivariate, e.g., the expression of a gene under many experimental conditions,

[1] The terms *graph* and *network*, without qualifications, are synonyms for the purposes of this thesis because throughout I represent networks in terms of graphs.

or the set of words contained in a document. Measurements are taken over time, and distributed across heterogeneous databases.

At any given epoch, each object is represented as a node in a graph. Relations are represented as edges in the graph, among nodes (i.e., objects) of the same type, whereas attributes are represented as edges in the graphs, among nodes of different types. From a statistical perspective, it is sometimes convenient to consider the *random matrices* corresponding to the various graphs at hand; that is, the adjacency matrices whose elements are scalar random variables that encode edge-weights. This is the perspective I favor throughout this thesis.

Example 1. *Figure 1.1 shows a collection of heterogeneous observations that we may use to gain insight into the biology, say, of yeast. The collection involves four different object types: proteins, genes, experimental conditions, and functional annotations. Relations are defined as measurements on pairs of objects of the same type: i.e., the matrices PP, GG, CC, and AA. Attributes are defined as measurements on pairs of objects of different types, e.g., entries of the matrix GC measure the expression of genes under the various experimental conditions.*

Such an integrated view of the data available for a given scientific problems invites us to think about the semantics and the substance of the relations among object types in the context of the application at hand, independently of whether or not observations about them are available. This process is beneficial as it is often suggestive of new research directions.

Example 2. *Proteins are composed of one or more subunits. In turn, each subunit is composed of one or more linear polypeptide molecules, which are polymers of twenty different amino acids, i.e., residues. Amino acids can be modified once they have been incorperated into a polypeptide and the presence of these modifications may have a strong influence on the functionality of the final protein molecule. Such modifications are called post-translational (Alberts et al., 2002). The matrix PG in Figure 1.1 encodes the mapping between proteins and gene tags after translation—the matrix GG encodes correlations between microarrays expression profiles of gene pairs. Should*

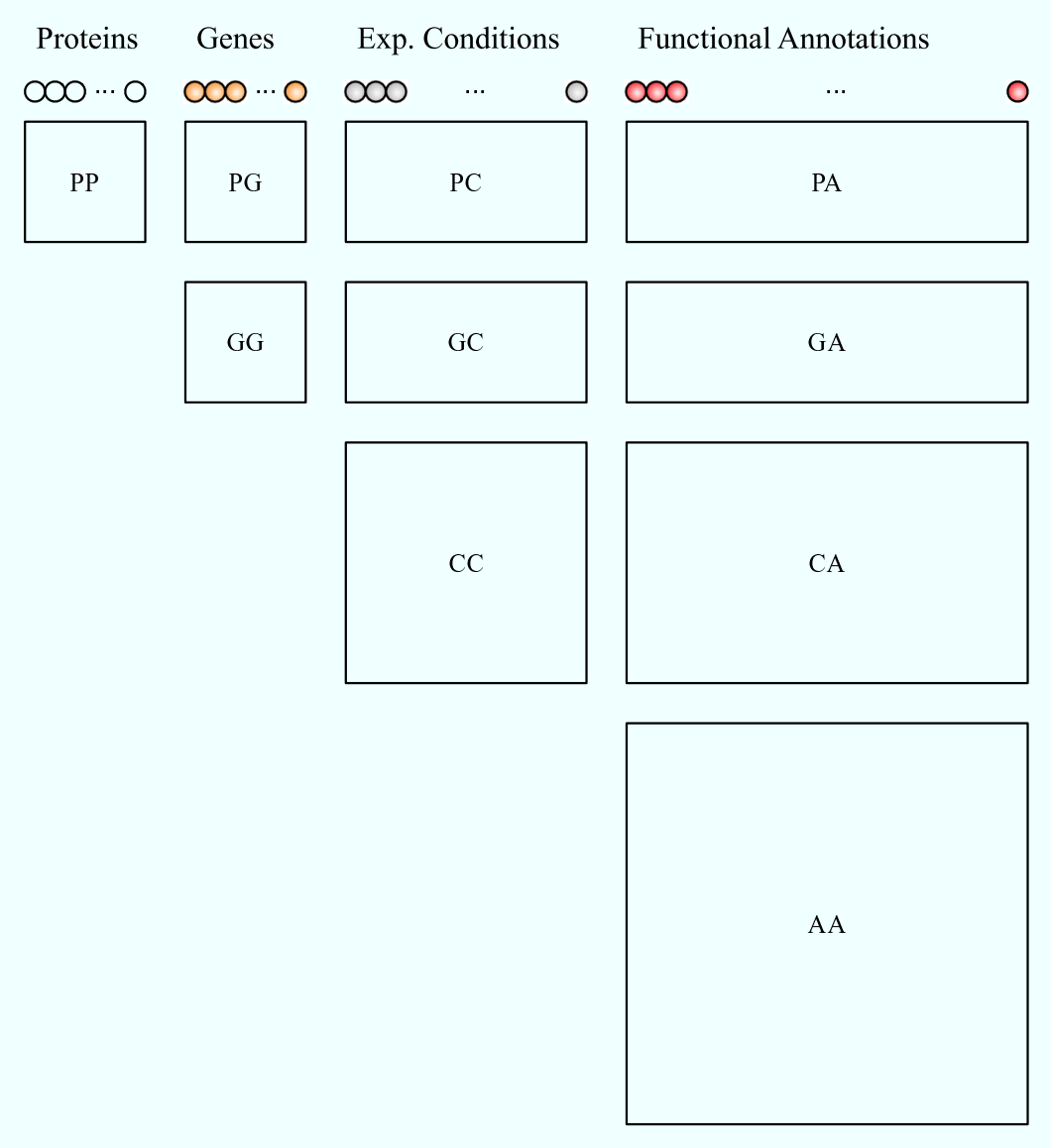

Figure 1.1: An example of complex data. The figure shows an integrated (but partial) view of the observations about a biological system. The types of objects are proteins, genes, experimental conditions, and functional annotations. The various rectangles are suggestive of the matrices that encode the edge weights of the networks (unipartite and bipartite) among pairs of nodes of the corresponding types.

post-translational modifications be negligible, we should see a one-to-one mapping between genes and proteins. Figure 1.1 suggests a way to get at such a mapping, which provides an alternative to what has been proposed in the literature (Tsur et al., 2005). That is, we could estimate the

mapping between proteins and gene tags by looking at the consistency of the interactions between stable protein complexes underlying the protein-to-protein network encoded in the matrix PP and the interactions underlying the gene-to-gene network in encoded in the matrix GG.

Other instances of complex data arise in diverse applications such as artificial intelligence (Carley, 2002), biology (Troyanskaya et al., 2003; Airoldi and Xing, 2006b), information retrieval (Barnard et al., 2003), natural language processing (Griffiths et al., 2005; Kontorovich et al., 2006), and statistical network analysis (Airoldi et al., 2007a).

Example 3. *The analysis of large collections of scientific publications involves a different set of object types: authors, documents, words and references (treated as documents' attributes). Nonetheless, the the data can be represented with a similar set of matrices, e.g., documents-to-words, documents-to-references, and authors-to-authors. Depending on the availability of data and on the scientific questions of interest, researchers typically focus on one, or at most a few, of such matrices (Erosheva et al., 2004; Airoldi et al., 2006e).*

Example 4. *The analysis of a dynamic communication network involves object types such as employees, emails and words. The data can be represented with a set of matrices very similar to those in Example 3. If the analysis takes place within a corporate environment, we may involve more object types such as tasks, resources. The matrices involving these new types, e.g., employees-to-tasks, employees-to-resources, tasks-to-tasks, and tasks-to-resources, would enrich our representation of the inner workings of the company, and allow to ask a different, possibly more interesting, set of questions (Carley, 2002).*

1.1.1 Abstract Representations

From a modeling perspective, it is useful to complement the discussion above with an overview of the data, and how they are represented. Consider a collection of *relations* measured on pairs of

nodes, and a collection of *attributes* measured on the same set of nodes. Such collections can be represented by a unipartite graph and by a bipartite graph, respectively. I choose notation that is suggestive of the elements of the *random matrices* that encode the edge weights in these graphs, rather than the more standard notation based on vertices, edges, and a map from edges to edge weights. A matrix representation of a such pair of unipartite and bipartite graphs is given, for example, by the matrices PP and PG in Figure 1.1. Relations correspond to edges in the unipartite graph PP, and connect paris of objects of the same type, e.g., proteins—the only type of objects in PP. Attributes correspond to edges in the bipartite graph PG, and connect pairs of objects of different types, e.g., proteins and gene-tags. The set of attainable values of relation and attribute measurements is application specific.[2]

For example, a collection of relations is denoted by

$$G_{1:R} = \{\ G_r : r = 1, \ldots, R\ \},$$

where the index r runs over R replicates. Each unipartite graph $G_r = (Y_r, \mathcal{N})$, is defined by a set of edge weights, Y_r, over a fixed set of nodes, $\mathcal{N}$, e.g., the proteins in PP. The random quantities that encode the edge weights between pairs of nodes (n, m) in $\mathcal{N} \otimes \mathcal{N}$ are denoted by $y_r(n, m)$, and take values in a separable, metric space.[3] I refer to Y_r as a random matrix whenever $Y_r(n, m)$ takes values in $\mathbb{R}$ for all edges in $\mathcal{E}$. A collection of attributes is denoted by

$$H_{1:R} = \{\ H_r : r = 1, \ldots, R\ \},$$

where the index r runs over R replicates. Each bipartite graph $H_r = (X_r, \mathcal{N}_{1:2})$, is defined by a set of edge weights, X_r, over two fixed sets of nodes of different types, $\mathcal{N}_1$ and $\mathcal{N}_2$, e.g., the proteins

[2] Always a separable (i.e., contains a countable, dense subset) metric space.

[3] It is possible to introduce the set of edges, and define Y_r as a mapping from edges to edge weights—an unnecessary complication at this stage.

and the gene-tags in PA. The random quantities that encode the edge weights between pairs of heterogeneous nodes (n, m) in $\mathcal{N}_1 \otimes \mathcal{N}_2$ are denoted by $x_r(n, m)$, and take values in a separable, metric space. When dealing with both unipartite and bipartite graphs for which $\mathcal{N} \equiv \mathcal{N}_1$, e.g., that is the case for PP and PA, it is sometimes convenient to denote the set of attributes in $\mathcal{N}_2$ by a collection of node-specific random quantities $x^r_{1:N}(m)$, where N is the number of nodes in $\mathcal{N}$, and m is one of the M distinct attributes in $\mathcal{N}_2$—the replicate index r has been moved to the exponent for clarity.

1.2 Goals of the Analysis

I distinguish two main types of analyses: *descriptive* versus *predictive*. In a descriptive analysis the goal is to find a model that captures the variability of the observations with high probability—in terms of the estimates for the underlying constants, and in terms of the inferred distributions over the latent quantities involved. In a predictive analysis the goal is to find a model that is good at predicting a specific set of attributes or relations from another set of attributes or relations. The ability of such a model in replicating the variability of the observations may be sacrificed in this case, since estimates and distributions assign high probability to the data do not necessarily lead to good predictions. The divergence of objectives between descriptive and predictive analyses is analogous to that between the probabilistic versions of principal component analysis (Jolliffe, 1986; Tipping and Bishop, 1999) and Fisher's linear discriminant analysis (Fisher, 1936, 1938; Ripley, 1996).

The models I consider in this thesis are slightly more complex. They posit a hierarchy of probabilistic *assumptions* on the way observables, (Y, X), and non-observables, Ξ, related to objects of different types are generated, and depend upon and interact with one another. Given these assumptions, the models summarize the complexity of the observations in terms of a set of *latent patterns*.

Patterns are defined in terms of a set of parameters, Θ, which are also non-observable but which are semantically distinct from Ξ. Small sets of underlying constants, e.g., $\mathcal{A}$ and $\mathcal{B}$, sit at the top of the hierarchy, constrain the space of non-observable quantities, (Ξ, Θ), and ultimately constrain the likelihood of the observations,

$$\ell(Y, X|\mathcal{A}, \mathcal{B}) \; = \int f(Y, X, \Xi, \Theta|\mathcal{A}, \mathcal{B}) \; P(d\Xi, d\Theta). \tag{1.1}$$

Likely patterns, Θ, and likely values of other non-observable quantities, Ξ, are searched for, and found, in the data. They may be used for organizing and simplifying complex information, determining object similarity, detecting outliers, and making predictions about attributes of, and relations among, the objects involved.

The analyses supported by these models boil down to a subset of four fundamental tasks: (1) allocation, that is, the search for a likely mapping of objects to patterns; (2) estimation, that is, the search for likely values of the underlying constants; (3) inference, that is, the search for likely values of patterns and other non-observables; (4) prediction, that is, the search for likely values of attributes and relations that need be predicted. For example, testing hypothesis about the existence of a specific pattern, $\theta_0 \in \Theta$, can be carried out by building, e.g., a plug-in confidence region $\widehat{\mathcal{R}} = \mathcal{R}(\widehat{\Theta})$ for Θ, and checking whether θ_0 belongs to it. The tasks relevant to a specific analysis can be carried out simultaneously, since the relevant quantities, i.e., observables, non-observables, and underlying constants, are tied together in a hierarchy of probabilistic assumptions by the generative algorithm.

Example 5. *Consider the output of a battery of microarray experiments on the same set of genes, $\mathcal{N}$, under different, R, experimental conditions, in yeast (Krogan et al., 2006). Proteins are uniquely identified by genes in the microarray experiments. Without entering into biological details, I wish to analyze probabilities of interactions between pairs of proteins, which are induced from correlations found in the gene expression experiments (Bhardwaj and Lu, 2005). Information about*

this unique, symmetric relation can be stored in a collection of R square tables, one for each experimental condition, whose entries are random variables with support $[0, 1]$ that encode the probability of an interaction between corresponding pairs of proteins. The analysis of the set of protein-protein interactions aims primarily at identifying stable protein complexes, i.e., clusters of proteins, since they have been shown to be important for carrying out cellular processes. Further, the number of protein complexes that are needed to explain the collection of protein interactions needs be identified. Lastly, the probabilities according to which pairs of such protein complexes interact with one another need be estimated.

An aspect of the methodology that is relevant to the discussion here is the presence in the proposed models of non-observables, Ξ, which encode *semantic elements of interest in a specific application*, e.g., the stable protein complexes of Example 5. This implies that such non-observables are potentially measurable, and, typically, few measurements about them are available—or can be made available at a cost. A special attention is given in the analyses to such latent quantities, and to other attributes or relations that are measured with error, e.g., experimental evidence or human annotators disagree on their values, on a small portion of the objects of study, e.g., they are expensive to obtain. I will often refer to partially available measurements about such attributes, relations and non-observables as *labels*. The portion of *labeled data* available is of interest for the estimation of the prediction error, and an explicit error model for the label is often desirable. Further, depending on the amount of labeled data available, different strategies for initializing the inference,[4] for fitting the underlying constants, and for inferring the distributions on the latent quantities given the data may be adopted.

Example 6. *Consider the set of hand-curated protein interactions produced by the Munich Institute for Protein Sequencing (Mewes et al., 2004). A single set of interactions between proteins has been experimentally verified. Information about this unique, symmetric relation can be stored in*

[4]Differences that have important consequences on the interpretability of the estimates and of the inferred distributions on the latent quantities.

one square table, whose entries are random variables with support $\{0, 1\}$ *that encode presence or absence of an interaction between corresponding pairs of proteins.*

1.3 Basic Modeling Elements

There are few central modeling ideas that inform the probabilistic algorithms presented in the following chapters. These ideas generalize model specifications that were used to gain insight into fundamental problems of computational biology, i.e., serial analysis of gene expression (Airoldi et al., 2006f) and protein interaction networks (Airoldi et al., 2006c), and into the analyses of large collections of scientific publications (Airoldi et al., 2006e) and of dynamic communication networks (Airoldi and Faloutsos, 2004; Airoldi et al., 2005d). They relate to the following four aspects of complex data: (1) the presence of a hierarchical structure in the likelihood, which includes both observable and non-observable random quantities, (2) the mixed membership assumption, according to which objects may participate in multiple latent patterns to different degrees, (3) the temporal dimension, and (4) the existence of multiple data types, and conditional dependencies among their distributions, in an integrated system.

These aspects are best illustrated below by discussing how they generalize popular data analysis models such as probabilistic principal component analysis (PPCA), factor analysis (FA), and state-space models (SSM).

1.3.1 Hierarchy and Latent Patterns

Let us consider a collection of attributes, $H = (X, \mathcal{N}_{1:2})$, and let us adopt the point of view of multivariate attribute measurements, $\vec{x}_{1:N}$, on the N objects in $\mathcal{N}_1$ about the M objects in $\mathcal{N}_2$.

Example 7. *The data generating process for* X *underlying factor analysis is instantiated by a*

probabilistic algorithm, $A1 : (\mathcal{N}_{1:2}, K, \Lambda, \Psi) \rightarrow \mathbb{R}^M$*,*

1. *For each object* $n \in \mathcal{N}_1$

   1.1. *Sample the latent factors* $\vec{\phi}_n \sim Normal_{\ K}\ (0, I)$

   1.2. *Sample the error* $\vec{\epsilon}_n \sim Normal_{\ M}\ (0, \Psi)$

   1.3. *Define the multivariate attribute* $\vec{x}_n = \Lambda\vec{\phi}_n + \vec{\epsilon}_n$*,*

where K *is typically referred to as the number of (scalar) factors,* Λ *is a deterministic matrix of factor loadings, and* Ψ *is an unconstrained variance-covariance matrix.*[5] *The algorithm suggests a hierarchical decomposition of the joint probability distribution of the attributes,* $X = \vec{x}_{1:N}$*, and the latent factors,* $\Theta = (\vec{\phi}_{1:N}, \vec{\epsilon}_{1:N})$*, given a set of underlying constants,* $\mathcal{A} = (\Lambda, \Psi)$*; that is, the integrand in Equation 1.2. By integrating the latent variables out of the joint we obtain the likelihood of the observations,*

$$\ell(X|\mathcal{A}) \ = \int f_1(\Theta|\mathcal{A})\ f_2(X|\Theta, \mathcal{A})\ d\Theta, \tag{1.2}$$

where f_1 *and* f_2 *are* K*- and* M*-dimensional Gaussian densities, respectively.*

In FA the latent factors are an example of *patterns*, the way I intend them; they are non-observable random quantities, defined in terms of a set of scalar parameters. Depending on the model, patterns may specify other mathematical objects such as probability distributions, smooth curves, and surfaces.

Confusion may arise about the notation for patterns, Θ, and underlying constants, $\mathcal{A}$, in those cases where latent patterns are defined to be deterministic. In such case the patterns would occupy a spot at the top of the hierarchy, similarly to the underlying constants, thus leaving us the choice to

[5] Note that PPCA differs from FA only in that the variance-covariance matrix of the errors, $\vec{\epsilon}_{1:N}$, is homoschedastic, that is, $\Psi = \sigma^2 I$ with σ scalar.

include Θ in $\mathcal{A}$ or not.[6] I shall clarify the use of notation whenever the occasion requires it. Further, I note that the latent patterns, Θ, and other non-observable random quantities with a semantic interpretation, Ξ, that appear in the general formulation of Section 1.2, are to be interpreted as part of a *hierarchical likelihood* since they typically model substantive elements of interest in the application at hand.

Example 8. *A simple mixture of spherical Gaussians for $\vec{x}_{1:N}$ can be specified by the following probabilistic algorithm, $A2 : (\mathcal{N}_{1:2}, K, \vec{\mu}_{1:K}, \Sigma_{1:K}) \rightarrow \mathbb{R}^M$*

1. *For each object $n \in \mathcal{N}_1$*

   1.1. *Sample the latent component indicator $i_n \sim Uniform\,(1, \ldots, K)$*

   1.2. *Sample the multivariate attribute $\vec{x}_n \sim Normal\ _M\,(\vec{\mu}_{i_n}, \Sigma_{i_n})$,*

where K is the number of mixture components, $(\vec{\mu}_{1:K}, \Sigma_{1:K})$ are the corresponding mean vectors and variance-covariance matrices, and $\Sigma_k = \sigma_k^2 I$ with σ scalar. The likelihood can be written as in Equation 1.2, where the attributes $\vec{x}_{1:N}$ are denoted by X, the latent component indicators $i_{1:N}$ by Θ, and the underlying constants $(\vec{\mu}_{1:K}, \Sigma_{1:K})$ by $\mathcal{A}$. In this case where f_1 and f_2 are discrete uniform and M-dimensional normal densities, respectively.

In the example above, the underlying constants $(\vec{\mu}_{1:K}, \Sigma_{1:K})$ qualify as patterns. It is conceivable to put probabilistic constraints on such quantities, e.g., by assuming that they are generated from some distributions. By doing so, I would introduce a new, more parsimonious set of underlying constants, $\mathcal{A}$, and promote $\Theta = (\vec{\mu}_{1:K}, \Sigma_{1:K})$ to be the non-observable, probabilistic patterns of the general formulation of Section 1.2.

[6] It could be argued, for instance, that the matrix of factor loadings, Λ, should be considered a part of the patterns underlying a set of attribute measurements, X, as much as the factors, Θ, themselves. However, in the hierarchical formulation I consider here, it is not difficult to imagine the use of a probabilistic model for Λ to endow the loadings with some desirable property (Airoldi and Lin, 2006).

Thus a generative algorithm and the corresponding hierarchical likelihood specify exactly how the various quantities of interest interact, in a probabilistic fashion, and encode structural hypothesis of the scientist.

1.3.2 Mixed Membership

The idea of *mixed membership* extends that of a mixture. Stated briefly, this assumption posits that the collection of measurements involving an object, i.e., both relations and attributes, may be ultimately explained in terms of multiple patterns to different degrees. A recurring element of my models is that such representations of latent patterns are associated with the components of a mixture, as in the example above. In this sense, both mixture models and mixed-membership models aim at describing the aggregate variability of a set of measurements in terms of a small set of latent patterns. There are two major salient differences, however, between a mixture model and a mixed membership model.

(i) In a mixture model the membership of objects to patterns is specified in terms of global weights. In a mixed-membership model the membership of objects to patterns is specified in terms of object-specific weights; these give a low-dimensional representation of the objects that can be used for, e.g., making predictions about object-specific quantities.

(ii) Measurements in a mixed membership model can be associated with more than one latent patterns. The role of sparsity in this context is to impose soft constraints in the estimation of the mapping between objects and latent patterns—I term this estiamtion the *allocation task*.

For instance, relations or attribute of an employee in Example 4 may be explained in terms of the latent patterns associated with more than one social group, and the interactions between two individual proteins can be explained by their taking part into more than one stable protein complex.

Example 8 provides us with another element of the general formulation of Section 1.2, that is, the set of non-observables that encodes semantic elements of interest. For instance, in the application to dynamic networks described in Example 4, or in the application to gene co-expression networks in Airoldi and Xing (2006b), the Gaussian mixture components are suggestive of latent social groups and latent stable protein complexes, respectively. In these specific applications, the non-observables $\Xi_{1:N}$ encode the *single* membership to a social group, or to a stable protein complex—whereas the latent patterns Θ provide parametric representations of social groups and complexes.

Example 9. *Going back to the previous example, the M scalar components of each multivariate attribute $\vec{x}_n$ are no longer constrained to be sampled from the same latent pattern, i.e., from the same mixture component, $(\vec{\mu}_k, \Sigma_k)$. The new algorithm, $A3$, which instantiates the mixed membership of K spherical Gaussians is as follows,*

1. *For each object $n \in \mathcal{N}_1$*

    1.1. *For each attribute $m \in \mathcal{N}_2$*

        1.1.1. *Sample the latent component indicator $i_n \sim Uniform\ (1, \ldots, K)$*

        1.1.2. *Sample the scalar attribute $x_n(m) \sim Normal\ (\mu_{i_n}(m), \sigma_{i_n}(m))$.*

Alternatively, it is possible to illustrate this richer mapping between observables measured on an object and mixture components entailed by the mixed membership assumption in terms of a general form for the likelihood. That is, we can rewrite the likelihood as an admixture for each univariate measurement. The mixture likelihood in Equation 1.2 can be specified as,

$$\ell(X|\mathcal{A}) = \prod_n \int f_1(\Theta|\mathcal{A}) f_2(\vec{x}_n|\Theta, \mathcal{A})\, d\Theta, \tag{1.3}$$

whereas the admixture likelihood corresponding to Algorithm $A3$ can be rewritten as,

$$\ell(X|\mathcal{A}) = \prod_{n,m} \int f_1(\Theta|\mathcal{A}) f_2(x_n(m)|\Theta, \mathcal{A}) \, d\Theta, \tag{1.4}$$

where f_1 and f_2 are Gaussian densities of appropriate dimensionality. Note that this is also the case for PPCA, but not for FA—because of possible structure in Ψ. In PPCA, however, the space of multivariate attributes is not to the convex cone spanned by the K non-observable quantities, because the factor loadings do not lie in the K-dimensional simplex—this is the case with f_1, which is a probability distribution on the K latent patterns, $\Theta_{1:K}$.

A final note concerns the mapping between observations and mixture components. Application-specific features of the mapping itself are typically of interest, since they impact the latent patterns found, and the results of the analysis more in general. One of the features often supported by the data is that such a mapping is sparse; that is, each univariate measurement can be ultimately explained in terms of a few latent patterns (i.e., mixture components). Further, in many applications the mapping is skewed; that is, that many univariate measurements can be ultimately explained in terms of a few patterns, whereas a few univariate measurements can be explained in terms of many of them.

Lastly, it is important to recognize that the choice between alternative specifications (e.g., parametric, semi-parametric, or ad-hoc) about the number of non-observable quantities, K, in these models is not a matter of mathematical elegance. Such a choice typically has a non-negligible impact on the substantive findings and their interpretation, therefore it should be motivated and discussed in terms of the specific scientific problem of interest, by the amount of information available about such non-observable quantities, as well as by the goals of the analysis (e.g., exploratory versus conclusive). For example, Krogan et al. (2006) find that the average size of stable protein complexes is about five proteins. That suggests the existence of a larger number of such

non-observable complexes, prior to any analysis of a new set of proteins, $\mathcal{P}$, than popular semi-parametric model specifications would entail; that is, $O(|\mathcal{P}|)$ rather than $O(\log|\mathcal{P}|)$.

1.3.3 Dynamics and Evolution

Several models of dynamic behavior exist in the literature, which can be used to model the evolution of latent patterns for a finite number of epochs, $\Theta^{(1:T)}$.

Example 10. *A linear, Gaussian state-space model extends the factor analysis model of Example 7, by linearly evolving the latent factors from one epoch to the next. The data generating process for $X^{(1:T)}$ is as follows,*

1. *At epoch $t = 0$*
   - 1.1. *For each object $n \in \mathcal{N}_1$*
     - 1.1.1. *Sample the latent factors $\vec{\phi}_n \sim Normal_{\ K}\ (0, I)$*
     - 1.1.2. *Sample the error $\vec{\epsilon}_n^{(0)} \sim Normal_{\ M}\ (0, \Psi)$*
     - 1.1.3. *Define the multivariate attribute $\vec{x}_n^{(0)} = \Lambda\vec{\phi}_n + \vec{\epsilon}_n^{(0)}$,*
2. *At epoch $0 < t < T$*
   - 2.1. *For each object $n \in \mathcal{N}_1$*
     - 2.2.1. *Evolve the latent factors $\vec{\phi}_n^{(t)} = F\vec{\phi}_n^{(t-1)}$,*
     - 2.2.2. *Sample the error $\vec{\epsilon}_n^{(t)} \sim Normal_{\ M}\ (0, \Psi)$*
     - 2.2.3. *Define the multivariate attribute $\vec{x}_n^{(t)} = \Lambda\vec{\phi}_n^{(t)} + \vec{\epsilon}_n^{(t)}$,*

where F is a $(K \times K)$ matrix that encodes the dynamics of the latent factors. As before, the algorithm suggests a hierarchical decomposition of the joint probability distribution of the attributes,

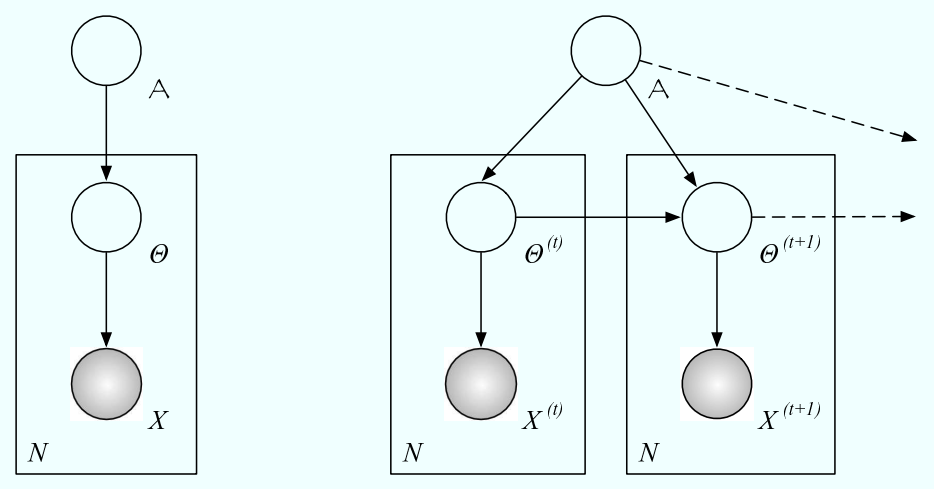

Figure 1.2: Graphical representations of a factor analysis model (left) and of a state-space model for observations at two consecutive epochs (right). White nodes denote non-observables, whereas shadowed nodes denote observables.

$X^{(1:T)} = \vec{x}_{1:N}^{(1:T)}$, *and the latent factors,* $\Theta^{(1:T)} = (\vec{\phi}_{1:N}^{(1:T)}, \vec{\epsilon}_{1:N}^{(1:T)})$, *given a set of underlying constants,* $\mathcal{A} = (F, \Lambda, \Psi)$ *that does not change over time.*[7] *The likelihood is then,*

$$\begin{aligned} \ell(X^{(1:T)}|\mathcal{A}) &= \int f_1(\Theta^{(0)}|\mathcal{A})\ f_2(X^{(0)}|\Theta^{(0)}, \mathcal{A}) \times \\ &\times \left(\prod_{t=1}^{T} f_0(\Theta^{(t)}|\Theta^{(t-1)}, \mathcal{A})\ f_2(X^{(t)}|\Theta^{(t)}, \mathcal{A}) \right) d\Theta^{(1:T)}, \end{aligned} \tag{1.5}$$

where f_1 *and* f_2 *are* K*- and* M*-dimensional Gaussian densities, respectively, and* f_0 *is the deterministic transformation in Step 2.2.1. of the data generating process. A graphical representation of FA and SSM is given in Figure 5.1, which highlights the simple connection between the two models.*

In my models, I consider three flavors of dynamics:

(1) a generalized state-space process (Brockwell and Davis, 1991; Xing, 2005a), possibly non-linear and non-Gaussian, which provides my models with a fully-parametric, tractable dy-

[7] The dynamic matrix F may be easily modeled as time dependent and/or stochastic, as the problem requires (Airoldi and Faloutsos, 2004; Airoldi et al., 2005d).

namics;

(2) a latent birth-death process (Karr, 1991; Xing, 2005b; Airoldi and Xing, 2006a) that allows for a possibly infinite number of patterns (semi-parametric) and generates complex pattern dynamics; and

(3) a co-evolutionary process (Carley, 1990, 1991) that induces highly non-linear dynamics by tying together the temporal behaviors of observables and non-observables.

Technical issues arise with the increasing complexity of the dynamical behaviors above.

As an alternative, it is possible to specify temporal patterns directly, as a part of (Ξ, Θ). Such a modeling strategy allows us to consider a sequences of observations about objects as being expressed as an admixture of complicated patterns, specified in a parametric or non-parametric fashion, while avoids technical issues that arise when the estimation of an explicit dynamics is considered.

1.3.4 Integration

Integration of the measurements on relations and attributes involving objects of different types may, and will, take many forms in the models considered throughout this thesis, and it seems unnecessary to list them all at this stage. It will suffice to distinguish two types of integration, one relates to descriptive versus predictive analyses, and the other relates to the integration of labels.

Example 11. *Consider the following generative process for a set of relations $G = (Y, \mathcal{N})$ among objects in $\mathcal{N}$, a set of multivariate attributes $H_x = (X, \mathcal{N}, \mathcal{T})$ on the same set of objects, and a set of labels $H_z = (Z, \mathcal{N}, \mathcal{C})$.*

1. *Sample the mixed-membership scores for objects in $\mathcal{N}$ according to*
   $(\vec{\pi}_{1:N}) \sim f_1(\vec{\pi}|\mathcal{A})$

2. *Sample the latent pattern indicators for object-specific relations and attributes, independently and given their mixed-memberships, according to*
   $(I_Y, I_X) \sim f_2(I|\vec{\pi}_{1:N})$

3. *Sample the observations given object-specific patterns according to*
   $(Y, X) \sim f_3(Y, X|I_X, I_Y, \Theta, \mathcal{A})$

4. *Sample predictive indicators for the objects' labels from the corresponding sets of object-specific pattern indicators that were sampled according to*
   $(I_Z) \sim f_4(I|I_Y, I_X)$

5. *Sample the labels given the predictive indicators according to*
   $Z \sim Generalized\ Linear\ Model\ (Z|I_Z, \mathcal{A})$

where $\Xi = (\vec{\pi}_{1:N}, I_Y, I_X, I_Z)$*, and the latent patterns* Θ *are deterministic. Relevant to the discussion here is the composition of the relations and attributes* (Y, X) *as independent sources of information in Steps 2–3, versus the composition of the labels* Z *as conditional source of information in Steps 4–5.*

In a descriptive analysis, sets of non-observables corresponding to different data sources always contribute equally to the data generation, and, in turn, observables always inform equally the inference process about the corresponding sets of non-observables. This is what happens with the relations and attributes (Y, X) in Example 11 and with the corresponding latent pattern indicators for relations and attributes (I_Y, I_X). In a predictive analysis, one set of non observables always contributes to the data generation conditionally on the values assumed by a second set of observables, and, in turn, the two sets of observables inform the inference process about non-observables unequally—namely, the information the latter set contributes to the inference process is used to describe *residual variability*, which cannot be explained by information contributed by the former

set of observables. This what happens to the labels Z in Example 11 with the corresponding latent pattern indicators I_Z.

1.4 Overview of the Research

Complexity of the observations is resolved into hierarchical mixtures of simple patterns that evolve over time, i.e., complex global behavior emerges from the combination of local (i.e., object-specific) interaction patterns and dynamics. This solution provides a principled approach to reconcile global properties the system with local phenomena of interest. Structured models similar to those shown in Figure 5.1 are often referred to as hierarchical models in the statistics literature (Kass and Steffey, 1989; Gelman et al., 1995). Estimation techniques include empirical Bayes (Morris, 1983; Carlin and Louis, 2005) and full Bayesian methods (Mosteller and Wallace, 1964, 1984; Airoldi et al., 2006a). The general model formulation I explore in this thesis subsumes many probabilistic models present in the literature, provides a *soft* and probabilistic version of many non-probabilistic algorithms, and most importantly provides the essential statistical elements for the analysis of complex data, random graphs and matrices, and dynamic networks.

In Chapter 2, I survey existing algorithms to generate popular *topologies* in unipartite graphs. I then present proper statistical models to generate such topologies, complete with likelihoods and estimators for the parameters involved. I conclude by exploring the lognormal and cellular graphs. In Chapter 3, I describe different ways to search for patterns and mechanisms underlying networks. In Chapter 4, I consider attributes, I describe an extension of the models to multivariate attributes and relations, and I describe strategies to integrate complex data into a large statistical model. In Chapter 5, I describe models of evolution for attributes and relations. Finally, in Chapter 6, I explore a selection of theoretical and computational issues associated with the general formulation of my models and describe aspects of future research.

1.4.1 Contributions of this Thesis

This thesis develops statistical methodology for the Bayesian analysis of data that arise in studies about complex networks and their evolution. The connection between modeling choices and substantive issues is kept at the forefront of the discussion. Furthermore, complexity in the various models is pursued only to the extent that it responds to needs that are rooted in the data and the goals of the analysis. Such a focus on the data and their properties is compatible with the development of a general modeling framework for the analysis of complex and evolving networks thanks to the central role played by few essential modeling elements—described in Section 1.3—that can be used to describe complex dynamic systems in general.

1. In many applications there is an large amount of information available with a temporal or sequential dimension. Methods that explicitly account for dynamics and evolution of the phenomena under scrutiny are much needed. Modeling approaches available to date, for which solid inference mechanisms are available, include hidden Markov models and generalized state-space models. New methodology and modeling strategies are needed that can account for richer evolutionary patterns of complex sets of measurements, i.e., relational and non-relational. Furthermore, there is a need for producing predictions that are based on several sources of information, which need be integrated; a solid probabilistic approach to this end is missing. This thesis develops a modeling framework that responds to these needs.

2. The models that can be specified working within the framework proposed in this thesis are extremely diverse and widely applicable. Many scientific studies lead to data sets that are represented as graphs, at some level, e.g., two-mode data lead to bipartite graphs, uni-modal data to graphs where we record relations between pairs of objects of the same type, multi-mode data where we record relations among objects of multiple types, multi-graphs where edges encode multivariate variables, and combination of these. Assumptions and intuitions

of interest may need be incorporated in application-specific models, but the modularity of my approach makes these *special* modeling issues a piece of the puzzle that can be addressed separately—by instantiated on of the integration strategies of Section 4.3—on top of a set of data source-specific models.

3. The proposed framework subsumes several models recently proposed in the machine learning and applied statistics literature and ties them together within a general formulation that is amenable to theoretical analysis. Therefore the proposed framework opens new analytical possibilities by allowing us to address theoretical aspects of interest in terms of the specifications of the general formulation. This high-level theoretical analysis disregards the nuances present in application-specific models and focuses on fundamental technical issues, such as identifiability, model selection, distribution free tests for assessing the goodness of fit, the geometrical understanding of allocation tasks, or the asymptotics of the family of hierarchical mixed-membership models. Even a limited explorations of these issues would advance the scientific understanding of the methodology. This exercise will ultimately benefit applications by providing theoretical insights to support application-specific modeling choices.

The grand vision is to establish a mature statistical theory of graphs and networks that can bridge theoretical computer science, a largely deterministic discipline, and statistical theory. This can be achieved, for example, by explicitly characterizing the relation between deterministic and probabilistic solutions to problems shared by both disciplines that involve graphs and networks. The ultimate goal is that of promoting the role of statistical Bayesian theory in the computing sciences and its modern applications.

1. This thesis provides solid foundations of a statistical theory of mixed-membership and exchangeable-edge models of graphs and networks and their evolution. Such foundations are missing, to the best of my knowledge. This is a goal worth pursuing in its own right.

2. This thesis promotes the role of Bayesian statistics in the theoretical computer science and data mining communities by providing new models and perspectives in applications of primary importance, for example, to biological sequences & networks, dynamic social networks, collections of scholarly publications, knowledge and corporate networks, and homeland security. The proposed general framework aims at fostering scientific progress by serving as a glue for several branches of the literature that are poorly aware of one another.

In conclusion, recent trends and events suggest an imminent shift of focus of the research community at large towards complex interacting dynamic systems, along with a rediscovered mindset that through integration we can finally deliver satisfactory solutions to long-standing real-world problems, as well as create new applications. This thesis presents methodology derived from applications for applications, and provides insights and understanding on when we can expect the methods we employ will work, and why. Specifically, I discuss models and methods that enable applications to biological databases, collections of scientific publications, and dynamic social and corporate networks. Success stories where my methods were key to answer real-world problems provide the background for the discussion. I argue that my efforts establish the foundations of a statistical/computational theory of complex networks and their evolution.

1.4.2 Limitations

This thesis develops a modeling framework to tackle specific applications. As a consequence topics and modeling approaches are omitted that I believe are important for the analysis of complex and evolving networks. I am currently working on an extension of the modeling framework developed here that addresses such topics and modeling approaches. A short list of what is missing in this thesis follows.

- Connections to statistical theory and methodology for the analysis of networks that does not

involve latent variables (Wasserman and Faust, 1994; Wasserman et al., 2007).

- A fully developed example of how the statistical methodology developed in this thesis offers a principled approach to tackle calibration and validation issues that arise in large-scale agent-based models and simulations of complex systems (Carley, 1990, 1991; Banks and Carley, 1996; Carley et al., 2006; Shreiber, 2006). However, I outline the main points of the argument in Section 5.2.

- Connections to random matrix theory (Metha, 2004). However, I devote Section 3.2 to situate in the context of this thesis some of the recent developments in the field that bear relevance to statistical network analysis; namely, the mathematical analysis of diffusion (Coifman et al., 2005a,b; Nadler et al., 2005; Lafon and Lee, 2006).

- Generative models of edge patterns (a.k.a. network motifs, node identity does not matter, e.g., see Milo et al. 2002), as opposed to the generative models of node patterns (node identity matters, e.g., see Chapter 3) developed in this thesis. At the model level, such an extension to Bayesian mixed membership models of edge patterns is trivial. However, non-trivial computational issues arise immediately, for example, in the evaluation of the likelihood—where a (combinatorial) search of all instances of the relevant edge motifs needs be performed, i.e., sub-graph matching.

- Models of complex dynamic behavior such as latent birth-death processes (Airoldi and Xing, 2006a), in preparation, and duplication-attachment processes (Wiuf et al., 2006).

- A complete analysis of the mathematical properties of exchangeable-edge models of Section 2.2. These models represent an important extension of the popular random graph model of Erdös and Rényi (1959) and Gilbert (1959), technically, by involving a layer of latent variables. Such an analysis is part of my current research (Airoldi and Carley, 2006; Airoldi and Shalizi, 2006).

Chapter 2

Random Graphs Revisited

Here I survey existing algorithms to generate popular *topologies* in unipartite graphs. Proper statistical models to generate such topologies are then presented, complete with likelihoods. I conclude with a presentation of novel probabilistic algorithms to generate lognormal and cellular graph topologies, along with their analysis.

Introduction and Motivation In order to shed some light on how the interactions among a set of objects of study, e.g., people or proteins, lead to the emergence of observed patterns and properties of interest, both local and global, e.g., groups and the small-world, several generative algorithms have been proposed. These algorithms abstract a small set of essential features of the objects and interactions, and try to replicate local or global patterns and properties of interest—either exactly or approximately, either in a deterministic fashion or with high probability. We consider such algorithms to be insightful whenever they can replicate the observable phenomena of interest, and the small set of essential features which they are based upon suggests us a possible explanation for them.

Example 12. *Milgram (1967) provided empirical evidence in support of the so called small world hyposthesis. Briefly, Milgram instructed a set of people (i.e., sources) in Nebraska, Kansas and Massachussets to send packets to any one of two specific individuals (i.e., targets) in Massachussets. The targets were described approximately in terms of a small set of characteristics such as location, profession, and other demographics. The sources were supposed to send the packets towards the target by sending it to a person they knew on a first name basis, i.e., to an acquaintance the source believed to be closer to the target. The game consisted in delivering the packet to the target with as few of these first-name links as possible. If the small world hypothesis held, the average lenght of the first-name chains of acquantances that connected a source to a target should be independent of the location of the sources. This is exactly what Migram found, the median length being somewhere around six—an independent statistical analysis of Milgram's data that includes incomplete chains suggests six to be a serious underestimate of the actual median length (Fienberg and Lee, 1975). In an abstract setting, we can represent people by means of nodes in a graph, and acquaintances by directed links from a node to another. The scientific phenomenon of interest is the small world; we would like to be able to explain it with a simple model that generates small world graphs with high probability. What kind of generative process should we posit? Watts and Strogatz (1998) propose a rewiring model to answer this question. In their model nodes are embedded in a metric space, $(\mathcal{X}, d)$, and each node is connected to its neighbors according to d by means of undirected edges. Then, with probability p each of the edges that connect a node with its d-neighbors is rewired at random, that is, is disconnected from a d-neighbor and connected to another node in the graph with equal probability. When the rewiring process is carried out for each node, the process ends. Although very simple, the rewiring model of Watts and Strogatz (1998) encodes a key intuition about how acquaintances may be established, that is, the fact that people form local circles of friends, and retain a few of them when they move across the country. This social process is suggestive of the rewiring model. It turns out that the rewiring of local neighbors alone is enough to generate small world graphs with high probability.*

Our ability to spot patterns and properties of interest crucially depends on the *graph metrics* available to us. Many metrics have been proposed over the years to measure various properties of graphs that could explain phenomena of interest, e.g., recurrent connectivity patterns, average path length, or degree distribution. Crucially, each of these metrics pre-encodes some intuition about the phenomena we conjecture may exist. In fact, such metrics are meant to capture numerical properties of those graphs where the phenomena of interest occur, that are absent in other graphs. In other words, we can only *attempt to measure* those patterns that *we believe are distinct* from background noise. The set of metrics available to us is then a byproduct of substantive intuitions about what we cannot see or measure. In a technical sense, each metric encodes a structural hypothesis, i.e., structural bias.

Example 13. *Milgram's (1967) empirical analysis and Watts and Strogatz's (1998) theoretical analysis use different metrics. Milgram considers the average length of first-name chains, and finds that they are consistently short. In particular, such a short average length (alternatively, such a small diameter) is less than what we whould expect to observe were the acquaintance network a purely random graph (Erdös and Rényi, 1959; Gilbert, 1959). However, there is possibly an infinite number of ways to generate graphs with a diameter smaller than that of a random graph. Which properties of small works graphs unequivocably distinguish them from others? In order to identify small world graphs, Watts and Strogatz consider the characteristic path length (closely related to Milgram's metric) and the clustering coefficient.*[1] *Stated formally, they find that the dimater of a graph drops from* $O(n)$ *to* $O(\log n)$ *even when a small fraction* p *of the edges is rewired,*[2] *whereas the clustering coefficient remains close to* $3/4$*. The rewiring process has little effect on several other metrics as well.*

[1] In Section 2.1 we show *to what extent* these two metrics can distingush graphs with a small world topology from graphs with other topological properties.

[2] Although, as noted by Bollobás and Riordan (2003), this fact is a particular instance of a classic result of random graph theory about the diameter of the giant component (Erdös and Rényi, 1960), there is much merit in the suggestive power of the simple generative model introduced by Watts and Strogatz (1998), who place such result in a context relevant to the scientific community at large.

Brief Overview of Results During the past few years the attention of the scientific community has increasingly focused on complex graphs and dynamic networks. As a consequence, many scientific investigations that involve graphs to some degree attempt an assessment of how sensitive the main findings are to topological properties of the graphs. The general trend, however, is for such investigations to leverage *popular models* of graphs, rather than focusing on their *topological properties*.

Example 14. *High throughput techniques have made way to the collection of data on many complementary aspects of the biology of the major species living on our planet, and integral approaches to the biological sciences are now possible. As a consequence of this, dependence among observations, data and model integration, and ultimately network science, have become fundamental to our ability to carry on the process of scientific discovery in this domain. Relevant to the discussion here is the fact that more and more research articles on complex biological networks, e.g., protein interaction networks, gene regulatory networks, and metabolic networks, make use of popular models of graphs such as the scale-free model (Barabasi and Albert, 1999), which is consistent with different graph topologies (Bollobás and Riordan, 2003), rather than investigating the topological properties of such networks directly.*

A crucial issue is then the mapping between popular models of graphs and the topological properties they possess. In particular, under scrutiny is whether the various *graph topologies* are different; if so, by how much; and whether the different generative algorithms for a specific graph topology lead to the same topological properties. This is in some sense a chicken and egg problem, since our ability to probe the space of topologies is limited, as discussed above, by the set of metrics we use. A brief exploration of these issues is presented in Section 2.1. I find that the popular models, e.g., scale-free and core-periphery, generate graphs with similar topological properties for non-pathological values of the relevant underlying constants. Furthermore, I find that alternative models that supposedly generate graphs with non-distinguishable topological properties, e.g., mod-

els of scale-free graphs by different authors, can be easily distinguished. These findings prompt us to make recommendations about how to provide successful assessments of the sensitivity of an analysis to topological properties of graphs. They also suggest that real-world graphs may be better modeled as mixtures of these popular models (Airoldi and Carley, 2005).

Another major issue is that the popular algorithms that have been proposed in the literature to generate topologies of interest can replicate local and global phenomena, but have no place for the data. That is, given a few underlying constants such algorithms generate observations that display a certain class of behavior, but it is never specified how to *estimate* values for those constants so that the generated behavior *fits a collection of data* the best. Being able to make a good use of the data is crucial whenever models and algorithms have to support analyses of real data. In Section 2.2 alternative mathematical representations of a graph are discussed. Within such context, I show how it is possible to posit a general class of probabilistic models, which I refer to as *exchangeable-edge* models, that generate graphs with pre-specified topological properties of interest, and at the same time allow for *formal inference and estimation* procedures. These models inform novel analyses of lognormal and cellular graphs.

I consider lognormal graphs in Section 2.2.2. I find that scale-free, or power-law, graphs provide us with a first-order approximation to graphs in this class. I then introduce a novel generative model for lognormal graphs that serves to show how models in this class (and hence scale-free or power-law graphs) may arise, and find that they do in two interesting sets of circumstances. First, I find that lognormal graphs may arise as an artifact of the network construction process. Namely, they arise in situations where edges are set between pairs of objects by thresholding correlation among their attributes even when such attributes are completely independent. In that sense, lognormal graphs are an artifact due to the way we measure the presence or absence of relations among objects in a population of interest. Second, I find that lognormal graphs may arise as a consequence of a limiting phenomenon when certain conditions hold on the way edges are established between a

new object and existing objects in the graph; namely, by means of a multiplicative process.

I consider cellular graphs in Section 2.2.3. I find that communities form because of the joint effect of two simple factors: (i) exclusivity, that is, the need for allocating resources to competing interests; and (ii) homophily, that is, the fact that social interactions are more likely to occur between individuals who share interests than between those who do not. Furthermore, I find that communities emerge quickly as exclusivity exceeds a certain threshold.

2.1 Stability and Separability of Metric Embeddings

The context behind the exploration I present here is given by two observations. First, the popularity gained by generative models of graphs have helped establish several flavors of *graph topologies*, in the scientific community at large. For example, few scientists today are unaware of notions such as *scale-free* and *small-world* networks. Because of their popularity, such notions are often leveraged in published research literature—for better or worse. Second, any scientific approach to modeling graphs faces the technical issue of *which minimal set of features can be used to characterize a graph*. The popular answer is to focus on a set of real-valued *metrics* (i.e., functions of edges and vertices of a graph, which are defined to capture specific topological properties) thus characterizing the graph as a vector.[3]

A crucial issue is then to study the mapping between popular models of graphs and the topological properties of the set of graphs they support. In particular, under scrutiny is whether the various graph topologies are different; if so, by how much; and whether the different generative algorithms for a specific graph topology lead to the same topological properties.

[3]In this thesis, I am interested in characterizations that entail a one-to-one map with the space of graphs; this issue is explored in detail in Section 2.2.

Table 2.1: Summary of published generative algorithms.

Type	Algorithm	Parameters
1.1.	Ring Lattice	$\Theta = (N, K_0)$
2.1.	Small World (Watts and Strogatz, 1998)	$\Theta = (N, K_0, Q_0)$
2.2.	Small World (Kleinberg, 1999a)	$\Theta = (N, K_0, K_1, R)$
2.3.	Small World (Airoldi, 2005)	$\Theta = (N, K_0, Q_0, Q_1, R)$
3.1.	Random (Erdös and Rényi, 1959)	$\Theta = (N, M)$
3.2.	Random (Gilbert, 1959)	$\Theta = (N, P)$
4.1.	Core–Periphery (Borgatti and Everett, 1999)	$\Theta = (N, P, P_0)$
4.2.	Core–Periphery (Airoldi and Carley, 2005)	$\Theta = (N, P, P_0)$
5.1.	Scale Free (Barabasi and Albert, 1999)	$\Theta = (N, P, N_0, P_0)$
5.2.	Scale Free (Airoldi and Carley, 2005)	$\Theta = (N, M, R)$
6.1.	Cellular (Airoldi and Carley, 2005)	$\Theta = (N, P, B, P_B, R)$

2.1.1 Experimental Evidence

Airoldi and Carley (2005) survey various flavors of graph topologies, or topology types, along with popular published algorithms that have been introduced to generate them; introduce novel algorithms that support a more diverse set of graphs, in terms the variability of their topological properties; and address the case of *cellular graph topology*. See Figure 2.1 for examples of the various topologies considered. Here, I briefly report on two statistical studies: (i) on the stability of topological properties of graphs to alternative algorithms that have been proposed to generate the same topology type; and (ii) on the separability of graphs with distinct topology types—characterized by a set of 18 metrics for the analysis of graphs, widely adopted in the social and physical sciences.

Stability of Topological Properties to Variations in the Algorithms The stability study is targeted towards the three most popular notions of random, small world, and scale free (also known as power-law) topologies. The overall plan is simple. First, for each of these three topologies, I will use the proposed algorithms to generate a collection of graphs, and I will label them accordingly to the specific algorithm variant that was used. Then, I will assess how well it is possible to distinguish graphs that were generated by algorithm variants with a batch of cross-validation

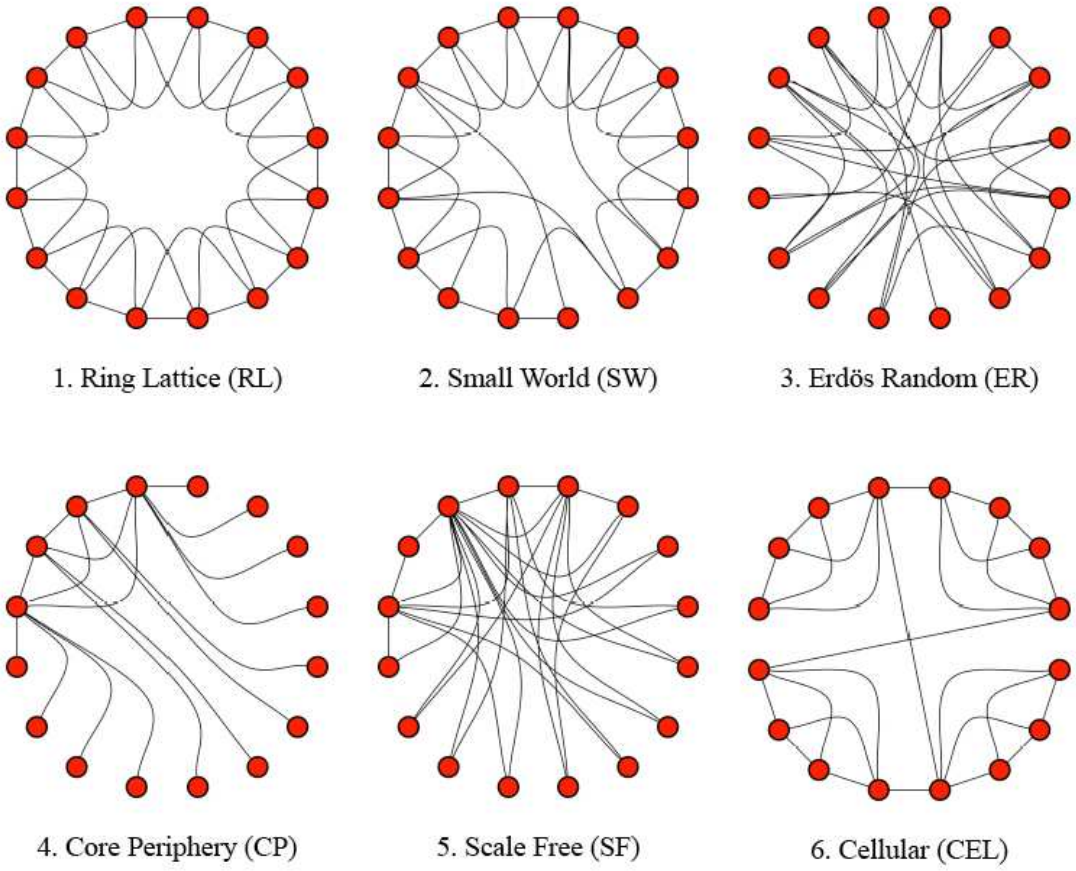

Figure 2.1: A glance at the relevant topologies on a ring. Note that in a ring there is a natural notion of distance that is distinct from the one entailed by shortest paths, i.e., the distance between nodes A and B is proportional to the arc-length that joins them, along the circle outlined by the ring.

experiments. I use a factorial experimental design; ten graphs were generated for each parameter configuration, and parameter configurations were set by defining a grid on the support of each parameter and then picking all combinations (Airoldi and Carley, 2005, Table 2). Graphs were generated among 250 nodes, and the choice of parameter configurations was further informed by controlling the density of edges. The rationale behind these choices is to make it hard for the classification algorithms to distinguish graphs based upon the scale and the density of the generated graphs. As a consequence, it is conceivable that any evidence of discriminatory power may be due to differential topological properties of the graphs generated by alternative algorithm variants.

I find that

(i) Extremal statistics (i.e., minimum and maximum) are good discriminators between the two algorithm variants that generate *random graphs*. The cross-validated accuracy is in the high 90% and this comes as no surprise given that Algorithm 3.1 of Table 2.1 leads to graphs with an exact density, whereas Algorithm 3.2 does not.

(ii) Properties of the degree distribution are good discriminators between algorithm variants that generate *scale free graphs*; the cross-validated accuracy is in the mid nineties; this can be explained by the non-negligible effect of the initial graph that is needed to initialize Algorithm 5.1, and it is consistent with the analysis of Bollobás and Riordan (2003).

(iii) Centrality of nodes and clustering coefficient are fairly good discriminators between algorithm variants that generate *small world graphs*; the cross-validated accuracy is in the mid eighties when we try to distinguish between sample graphs of Algorithms 2.1 and 2.2 or between Algorithms 2.2 and 2.3, and it drops to the high seventies when we try to distinguish between Algorithms 2.1 and 2.3. This is expected since Algorithms 2.1 and 2.3 lead to graphs with more variable neighborhood structures than Algorithm 2.2.

The published generative algorithms described in Table 2.1 play a role in this small experiment for which they were not designed. They were originally proposed to illustrate mechanisms of aggregation suggestive of social and artificial regularities underlying observed phenomena. Because of their popularity, however, the scientific community is adopting these mechanisms for analyses that are very different in scope from those they were intended for. I find this practice dangerous, both in terms of the reproducibility of the analyses, and in terms of the support such simplistic algorithms can offer to substantive conclusions. This is the message I mean to offer with the experiments presented here.

Table 2.2: Pairwise comparisons; entries quote the errors achieved in discriminating graphs generated according to pairs of algorithms. Errors are estimated using cross-validation. The size is fixed at 250 nodes; the density is controlled by design.

	Lattice	Random	Small World	Scale Free	Cellular	Core-Per.
Lattice	N/A	0.2700	0.0745	0.00	0.00	0.00
Random		0.00	0.4122	0.2794	0.3255	0.25
Small World			0.2478	0.0866	0.1312	0.0531
Scale Free				0.0007	0.2645	0.3333
Cellular					0.1746	0.3715
Core-Per.						0.50

Popular Notions of Graphs Topologies and their Topological Properties The separability study is motivated by another question related to data analysis. Often times, funding agencies and scholars[4] find it interesting to investigate the following: *given a collections of measurements on pairs of objects in a population of interest, what is the popular notion of topology that best represents such observations?* The question seems to imply, or assume, that topology types are uniquely defined in terms of few topological properties—those which the corresponding models are based upon. The goal of the separability study is to assess the extent to which this is an ill-posed question.

I performed two batches of experiments, whose results are reported in Tables 2.2 and 2.3. The first batch of experiments considered about 6000 graphs, generated according to the factorial experimental design used in the stability study, but extended to all the algorithms in Table 2.1. The size of all graphs was set at 250 nodes, and the density was controlled by design. The results are presented in Table 2.2, in the form of a pair-wise comparisons. Each entry gives the errors achieved in discriminating graphs generated according the corresponding pair of published algorithms—diagonal entries report the errors corresponding to the stability study discussed above. Errors were obtained with a naïve Bayes classifier; similar results were obtained with decision trees, support vector machines, and logistic-regression. The second batch of experiments considered about 40000 graphs,

[4] I omit citations here, although it is easy to identify notable examples.

Table 2.3: Pairwise comparisons and best three discriminators; entries quote the errors achieved in discriminating graphs generated according to published algorithms. Errors are estimated using cross-validation, with the quoted topological measures as the unique features. The size and density are variable, not controlled by design.

	1st Property	2nd Property	3rd Property
Random vs. Small World	Net. Constr. (max) 0.2740	Connectedness 0.2960	Net. Constr. (dev) 0.3800
Random vs. Scale Free	Eig. Cnt. (min) 0.3690	Close. Cnt. (min) 0.4040	Inv-Close. Cnt. (min) 0.4100
Random vs. Cellular	Eig. Cnt. (avg) 0.3110	Inv-Close. Cnt. (min) 0.3140	Close. Cnt. (min) 0.3170
Random vs. Core-Per.	Centrality (dev) 0.3470	Centrality (max) 0.3500	Close. Cnt. (dev) 0.3530
Small World vs. Scale Free	Net. Constr. (max) 0.2590	Eig. Cnt. (min) 0.2720	Eig. Cnt. (avg) 0.3410
Small World vs. Cellular	Net. Constr. (max) 0.1380	Connectedness 0.1750	Eig. Cnt. (avg) 0.2200
Small World vs. Core-Per.	Net. Constr. (max) 0.2400	Connectedness 0.2620	Centrality (max) 0.2870
Scale Free vs. Cellular	Centrality (max) 0.2860	Close. Cnt. (max) 0.2860	Inv-Close. Cnt. (max) 0.3060
Scale Free vs. Core-Per.	Centrality (dev) 0.3480	Close. Cnt. (dev) 0.3530	Connectedness 0.4170
Cellular vs. Core-Per.	Centrality (max) 0.2250	Inv-Close. Cnt. (max) 0.2520	Close. Cnt. (max) 0.2580

sampled from a pool of one million graphs generated according to the full factorial block-design of Frantz and Carley (2005b), which makes use of all the algorithms in Table 2.1. In the sample, the size of the graphs is controlled for, and so is the density. The results[5] are presented in Table 2.3. Each entry provides the discrimination errors achieved with a decision tree; similar results were obtained using a naïve Bayes classifier, support vector machines, and logistic regression.

The analysis of the results of both separability studies suggest that (i) the generative algorithms presented in Table 2.1 often lead to *unrealistic variability profiles* for specific metrics over a fairly large range of parameter values—either by design or by construction; (ii) as we consider the collections of graphs generated according to popular notions topology types in a large space of topological properties, *the boundary between pairs of topology types is not sharp*, and most of the graphs display mixed characteristics.

2.1.2 Discussion

The experimental evidence suggests that scientific questions about the data that rely on popular notions of graph topologies have to be treated carefully. Topology types are operationally defined by specific data generating processes that were devised to illustrate the effects of compelling aggregation mechanisms on a small set of topological properties of a graph. The proposed statistical analysis of such algorithms shows that (i) alternative options available for generating the same topology type are distinguishable in terms of the set topological properties they entail, and (ii) processes that supposedly lead to distinct topologies types, actually generate graphs that share many topological properties. *Therefore, while these algorithms deliver insights about phenomena of interest, it is dangerous to employ them for other purposes, as it is often done in practice.* In the context of statistical testing, for example, those algorithms may be used to produce p-values for metrics of interest. Topological properties of a graph under the null hypothesis can be evaluated

[5] I wish to thank Ian C. Fette for facilitating this study by sharing useful code.

by sampling graphs according to one of the popular algorithms listed in Table 2.1. Concluding, the experimental evidence suggests that (i) we need a larger set of topological properties to be able to characterize a graph exactly, e.g., along the lines of a representation theorem, and (ii) we need a richer set of statistical models for the purpose of data analysis.

An alternative approach to the scientific analysis of complex and dynamic graphs keeps the *topological properties* of *the data*—an observed collection of paired measurements—at the forefront of the analysis. Along these lines, statistical models of graphs with desirable topological properties, and their relation to the popular notions of topology, are explored in Section 2.2.

2.2 Exchangeable–Edge Models

A major issue with many of the popular algorithms that have been proposed in the literature to generate topologies of interest is that while they are meant to replicate local and global phenomena a procedure to estimate values for the constants that correspond to those phenomena, as well as a procedure to fit corresponding models to available data and assess their fit, are seldom specified. Being able to make a good use of the data is crucial whenever models and algorithms have to support substantive analyses and conclusions.

Example 15. *The US Army believes that the efficiency of communications during combat is directly correlated with the outcome of a battle. The efficiency in this context is defined in terms of a set of relevant network metrics, and being able to monitor these metric is the task of interest. In particular, it is crucial to be able to detect whenever communication patterns start displaying a level of variability that is considered abnormal, non-optimal, and ultimately dangerous. This problem can be stated formally as a statistical change-point problem, where detection has to occur in real-time, on a stream of data about communications in the network. In this context, a probabilistic model of a communication network,* $P(G_t|\Theta)$*, corresponds to a statistical model for a random*

variable, $P(X_t|\Theta)$, in the classical formulation. Statistical procedures that lead to estimates of the underlying constants, Θ, with desirable properties, e.g., consistency and unbiasedness, are critical for detecting deviations from normality, in both formulations.

Here, I discuss mathematical characterizations of a graph. Within this context, I show how it is possible to posit a general class of probabilistic models, which I refer to as *exchangeable-edge* models, that generate graphs with pre-specified topological properties of interest, and at the same time allow for *formal inference and estimation* procedures. I demonstrate the utility and flexibility of these models by introducing a novel analytical perspective of lognormal and cellular graphs.

Challenges to the Mathematical Characterization of Graphs The minimal representation of a graph G is given in terms of a set of vertices, $\mathcal{N}$, and a set of edges, $\mathcal{E}$, encoded by an adjacency matrix, Y, where the entry $Y(n,m) \in \{0,1\}$ encodes the presence or absence of the corresponding directed edge, $n \to m$. I make no distinction between *edges* and *edge weights* in the presentation below.

The matrix Y characterizes the topological properties of the graph it represents. For example, the degree of a binary graph (this simplifies things by requiring no distinction between in- and out-degree) is defined as a vector $\vec{d}_G$ with generic element

$$d_G(n) = \sum_m Y(n,m) \quad \text{for } n \in \mathcal{N}.$$

In general, the collection of eigenvalues, $\lambda_{1:N}$, and eigenvectors, $\vec{u}_{1:N}$, of Y give us an exact characterization of the graph as follows,

$$Y = \sum_{n \in N} \lambda_n \vec{u}_n \vec{u}_n^\top,$$

where N is the number of nodes in $\mathcal{N}$. In this sense, exact characterizations of a matrix provide us

with exact characterizations of a graph.

The following question is of interest; *what set of topological properties is necessary and sufficient to characterize the matrix Y, exactly?* Here is the challenge; from a statistical modeling perspective we would like to characterize a graph in terms of its essential topological properties. This would allow us to inform models of complex and dynamic graphs by analyzing such essential properties measured on observed graphs. On the other hand, the results of the previous section suggest that there is a disconnect between the popular classes of graphs for which generative models have been published and characteristics of their topological properties—as captured by existing metrics. It is not known whether a set of topological properties exist that exactly characterizes a graph, or, alternatively, what classes of graphs can be defined in terms of sets of constraints on a set of topological properties. These questions provide the context for the investigations that follow. I seek either exact or approximate characterizations, either for graphs with a finite set of nodes or in the infinite limit of large graphs.

2.2.1 Specifications and Likelihood

A first step is to extend the random graph models of Erdös and Rényi (1959) and Gilbert (1959) to include a set of latent variables. This will make the edges exchangeable, i.e., conditionally independent given values of these latent variables, rather than independent. The latent variables are themselves an IID sample from a common distribution. This extension allows me to reproduce the behavior of the original random graph models, and to induce a new array of interesting global behaviors such as those encoded in *lognormal graphs* (Section 2.2.2) and *cellular graphs* (Section 2.2.3). Furthermore, it is possible to write down and evaluate the likelihood corresponding to such models.

It is possible to generate a diverse set of graphs by means of exchangeable-edge models; a fairly

general class of statistical models of random graphs. An exchangeable-edge model of a graph is specified as follows,

$$Y(n,m) \,\big|\, \Theta \;\sim\; P_\Theta \tag{2.1}$$

$$\Theta \,\big|\, \mathcal{A} \;\sim\; P_\mathcal{A}, \tag{2.2}$$

for each element (n,m) of the matrix Y; where n, m are nodes in $\mathcal{N}$; Θ is a collection of latent variables; $\mathcal{A}$ is a collection of hyper-parameters; and $P_\Theta, P_\mathcal{A}$ are probability distributions. The likelihood can be written as follow,

$$\ell(Y|\mathcal{A}) = \int P_\mathcal{A}(\Theta|\mathcal{A}) \prod_{n,m} P_\Theta(Y(n,m)|\Theta)\, d\Theta. \tag{2.3}$$

In a more general formulation Y may be multivariate, e.g., may encode multivariate sociometric relations (Sampson, 1968), longitudinal, e.g., may encode a temporal sequence of graphs (Priebe et al., 2005), or even more complex.

Example 16. *In many applications, it is convenient to distinguish between latent variables that are partially, or potentially measurable and correspond to substantive concepts, e.g., tight groups of agents in the analysis of social networks. Following the discussion in Section 1.2, we may denote by* Ξ *the partially observable variables with a substantive interpretation, e.g., club membership, and by* Θ *other latent variables, e.g., propensity to participate to social activities. Latent variables in both these collections are agent-specific; namely,* $\Theta = \vec{\theta}_{1:N}$ *is a collection of agent-specific vectors whose components specify the grade of membership of agents to clubs, whereas* $\Xi = \xi_{1:N}$ *is a collection of agent-specific scalars that specify agents' propensities to socialize with members of other clubs. The corresponding exchangeable edge model for a sequence of graphs* $G_{1:T} = (Y_{1:T}, \mathcal{N})$*, that encodes social interactions recorded over* T *weeks among the same set of agents,*

$\mathcal{N}$, takes the form of a hierarchical mixed-membership model. At time t, it posits that

$$Y_t(n,m) \mid (\xi_{nm}, \theta_{nm}) \sim P_{(\xi,\theta)} \tag{2.4}$$

$$\xi_{nm} \mid \mathcal{A} \sim P_{\mathcal{A}} \tag{2.5}$$

$$\theta_{nm} \mid \mathcal{B} \sim P_{\mathcal{B}}, \tag{2.6}$$

for each observed social interactions (n,m) recorded in the matrix Y_t during the t-th week; where n, m are agents in $\mathcal{N}$; ξ_{nm} denotes (ξ_n, ξ_m); θ_{nm} denotes (θ_n, θ_m); and $P_{(\xi,\theta)}, P_{\mathcal{A}}, P_{\mathcal{B}}$ are probability distributions. Figure 2.2 illustrates the graphical representation of this model, using plates. Note that, in this model, both the degree of membership of agents to clubs, and the propensity to socialize with other agents do not change from one week to the next. Furthermore, the grade of membership and the propensity to socialize are non-observable, competing explanations of the observed interactions.

The main advantage of these models is that edges are exchangeable, that is, weakly dependent [6] rather than independent. Working within the skeletal specifications of Equations 2.1 and 2.2, we can introduce layers in the hierarchy and posit stochastic block models (Airoldi et al., 2006d, and Section 3.1), latent space models (Hoff et al., 2002), and diffusion models (Coifman et al., 2005a,b, and Section 3.2). In Chapter 5, I discuss how to introduce flavors of dynamics and evolution in a few cases—temporal models will break the exchangeability of edges within a graph and introduce different dependence structures.

[6] De Finetti's theorem implies exchangeable edges can be characterized as being independent conditionally on a collection of latent variables (Schervish, 1995).

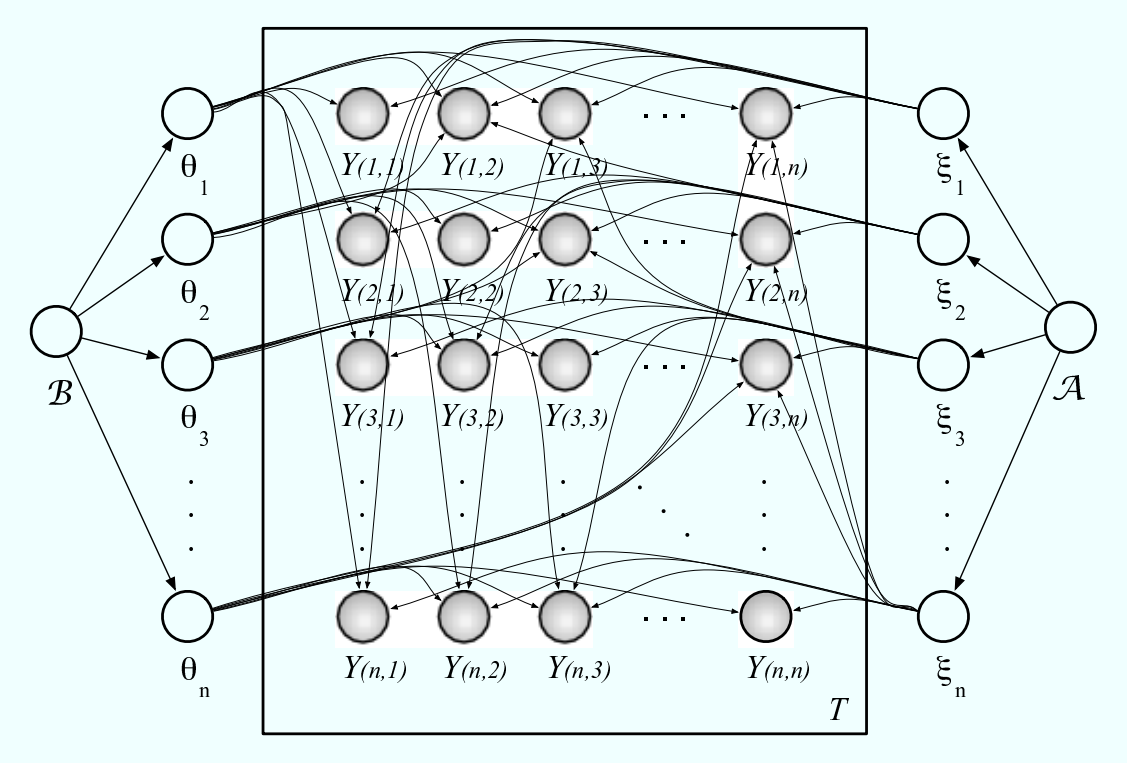

Figure 2.2: Graphical representation of the exchangeable edge model in Example 16.

2.2.2 Lognormal Graphs

Researchers in a diverse set of disciplines have offered evidence and arguments that ultimately purport the power-law distribution as an inevitable regularity of natural and artificial phenomena alike. In this section, I investigate the basis for such claims.

Main Argument—Part I: Building Association Graphs Consider the following exchangeable-edge generating process for a (binary, undirected) graph $G = (Y, \mathcal{N})$:

Algorithm A1

1. for $n \in \mathcal{N}$

2. for $k = 1, \ldots, K$

3. $b_n(k) \sim Bernoulli\ \big(\ p\ \big)$

4. for $n, m \in \mathcal{N}$

5. $Y(n,m) \sim Dirac\ \big(\ f(\vec{b}_n, \vec{b}_m) > \tau\ \big)$

The constants underlying Algorithm A1 are $\Theta = (K, p, f, \tau)$, where K is the number of node-specific attributes, $b_n(k)$; p is the probability that any one of the attributes is present; $f(\cdot, \cdot)$ is a measure of similarity; and τ is the f-similarity threshold beyond which two nodes are set as connected. Algorithm A1 replicates a common network construction process.[7] In such a scenario, the node-specific collections of $\vec{b}_n$ represent noisy measurements about those aspects of a node that matter when decisions about the presence of a tie with other nodes have to be made.

Below I explore the degree distribution of an association graph that arises in a case where there is no association structure among nodes, in terms of their multivariate attribute representations, $\vec{b}_{1:N}$. In order to derive the degree distribution I completely specified Algorithm A1 by setting the size of the graph to 100 nodes; I set the number of latent aspects, K, to ten, and the probability that any one of the attributes is present, p, to 0.5; in different experiments I explored various measures of association based on Pearson's correlation coefficient, few of many possible choices for f; last, in order to find (purportedly) significant associations, I set the f-threshold, τ, to be the 95-th percentile of the observed associations—about 5000 for each graph. I then sampled many graphs. As an example, the limiting average degree distribution of one of the graphs is shown in Figure 2.3, on a log–log scale; the corresponding matrix, Y, is shown in Figure 2.4. The results of the simulations consistently suggest a quadratic relation between nodes' degree and their frequency, on the log–log scale. That is, the simulated association graphs have a lognormal degree distribution, whenever no association exist among their attribute representations, $\vec{b}_{1:N}$.

[7] An alternative formulation of Algorithm A1 posits distinct attribute-specific probabilities for each node, $p = p_n(k)$, which are sampled independently from a standard Gaussian distribution and then projected into the $[0, 1]$ interval. The attributes, $b_n(k)$, are then independent samples from a Bernoulli, conditionally on such probabilities.

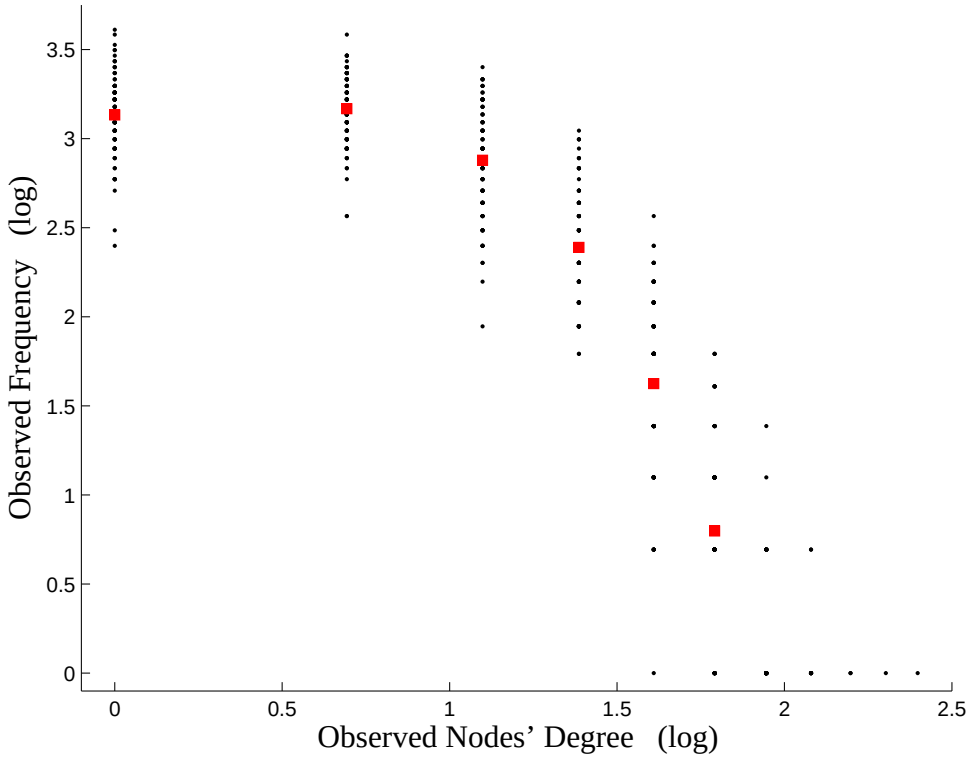

Figure 2.3: Observed degrees versus frequency of nodes with such a degree, over a set of 100 graphs sampled according to Algorithm A1. Squares correspond to averages.

Alternative measurement schemes exist, which are based upon alternative measures of association that may be adopted to decide when to establish an edge. However, the empirical results are robust to such alternatives. This empirical observation may be formalized by looking at the distribution of the association measure f as follows.

Conjecture 1. *Any strategy that builds a graph by thresholding a measure association, X, will induce a degree distribution proportional to the tail of the probability density of P_X, where the tail is defined by the threshold τ.*

The general mechanism through which probability statements about pairs of nodes, e.g., in terms of the correlations $Pr\ (r(n,m) > \tau)$, translate into probability statements about the individ-

Figure 2.4: The matrix $Y_{(100\times100)}$ of a lognormal graph sampled via Algorithm A1.

ual nodes, e.g., in terms of the degree $Pr\ (d(n) \leq k)$, is fairly simple. Consider the degree of a node, n, in a graph generated via Algorithm A1,

$$d(n) = \{\#\text{ nodes } m \text{ such that } r(n,m) > \tau\}$$

At a first degree of approximation, $d(n)$ follows a Binomial distribution with parameters $N = |\mathcal{N}|$, i.e., the number of nodes in the graph, and, $p = Pr\ (r(n,m) > \tau)$, i.e., the probability of imputing an edge. The heavy-tail of the degree distribution follows from the fact that the probabilities of imputing edges, $p_{1:N}$, are different for the various nodes.

As we restrict the focus to those thresholding schemes that are based upon Pearson's correlation coefficient, r, we can derive more precise results. Table 2.4 describes four measurement schemes, based on different functions of r, in terms of the probability density function they induce on the measure of association, $f(r)$, and of the range of possible values for f, i.e. the support. Figure 2.5

shows the estimated probability density functions corresponding to these four association measures based upon Pearson's correlation coefficient describes in Table 2.4. Define r_N to be Pearson's correlation coefficient, computed on a sample of size N of paired measurements (X_n^1, X_n^2). The asymptotic distribution $(N \to \infty)$ of r_N is bell-shaped, left-skewed. Fisher (1915) derived the distribution of Pearson's correlation coefficient when the data are bivariate Gaussian. In subsequent work he showed that the transformation,

$$\zeta = \log \sqrt{\frac{1+r}{1-r}}, \quad r \in [-1, +1] \tag{2.7}$$

is useful in stabilizing the skewness of r_∞, and induces a Gaussian distribution on the support of ζ, unbiased for $\sigma_{(X^1,X^2)}$, with variance $(N-3)^{-1/2}$ (Fisher, 1921).

Concluding, experimental evidence suggests that graphs with a heavy-tailed degree distribution may be generated as artifacts of the way we measure association, e.g., by thresholding, even in those situations where no real association exist. Observing a heavy-tailed degree distribution in a graph should not be regarded as interesting substantive finding in the absence of a deeper analysis.

Main Argument—Part II: Limiting Graph Structures Consider the behavior of graphs as new nodes and edges appear. In the limit of large graphs (i.e., many nodes), and for a wide range of aggregation regimes (i.e., different edge addition rules), lognormal graphs are an attractor. Their basin of attraction is large. Below, the limiting behavior of scale-free graphs (Barabasi et al., 1999;

Table 2.4: Characteristics of four measurement schemes to establish associations, by thresholding, based upon Pearson's correlation coefficient.

Measure	Support	PDF	Notes
r	$[-1, +1]$	bell-shaped, left-skewed	See Fisher (1915)
ζ	$(-\infty, +\infty)$	Gaussian	See Anderson (1996)
r^2	$[0, +1]$	bell-shaped, right-skewed	
e^ζ	$[0, +\infty)$	lognormal	

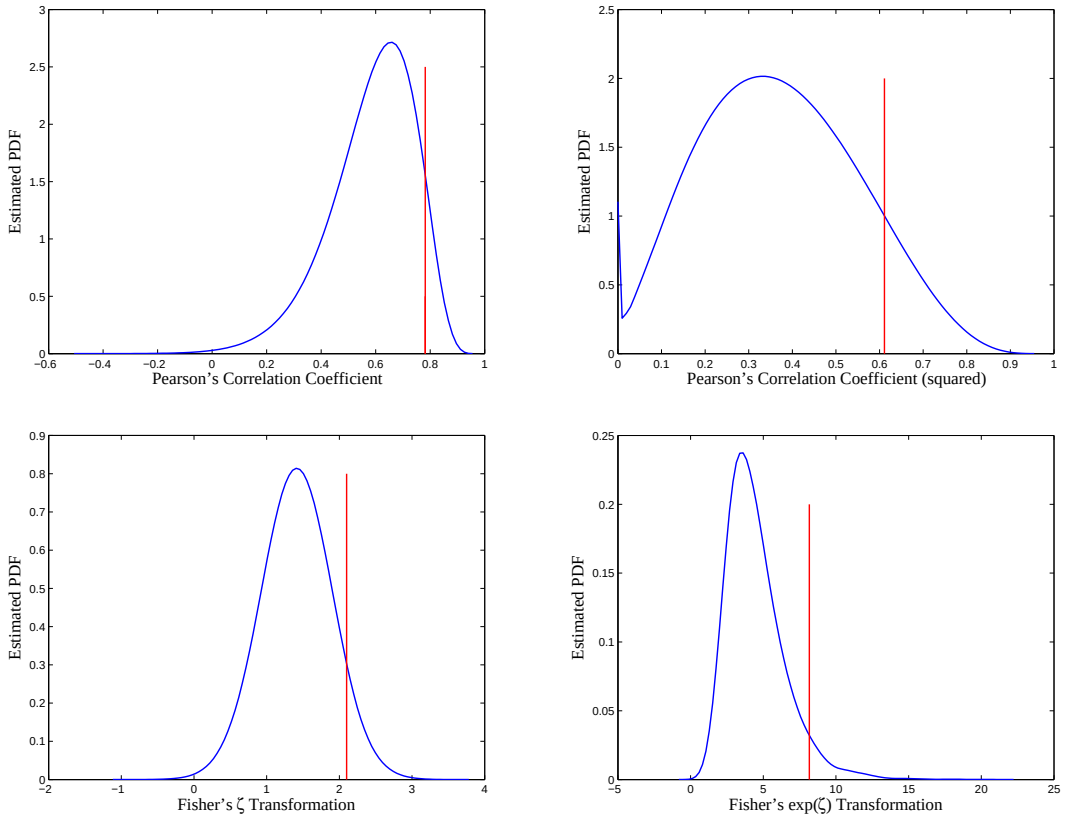

Figure 2.5: Estimated probability density functions corresponding to the four association measures based upon Pearson's correlation coefficient describes in Table 2.4. The vertical (red) bar indicates the 95-th percentile; e.g., r_0 s.t. $P(R \leq r_0) \geq 0.95$.

Faloutsos et al., 1999; Huberman and Adamic, 1999) is revisited in the light of this remark, in the larger context of statistical convergence. Consider the following algorithm.

Algorithm A2

1. start with $G = (Y = 0, \mathcal{N} = \emptyset)$
2. repeat
3. add node n to $\mathcal{N}$
4. sample its degree $d(n) \sim P(\Theta)$

5. connect node n to $d(n)$ existing nodes in $\mathcal{N}$ with equal probability

To illustrate the main point, it is enough to note that Algorithm A2 entails an *expected rate* of change of the degree of a node, $d^{t+1}(n)/d^t(n)$, constant over time. In this case,

$$\begin{aligned} k^{t+1}(n) &= k^t(n) \cdot d^t(n) && (2.8)\\ \log k^{t+1}(n) &= \log k^0(n) + \Sigma_{s=1}^t \log d^s(n) && (2.9)\\ k^{t+1}(n) - k^t(n) &= k^t(n) \cdot (d^t(n) - 1) && (2.10)\\ \Delta k^{t+1}(n) - k^t(n) &= k^t(n) \cdot \tilde{d}^t(n). && (2.11) \end{aligned}$$

In the algorithm proposed by Barabasi et al. (1999) the *rate* of change of the degree is the quantity that remains constant over time, so that

$$\mathbb{E}_\Theta[\tilde{d}^t(n)] = \frac{A}{m_0\, t + k^0}$$

It is possible to show the connection in a different way, by looking at the limiting degree distribution of graphs where the *expected rate* of change of the degree is decreasing over time,

$$\mathbb{E}_\Theta[\tilde{d}^t(n)] = O(t^{-\alpha}), \quad \alpha > 0.$$

Remark 1. *Power-law graphs (also referred to as scale-free) provide a first-order approximation to lognormal graphs, in terms of their degree distribution.*

Consider the density of the degree distribution of a lognormal graph,

$$p\left(\, x \,\middle|\, \mu, \sigma^2 \,\right) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma^2}\left(\log(x)-\mu\right)^2}, \quad \mu, \in \mathbb{R}, \text{ and } x, \sigma \in \mathbb{R}^+. \qquad (2.12)$$

This entails a quadratic relation between log density and log degree,

$$\begin{aligned}\log p\left(\, x \,\middle|\, \mu, \sigma \,\right) &= -\log(x) - \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\left(\, \log(x) - \mu \,\right)^2 \\ &= -\log(\sigma\sqrt{2\pi}) - \frac{\mu^2}{2\sigma^2} - \left(1 - \frac{2\mu}{2\sigma^2}\right)\log(x) - \frac{1}{2\sigma^2}\log^2(x).\end{aligned}$$

Taylor-expand $\log p$, with respect to $\log x$, to the first order,

$$\log p\left(\, x \,\middle|\, \mu, \sigma \,\right) \approx \log p\left(\, x_0 \,\middle|\, \mu, \sigma \,\right) + \frac{\mu - x_0}{\sigma^2}\left(\, \log(x) - x_0 \,\right).$$

If we choose $x_0 = \mu$, the mean, the expression above simplify to

$$\log p\left(\, x \,\middle|\, \mu, \sigma \,\right) \approx -\log(x) - \log(\sigma\sqrt{2\pi}),$$

which implies a linear relation between log density and log degree, that is, the relation underlying the degree distribution of a power-law graph.

Thus we see that exchangeable-edge models illustrate how lognormal graphs arise in two interesting sets of circumstances: (i) as an artifact of the way we measure associations; and (ii) in the limit, as a consequence of multiplicative aggregation processes. Inasmuch as they approximate lognormal graphs, power-law graphs may arise in the same circumstances. This fact may help to explain their ubiquity.

2.2.3 Cellular Graphs

Despite the fact they are so pervasive, no characterization exists that explains *why* cellular networks arise as a structure of collective organization. What are the conditions that naturally conjure cellular structures among individuals? Below, I introduce a simple exchangeable-edge model that

suggests a possible answer.

I posit a stylized model of a population of agents with limited resources. In particular, I assume that one of the resources (e.g., time) is instrumental in acquiring other resources (e.g., knowledge). Individual agents are endowed with a limited amount of the former (e.g., hours in a day). This limitation imposes choices on the agents about how to allocate the instrumental resource. In a model with time as the limited resource, for example, the choices to be made by the agents may concern which interests to cultivate. Given the set of individual choices, the network among agents emerges as a consequence of a simple social aggregation process that induces ties between pairs of agents with similar interests.

Algorithm A3

1. for $n \in \mathcal{N}$
2. $\vec{p}_n \sim$ logit $MVN\left(\vec{0}, \Sigma(\alpha)\right)$
3. for $k = 1, \ldots, K$
4. $b_n(k) \sim Bernoulli\left(p_n(k)\right)$
5. for $n, m \in \mathcal{N}$
6. $Y(n,m) \sim Dirac\left(f(\vec{b}_n, \vec{b}_m) > \tau\right)$

The constants underlying Algorithm A2 are $\Theta = (K, f, \alpha, \tau)$, where K is the number of node-specific attributes, $b_n(k)$; $f(\cdot,\cdot)$ is a measure of similarity; τ is the f-similarity threshold beyond which two nodes are set as connected; and α is the *exclusivity parameter* that I shall now discuss.

Algorithm A2 is fairly similar to the algorithm I used to generate lognormal graphs. The N agents, i.e., the nodes, are associated with binary strings, $\vec{b}_{1:N}$, whose components indicate the presence or absence of a specific interest, out of K possible. There are two changes here; (i) the

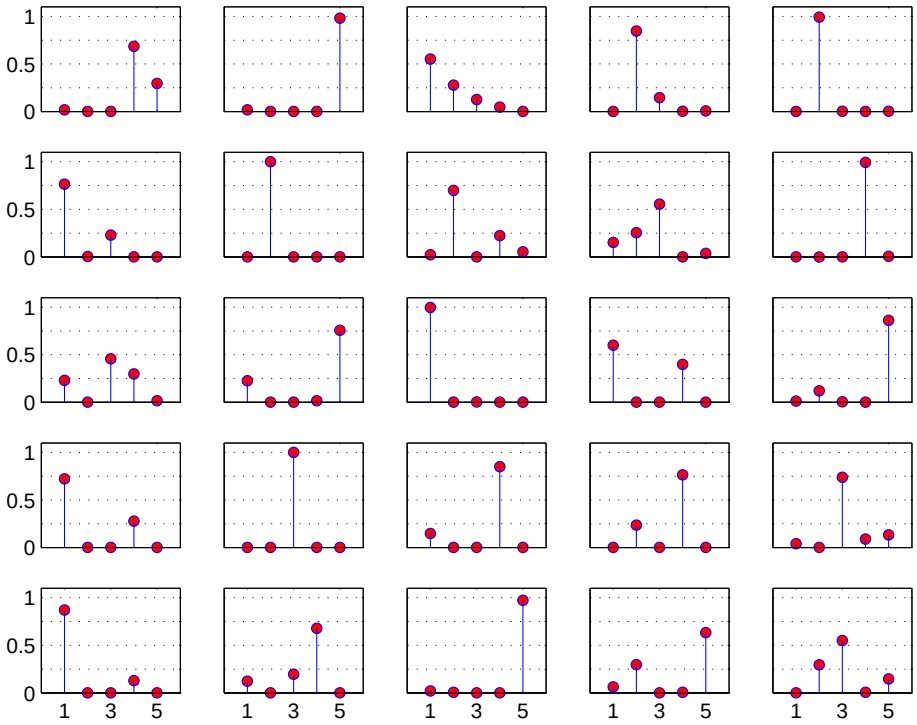

Figure 2.6: Example interest vectors for 25 nodes.

fact the elements $b_n(k)$ are sampled with agent-specific probabilities; and (ii) the agent-specific probability vectors, $\vec{p}_n$, have weakly dependent components—this is crucial. Specifically, the variance-covariance matrix $\Sigma = \Sigma(\alpha)$ has entries

$$\sigma_{nn} = \frac{\alpha^2(K-1)}{(K\alpha)^2(K\alpha+1)}, \quad \text{and} \quad \sigma_{nm} = \frac{-2\alpha}{(K\alpha)^2(K\alpha+1)}. \tag{2.13}$$

specified in terms of the scalar parameter α. The moments of the multivariate normal distribution in step 2 of Algorithm A3 are reminiscent of those of a Dirichlet distribution. The main difference between the the multivariate normal and the Dirichlet is that the support of the former is the unit

Figure 2.7: The matrix $Y_{(100\times100)}$ of a cellular graph sampled via Algorithm A3.

hyper-cube after a logistic transformation of its coordinates,

$$p_n(k) = \frac{\exp\{z_k\}}{1 + \sum_k \exp\{z_k\}}, \quad \text{for } i = 1...K,$$

whereas the support of the latter is the K-dimensional simplex. The variance-covariance structure $\Sigma(\alpha)$ enforces what I term *exclusivity of interests* in the following sense; *by dedicating time to acquire a specific interest, agents implicitly choose not to pursue other interests*. The parameter α has support in $(0, \infty)$; the closer α is to zero, the stronger it promotes exclusivity. Furthermore, $\alpha < 1$ promotes exclusivity, whereas $\alpha > 1$ implies that agents are likely to devote an equal amount of time to each one of the K available interests.

Figure 2.6 show an example of interest vectors, $\vec{b}_n$, for 25 nodes, in a simulation where the number of interests, K, is five. Figure 2.7 shows one of the cellular graphs generated by Algorithm A3. The parameters were: 100 nodes, $K = 5$, f is Pearson's correlation coefficient, and τ is the 75-th percentile of the observed associations. Figure 2.8 shows an aspect of the emergence of

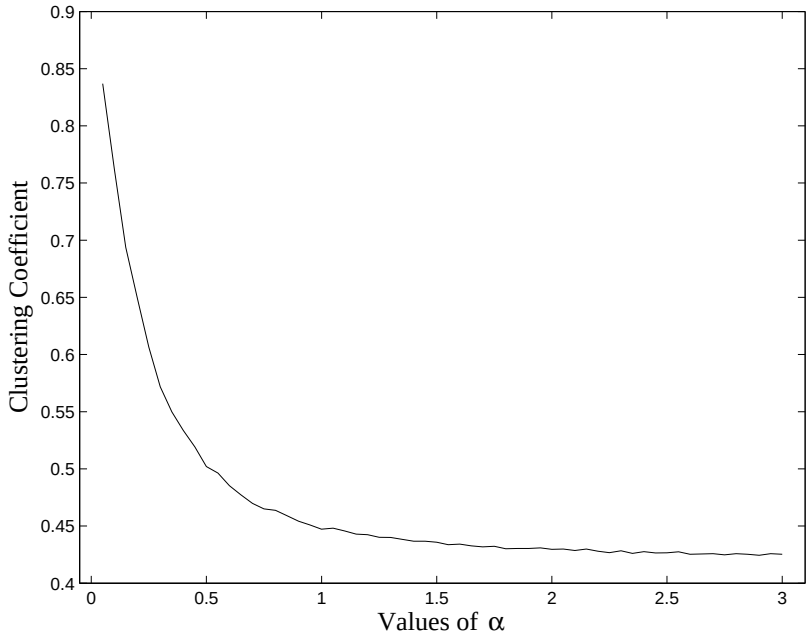

Figure 2.8: The clustering coefficient as a function of α obtained with Algorithm A3.

communities induced by Algorithms A3. In particular, a sharp drop in the clustering coefficient occurs as the exclusivity parameter α increases from ≈ 0 towards 1; when $\alpha \geq 1$ the community structure is no longer present.

By using Algorithm A3 I show how communities form because of the joint effect of two simple factors: (i) exclusivity, that is, the need for allocating resources to competing interests—which may be induced by the finiteness of such resources; and (ii) homophily, that is, the fact that social interactions are more likely to occur between individuals who share interests than between those who do not. Communities emerge, quickly, as $\alpha < 1 \rightarrow 0$.

2.3 Convex Generation of Graphs with Degree Constraints

Here I express the problem of *finding the most likely graph with a given, arbitrary degree distribution* as a convex optimization problem.

There are many situations where being able to sample graphs with a given arbitrary degree distribution is crucial. One of the main issues with the existing algorithms is their failure rate, that is, the number of times such algorithms have to be restarted in order to produce a graph that satisfies the given constraint on the degree distribution (Milo et al., 2004c) . Expressing the problem of finding likely graphs that satisfy the constraint on the degree distribution as a convex optimization problem will resolve this issue. In fact, once we have a starting point in a high probability region of the space of feasible graphs, sampling graphs at random that satisfy a given constraint is easier—leads to a much lower failure rate.

Consider a undirected, unipartite graphs, $G = (Y, \mathcal{N}) \in \mathcal{G}$. A basis for $\mathcal{G}$ is given by the collection of graphs $E_{nm} = (Y_{nm}, \mathcal{N})$, indexed by pairs of nodes (u, v) in $\mathcal{N}$. The element (i, j) of the adjacency matrix of E_{nm} is defined as,

$$Y_{nm}(i,j) = \begin{cases} 1 & \text{if } (i = n, j = m) \text{ or } (i = m, j = n) \\ 0 & \text{otherwise.} \end{cases}$$

It is then possible to write the problem of finding a graph, Y, with a pre-specified degree sequence, $\vec{d}$, as follows:

$$\begin{aligned} \text{opt}_{\vec{\alpha}} \quad & Y = \sum_{e \in \mathcal{E}_G} \alpha_e \, Y_e \\ & \text{s.t. } \; Y \cdot \vec{1} = \vec{d} \text{ and } \; \alpha_e \in \{0; 1\} \text{ for all } e, \end{aligned} \tag{2.14}$$

where $\mathcal{E}_G$ is the set of all possible edges among pairs of nodes of G. Simulated annealing, for exam-

ple, can be used to find the unique solution. Problem 2.14 is formulated as a convex optimization (Boyd and Vandenberghe, 2004).

* * *

In this chapter, I introduced exchangeable-edge models of graphs. I argue that:

- they represent an important extension of the popular random graph model of Erdös and Rényi (1959) and Gilbert (1959), technically, by involving a layer of latent variables;
- they are proper statistical (Bayesian) models, in the sense that we can write down and evaluate their likelihood, and can therefore be used for principled analyses of data.

To substantiate these claims, I presented a novel analysis of lognormal and cellular networks based upon them. Methodology for statistical network analysis is presented in Chapter 3 as well. The development of exchangeable-edge models has led me to the useful formulation of the problem of sampling graphs with given degree constraints as a convex optimization problem.

Chapter 3

Discovering Latent Patterns

In this Chapter, I work within the general formulation of exchangeable-edge models of Section 2.2 to specify Bayesian mixed-membership models of random graphs that are used to discovery latent patterns.

Introduction and Motivation Statistical models are used to inform scientific analyses of graphs and networks that encode observations about phenomena of interest (Holland and Leinhardt, 1975; Wasserman, 1980; Fienberg et al., 1985; Wasserman and Pattison, 1996; Watts and Strogatz, 1998; Cooper and Frieze, 2003; Kemp et al., 2006). Often, we can specify models via probabilistic algorithms that generate nodes and/or edges in a hierarchical fashion, starting from a small set of underlying constants. Specifications of hierarchical dependencies among such constants, other non-observable quantities possibly generated in intermediate steps, and the data provide a channel to inform the analysis with structural assumptions that are relevant for a specific application. For instance, in social networks analysis the *social context* in which actors interact, or the *group* which actors are members of, are examples of such non-observable (or partially observable) quantities that may be useful to explain, e.g., email communications among a group of employees within a

company, or interactions among a set of proteins in a certain tissue under specific experimental conditions.

The three main approaches proposed in the social, mathematical and computing sciences literatures, that make use of non-observable quantities[1] to express specific concepts relevant to an application domain, as (i) latent space models, (ii) block models, and (iii) diffusion models. More in detail, in (i) the N nodes in the graph are projected onto a latent space, Θ, in a way that edges are preserved whenever the distance of their projections is high enough, e.g., $d(\theta_n, \theta_m)$ exceeds a threshold. In (ii) the graph is summarized in terms of a noisy block structure, B, along with the memberships of nodes to blocks, $\pi_{1:N}$, as detailed in Section 3.1. In (iii) mathematical functional defined on the graph are proposed to study the diffusion process of real or informational artifacts among the nodes. I will focus on extending block models for a major portion of this chapter. I will then revisit latent space models and diffusion models towards the end of the chapter.

3.1 Admixture of Latent Blocks Model

Relational information arise in a variety of settings, e.g., in scientific literature papers are connected by citation, in the word wide web the webpages are connected by hyperlinks, and in cellular systems the proteins are often related by physical protein-protein interactions revealed in yeast-two-hybrid experiments. These types of relational data violate the assumptions of independence or exchangeability of objects adopted in many conventional analyses. In fact, the relationships themselves between objects are often of interest in addition to the object attributes. For example, one may be interested in predicting the citations of newly written papers or the likely links of a web-page, or in clustering cellular proteins based on patterns of interactions between them.

[1] A mainstream approach to statistical network analysis that (for the most part) does not make use of latent variables is presented in the book by Wasserman and Faust (1994).

In many such applications, clustering the objects of study or projecting them in a low dimensional space (e.g., a simplex) is only one of the goals of the analysis. Being able to estimate the relational structures among the clusters themselves is often as important as object clustering. For example, from observations about email communications of a study population, one may be not only interested in identifying groups of people of common characteristics or social states, but also at the same time exploring how the overall communication volume or pattern among these groups can reveal the organizational structures of the population. Furthermore, in modern network analysis tasks described above, it is desirable to also relax the unary-aspect assumption on each node imposed by extant models. To this extent, I introduce a new class of models based the principle of *stochastic block models of mixed membership*, which combines features of the mixed-membership models (Erosheva and Fienberg, 2005) and the block models (Holland et al., 1983; Anderson et al., 1992; Nowicki and Snijders, 2001; Doreian et al., 2004) via a hierarchical Bayesian framework, and offers a flexible machinery to capture rich semantic aspects of various network data—see Section 4.2.2 for a general formulation.

Below, I describe an instantiation of this class of models, referred to as *admixture of latent blocks* (ALB) to reasons to be explained shortly, for analyzing networks of objects with multiple latent roles, e.g., social activities in case the objects refer to people (Airoldi et al., 2007b), or biological functions in case the objects refer to proteins (Airoldi et al., 2006c). As mentioned above, classical network models such as the stochastic block models only allow each nodes to bear a single role. Our model alleviates this constraint, and furthermore posits that each nodes can adopt different roles when interacting with different other nodes. In Section 4.2 of Chapter 4, I will describe the general model formulation for multivariate relations along with the general model formulation for multivariate attributes.

Historical Notes A popular class of probabilistic models for relational data analysis are based on the stochastic block model (SBM) formalism for psychometric and sociological analysis pioneered

by Holland and Leinhardt (1975), and later extended in various contexts (Fienberg et al., 1985; Wasserman and Pattison, 1996; Snijders, 2002; Hoff et al., 2002; Doreian et al., 2004). In machine learning, Markov random networks have been used for link prediction (Taskar et al., 2003) and the traditional block models have been extended to include nonparametric Bayesian priors (Kemp et al., 2004, 2006) and to integrate relations and text (McCallum et al., 2007). Typically, these models posit that every node in a study network is characterized by a unary *latent aspect* that accounts for its interaction patterns to peers in the networks; and conditioning on the observed network topology one can reason about these *latent aspects* of nodes via posterior inference. These formulations are closely related to the one introduced here.

Largely disjoint from the network analysis literature, methodologies for latent aspect modeling have also been widely investigated in the contexts of different informational retrieval problems concerning modeling the high-dimensional non-relational attributes such as text content or genetic-allele profile. In many of these domains, variants of a mixed membership formalism have been proposed to capture a more realistic assumption about the observed attributes, that the observations are resulted from contributions from multiple latent aspects rather than a unary aspects as assumed in most extant network models such as SBM. The mixed membership models have emerged as a powerful and popular analytical tool for analyzing large databases involving text (Blei et al., 2003), text and references (Cohn and Hofmann, 2001; Erosheva et al., 2004), text and images (Barnard et al., 2003), multiple disability measures (Erosheva and Fienberg, 2005; Manton et al., 1994), and genetics information (Rosenberg et al., 2002; Pritchard et al., 2000; Xing et al., 2003c). These models often employ a simple generative model, such as a bag-of-words model or a naive Bayes, embedded in a hierarchical Bayesian framework involving a latent variable structure that combines multiples latents aspects. This scheme induces dependencies among the objects' relational behaviors in the form of probabilistic constraints over the estimation of what might otherwise be an extremely large set of parameters.

3.1.1 Goals of the Analysis

I am concerned with modeling data represented as a collection of directed unipartite graphs. A unipartite graph is a graph whose nodes are of a single type, e.g., individual human beings in case of a person-to-person communication network, as opposed to bipartite and multipartite graphs, where the nodes are of two or multiple types, e.g., genes-to-experiments (Blei et al., 2003; Airoldi et al., 2006f) or employees-to-tasks-to-resources (Carley, 2002). Consider a collection of unipartite graphs whose edges encode measurements on pair of nodes about a response variable. Multiple graphs encode replicates, M, of the same relation. Denote the collection of graphs by $\mathcal{G} = \{G_m : r = 1, \ldots, M\}$, where each graph, G_m, is defined over a common set of nodes, $\mathcal{N}$. The random variables that encode edge weights are denoted by $R_m(p, q)$, where (p, q) is a pair of nodes in $\mathcal{N}$.

Example 17. *Sampson (1968) described a collection of relationships measured among a group of monks in a monastery. He observed responses about typically asymmetric relations such as "Do you like monk X?", at a sequence of epochs. This information is representable as a collection of M graphs, where the edges encode,the binary "like" responses.*

Example 18. *Mewes et al. (2004) describe the set of hand curated protein interactions produced by the Munich Institute for Protein Sequencing. A single set of interactions between proteins has been experimentally verified. This information is representable as a single graph where the random variables associated with the edges are binary. See the reanalyses in (Airoldi et al., 2006c) for further details.*

The analysis of such data typically focuses on the following objectives: (1) identifying clustering of nodes; (2) determining the number of clusters; and (3) estimating the probability distribution of interactions among actors within and between clusters. For instance, in the monestary social network of Example 17, objective 1 translates to identifying the solid factions among monks, In

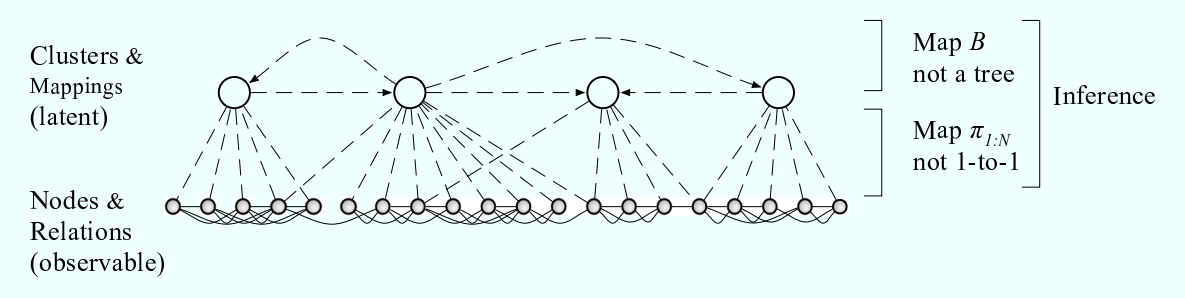

Figure 3.1: The scientific problem at a glance. The goal of the analysis is to make inference on two mappings; nodes-to-clusters (via $\vec{\pi}_{1:N}$) and clusters-to-clusters (via B). The facts that B does not necessarily encode a tree, and that $\vec{\pi}_{1:N}$ is not necessarily one-to-one distinguish our formulation from typical hierarchical and hard clustering.

addition one wants to determine how many factions are likely to exist in the monastery, and how the factions relate to one another. Typically, unsupervised learning experiments are performed, or semi-supervised learning experiments with minimal information available in terms of membership of, say, monks to factions. Working in the hierarchical Bayes framework, we can either specify the constants underlying the distribution of random quantities at the top level of the hierarchy (i.e., the hyper-parameters) or estimate them via empirical Bayes methods. This methodology accommodates hypothesis testing about the existence of specific relational structure among clusters.

3.1.2 Model Specifications

The approach detailed below employs a hierarchical Bayesian formalism that encodes statistical assumptions underlying a network generative process. This process generates the observed networks according to the latent distribution of the hypothetical group-involvement of each monk, as specified by a mixed-memembership multinomial vector $\pi := [\pi_1, \ldots, \pi_K]'$ where π_i denotes the probability of a monk belonging to group i; and the probabilities of having interactions between different groups, as defined by a matrix of Bernoulli rates $B_{(K\times K)} = \{B_{ij}\}$ where B_{ij} represents the probability of having a link between a monk from group i and a monk from group j. Each monk is associated with a unique π, meaning that he can be simultaneously belonging to multi-

ple groups, and the degree of involvements in different groups is unique for each monk; and π of different monks independently follow a Dirichlet distribution parameterized by α.

More generally, for graph m and each node, let indicator vector [2] $\vec{z}^{m}_{p\to q}$ denote the group membership of node p when it is to approach with node q; let $\vec{z}^{m}_{p\leftarrow q}$ denote the group membership of node q when it is approached by node p; let $N := |\mathcal{N}|$ denote the number of nodes in the graph; and let K denote the number of distinct groups a node can belong to. An admixture of latent blocks (ALB) model posit that a sequence of M networks, $G_{1:M} = (R_{1:M}, \mathcal{N})$, can be instantiated according to the following procedure:

Algorithm $A1 : (\mathcal{N}, M, K, \vec{\alpha}, B) \to R_{1:M}$.

1. For each node $p \in \mathcal{N}$

   1.1. Sample $\vec{\pi}_p \sim Dirichlet\ (\vec{\alpha})$.

2. For each interaction network $m = 1, \ldots, M$

   2.1. For each pair of nodes $(p, q) \in \mathcal{N} \otimes \mathcal{N}$

      2.1.1. Sample group $\vec{z}^{m}_{p\to q} \sim Multinomial\ (\vec{\pi}_p, 1)$

      2.1.2. Sample group $\vec{z}^{m}_{p\leftarrow q} \sim Multinomial\ (\vec{\pi}_q, 1)$

      2.1.3. Sample $R_m(p, q) \sim Bernoulli\ (\vec{z}^{m\ \top}_{p\to q} B\ \vec{z}^{m}_{p\leftarrow q})$

It is noteworthy that in the above model, the group membership of each node is *context dependent*, that is, each nodes can assume different membership when interacting to or being interacted by different peers. Therefore, each node is statistically an admixture of group-specific interactions, and I denote the two sets of latent group indicators corresponding to the m-th observed network

[2] An indicator vector of memberships in one of the K groups is defined as a K-dimensional vector of which only one element whose index corresponds to the id of the group to be indicated equals to one, and all other elements equal to zero.

by $\{\vec{z}^m_{p\rightarrow q} : p, q \in \mathcal{N}\} =: Z^{\rightarrow}_m$ and $\{\vec{z}^m_{p\leftarrow q} : p, q \in \mathcal{N}\} =: Z^{\leftarrow}_m$. Marginalizing out the latent group indicators, it is easy to show that the probability of observing an interaction between node p and q across the M networks is $\bar{\sigma}_{pq} = \vec{\pi}_p{}^{\top} B\, \vec{\pi}_q$.

Under an ALB model outlined above, the joint probability distribution of the data, $R_{1:M}$, and the latent variables $(\vec{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M})$ can be written in the following factored form:

$$\begin{aligned} p(R_{1:M}|\vec{\alpha}, B) &= \int_{\Pi\otimes\mathcal{Z}} p(R_{1:M}, \vec{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M}|\vec{\alpha}, B)\, d\vec{\pi}\, dZ \\ &= \int_{\Pi\otimes\mathcal{Z}} \Big(\prod_m \prod_{p,q} p_1(R_m(p,q)|\vec{z}^m_{p\rightarrow q}, \vec{z}^m_{p\leftarrow q}, B)\; p_2(\vec{z}^m_{p\rightarrow q}|\vec{\pi}_p, 1) \\ &\quad \times p_2(\vec{z}^m_{p\leftarrow q}|\vec{\pi}_q, 1) \Big) \prod_p p_3(\vec{\pi}_p|\vec{\alpha})\, d\vec{\pi}\, dZ \end{aligned} \tag{3.1}$$

where p_1 is Bernoulli, p_2 is multinomial, and p_3 is Dirichlet.

To compute the likelihood of the observed networks, one needs to marginalize out the hidden variables $\vec{\pi}$ and Z for all notes, which is intractable for even for small graphs. In Section 3.1.3, I describe a variational scheme to approximate this likelihood for parameter estimation.

Dealing with Sparsity Most networks in real world are sparse, meaning that most pairs of nodes do not have edges connecting them. But in many network analyses, observations about interactions and non-interactions are equally important in terms of their contributions to model fitness. In other words, they would compete for a statistical explanation in terms of estimates for parameters $(\vec{\alpha}, B)$, and would both influence the distribution of latent variables such as $\vec{\pi}_{1:N}$. A non desirable consequence of this, in scenarios where interactions are rare, is that parameter estimation and posterior inference would explain patterns of non-interaction rather than patterns of interaction.

In order to be able to calibrate the importance of rare interactions, we introduce the sparsity parameter $\rho \in [0, 1]$, which models how often a non-interaction is due to measurement noise (which

is common in certain experimentally derived networks such as the protein-protein interaction networks) and how often it carries information about the group memberships of the nodes. This leads to a small extension of the generative process outlined in the last subsection. Specifically, instead of drawing an edge directly from a Bernoulli with rate $\vec{z}^{m\top}_{p\to q} B\, \vec{z}^{m}_{p\leftarrow q}$, now we sample an interaction with probability $\sigma^m_{pq} = (1-\rho)\cdot \vec{z}^{m\top}_{p\to q} B\, \vec{z}^{m}_{p\leftarrow q}$; therefore the probability of having no interaction this pair of nodes is $1-\sigma^m_{pq} = (1-\rho)\cdot \vec{z}^{m\top}_{p\to q}(1-B)\, \vec{z}^{m}_{p\leftarrow q} + \rho$. This is equivalent to re-parameterizing the interaction matrix B. During estimation and inference, a large value of ρ would cause the interactions in the matrix to be weighted more than non-interactions in determining the estimates of $(\vec{\alpha}, B, \vec{\pi}_{1:N})$.

3.1.3 Estimation and Inference

I use an empirical Bayes framework for estimating the parameters $(\vec{\alpha}, B)$, and employ a mean-field approximation scheme (Jordan et al., 1999) for posterior inference of the (latent) mixed-membership vectors, $\vec{\pi}_{1:N}$. Model selection can be performed to determine the plausible value of K—the number of groups of nodes—based on a strategy described in Airoldi et al. (2006e).

In order to estimate $(\vec{\alpha}, B)$ and infer the posterior distributions of $\vec{\pi}_{1:N}$ we need to be able to evaluate the likelihood, which involves the non-tractable integral over Z and $\vec{\pi}_{1:N}$ in Equation 3.1. Given the large amount of data available for most networks, we focus on approximate posterior inference strategies in the context of variational methods, and we find a tractable lower bound for the likelihood that can be used as a surrogate for inference purposes. This leads to approximate MLEs for the hyper-parameters and approximate posterior distributions for the (latent) mixed-membership vectors.

Variational Expectation-Maximization The approximate variant of EM I describe here is often referred to as *Variational EM* (Beal and Ghahramani, 2003; Blei et al., 2003). Begin by rewriting

$Y = R_{1:M}$ for the data, $X = (\vec{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M})$ for the latent variables, and $\Theta = (\vec{\alpha}, B)$ for the model's parameters. Briefly, it is possible to lower bound the likelihood, $p(Y|\Theta)$, making use of Jensen's inequality and of any distribution on the latent variables $q(X)$,

$$\begin{aligned} p(Y|\Theta) &= \log \int_{\mathcal{X}} p(Y, X|\Theta)\, dX \\ &= \log \int_{\mathcal{X}} q(X)\, \frac{p(Y, X|\Theta)}{q(X)}\, dX \qquad \text{(for any } q\text{)} \\ &\geq \int_{\mathcal{X}} q(X)\, \log \frac{p(Y, X|\Theta)}{q(X)}\, dX \qquad \text{(Jensen's)} \\ &= \mathbb{E}_q \left[\log p(Y, X|\Theta) - \log q(X) \right] \;=:\; \mathcal{L}(q, \Theta) \end{aligned} \tag{3.2}$$

In EM, the lower bound $\mathcal{L}(q, \Theta)$ is then iteratively maximized with respect to Θ, in the M step, and q in the E step (Dempster et al., 1977). In particular, at the t-th iteration of the E step we set

$$q^{(t)} = p(X|Y, \Theta^{(t-1)}), \tag{3.3}$$

that is, equal to the posterior distribution of the latent variables given the data and the estimates of the parameters at the previous iteration.

Unfortunately, the posterior in Equation 3.3 for the admixture of latent blocks model cannot be computed. Rather, a direct parametric approximation to it needs be defined, $\tilde{q} = q_{\Delta}(X)$, which involves an extra set of *variational parameters*, Δ, and entails an approximate lower bound for the likelihood $\mathcal{L}_{\Delta}(q, \Theta)$. At the t-th iteration of the E step, the Kullback-Leibler divergence between $q^{(t)}$ and $q^{(t)}_{\Delta}$, is then minimized with respect to Δ, using the data.[3] The optimal parametric approximation is, in fact, a proper posterior as it depends on the data Y, although indirectly, $q^{(t)} \approx q^{(t)}_{\Delta^*(Y)}(X) = p(X|Y)$.

[3]This is equivalent to maximizing the approximate lower bound for the likelihood, $\mathcal{L}_{\Delta}(q, \Theta)$, with respect to Δ.

Lower Bound for the Likelihood According to the mean-field theory (Jordan et al., 1999; Xing et al., 2003b), one can approximate an intractable distribution such as the one defined by Equation 3.1 by a fully factored distribution $q(\vec{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M})$ defined as follows:

$$\begin{aligned} & q(\vec{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M} | \vec{\gamma}_{1:N}, \Phi^{\rightarrow}_{1:M}, \Phi^{\leftarrow}_{1:M}) \\ = & \prod_{p} q_1(\vec{\pi}_p|\vec{\gamma}_p) \prod_{m} \prod_{p,q} \Big(q_2(\vec{z}^{m}_{p\rightarrow q}|\vec{\phi}^{m}_{p\rightarrow q}, 1)\; q_2(\vec{z}^{m}_{p\leftarrow q}|\vec{\phi}^{m}_{p\leftarrow q}, 1) \Big), \end{aligned} \tag{3.4}$$

where q_1 is a Dirichlet, q_2 is a multinomial, and $\Delta = (\vec{\gamma}_{1:N}, \Phi^{\rightarrow}_{1:M}, \Phi^{\leftarrow}_{1:M})$ represent the set of free *variational parameters* need to be estimated in the approximate distribution.

Minimizing the Kulback-Leibler divergence between this $q(\vec{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M}|\Delta)$ and the original $p(\vec{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M})$ defined by Equation 3.1 leads to the following approximate lower bound for the likelihood.

$$\begin{aligned} \mathcal{L}_{\Delta}(q, \Theta) = & \mathbb{E}_q \Big[\log \prod_{m} \prod_{p,q} p_1(R_m(p,q)|\vec{z}^{m}_{p\rightarrow q}, \vec{z}^{m}_{p\leftarrow q}, B) \Big] \\ + & \mathbb{E}_q \Big[\log \prod_{m} \prod_{p,q} p_2(\vec{z}^{m}_{p\rightarrow q}|\vec{\pi}_p, 1) \Big] + \mathbb{E}_q \Big[\log \prod_{m} \prod_{p,q} p_2(\vec{z}^{m}_{p\leftarrow q}|\vec{\pi}_q, 1) \Big] \\ + & \mathbb{E}_q \Big[\log \prod_{p} p_3(\vec{\pi}_p|\vec{\alpha}) \Big] - \mathbb{E}_q \Big[\prod_{p} q_1(\vec{\pi}_p|\vec{\gamma}_p) \Big] \\ - & \mathbb{E}_q \Big[\log \prod_{m} \prod_{p,q} q_2(\vec{z}^{m}_{p\rightarrow q}|\vec{\phi}^{m}_{p\rightarrow q}, 1) \Big] - \mathbb{E}_q \Big[\log \prod_{m} \prod_{p,q} q_2(\vec{z}^{m}_{p\leftarrow q}|\vec{\phi}^{m}_{p\leftarrow q}, 1) \Big] . \end{aligned}$$

Working on the single expectations leads to the following expression,

$$\begin{aligned}
\mathcal{L}_{\Delta}(q,\Theta) &= \sum_{m}\sum_{p,q}\sum_{g,h}\phi^{m}_{p\to q,g}\phi^{m}_{p\leftarrow q,h}\cdot f\ \big(\ R_m(p,q), B(g,h)\ \big) \\
&+ \sum_{m}\sum_{p,q}\sum_{g}\phi^{m}_{p\to q,g}\ \big[\ \psi(\gamma_{p,g}) - \psi(\sum_{g}\gamma_{p,g})\ \big] \\
&+ \sum_{m}\sum_{p,q}\sum_{h}\phi^{m}_{p\leftarrow q,h}\ \big[\ \psi(\gamma_{p,h}) - \psi(\sum_{h}\gamma_{p,h})\ \big] \\
&+ \sum_{p}\log\Gamma(\sum_{k}\alpha_k) - \sum_{p,k}\log\Gamma(\alpha_k) + \sum_{p,k}(\alpha_k - 1)\ \big[\ \psi(\gamma_{p,k}) - \psi(\sum_{k}\gamma_{p,k})\ \big] \\
&- \sum_{p}\log\Gamma(\sum_{k}\gamma_{p,k}) + \sum_{p,k}\log\Gamma(\gamma_{p,k}) - \sum_{p,k}(\gamma_{p,k} - 1)\ \big[\ \psi(\gamma_{p,k}) - \psi(\sum_{k}\gamma_{p,k})\ \big] \\
&- \sum_{m}\sum_{p,q}\sum_{g}\phi^{m}_{p\to q,g}\log\phi^{m}_{p\to q,g} - \sum_{m}\sum_{p,q}\sum_{h}\phi^{m}_{p\leftarrow q,h}\log\phi^{m}_{p\leftarrow q,h}
\end{aligned}$$

where

$$f\ \big(\ R_m(p,q), B(g,h)\ \big) = R_m(p,q)\log B(g,h) + \big(\ 1 - R_m(p,q)\ \big)\ \log\ \big(\ 1 - B(g,h)\ \big);$$

m runs over $1,\ldots,M$; p,q run over $1,\ldots,N$; g,h,k run over $1,\ldots,K$; and $\psi(x)$ is the derivative of the log-gamma function, $\frac{d\log\Gamma(x)}{dx}$.

The Expected Value of the Log of a Dirichlet Random Vector The computation of the lower bound for the likelihood requires us to evaluate $\mathbb{E}_q\ \big[\ \log\vec{\pi}_p\ \big]$ for $p = 1,\ldots,N$. Recall that the density of an the exponential family distributions with natural parameter $\vec{\theta}$ can be written as

$$\begin{aligned}
p(x|\alpha) &= h(x)\cdot c(\alpha)\cdot\exp\big\{\ \sum_{k}\ \theta_k(\alpha)\cdot t_k(x)\ \big\} \\
&= h(x)\cdot\exp\big\{\ \sum_{k}\ \theta_k(\alpha)\cdot t_k(x) - \log c(\alpha)\ \big\}\ .
\end{aligned}$$

Omitting the node index p for convenience, the density of the Dirichlet distribution p_3 can be rewritten as an exponential family distribution,

$$p_3(\vec{\pi}|\vec{\alpha}) = \exp\left\{ \sum_k (\alpha_k - 1)\log(\pi_k) - \log \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} \right\},$$

with natural parameters $\theta_k(\vec{\alpha}) = (\alpha_k - 1)$ and natural sufficient statistics $t_k(\vec{\pi}) = \log(\pi_k)$. Let $c'(\vec{\theta}) = c(\alpha_1(\vec{\theta}), \ldots, \alpha_K(\vec{\theta}))$; using a well known property of the exponential family distributions (Schervish, 1995) it follows that

$$\begin{aligned}
\mathbb{E}_q \left[\log \pi_k \right] &= \mathbb{E}_{\vec{\theta}} \left[\log t_k(x) \right] \\
&= -\frac{\partial}{\partial \theta_k} \log c' \left(\vec{\theta} + 1 \right) \quad \text{(Schervish, 1995, Thm 2.64)} \\
&= \psi \left(\theta_k + 1 \right) - \psi \left(\sum_k \theta_k + K \right) \\
&= \psi \left(\alpha_k \right) - \psi \left(\sum_k \alpha_k \right),
\end{aligned}$$

where $\psi(x)$ is the derivative of the log-gamma function, $\frac{d \log \Gamma(x)}{dx}$.

Variational E Step The approximate lower bound for the likelihood $\mathcal{L}_\Delta(q, \Theta)$ can be maximized using exponential family arguments and coordinate ascent (Wainwright and Jordan, 2003).

Isolating terms containing $\phi^m_{p\to q,g}$ and $\phi^m_{p\leftarrow q,h}$ we obtain $\mathcal{L}_{\phi^m_{p\to q,g}}(q, \Theta)$ and $\mathcal{L}_{\phi^m_{p\to q,g}}(q, \Theta)$. The natural parameters $\vec{g}^m_{p\to q}$ and $\vec{g}^m_{p\leftarrow q}$ corresponding to the natural sufficient statistics $\log(\vec{z}^m_{p\to q})$ and $\log(\vec{z}^m_{p\leftarrow q})$ are functions of the other latent variables and the observations. We find that

$$\begin{aligned}
g^m_{p\to q,g} &= \log \pi_{p,g} + \sum_h z^m_{p\leftarrow q,h} \cdot f \left(R_m(p,q), B(g,h) \right), \\
g^m_{p\leftarrow q,h} &= \log \pi_{q,h} + \sum_g z^m_{p\to q,g} \cdot f \left(R_m(p,q), B(g,h) \right),
\end{aligned}$$

for all pairs of nodes (p, q) in the m-th network; where $g, h = 1, \ldots, K$, and

$$f\ \big(\ R_m(p,q), B(g,h)\ \big) = R_m(p,q) \log B(g,h) + \big(\ 1 - R_m(p,q)\ \big) \log\ \big(\ 1 - B(g,h)\ \big)\ .$$

This leads to the following updates for the variational parameters $(\vec{\phi}^m_{p\to q}, \vec{\phi}^m_{p\leftarrow q})$, for a pair of nodes (p, q) in the m-th network:

$$\begin{aligned}
\hat{\phi}^m_{p\to q,g} &\propto e^{\ \mathbb{E}_q\left[g^m_{p\to q,g}\right]} \\
&= e^{\ \mathbb{E}_q\left[\log \pi_{p,g}\right]} \cdot e^{\ \sum_h \phi^m_{p\leftarrow q,h} \cdot\ \mathbb{E}_q\left[f\left(R_m(p,q),B(g,h)\right)\right]} \\
&= e^{\ \mathbb{E}_q\left[\log \pi_{p,g}\right]} \cdot \prod_h \left(\ B(g,h)^{R_m(p,q)} \cdot\ \big(\ 1 - B(g,h)\ \big)^{1-R_m(p,q)}\right)^{\phi^m_{p\leftarrow q,h}} \\
\hat{\phi}^m_{p\leftarrow q,h} &\propto e^{\ \mathbb{E}_q\left[g^m_{p\leftarrow q,h}\right]} \\
&= e^{\ \mathbb{E}_q\left[\log \pi_{q,h}\right]} \cdot e^{\ \sum_g \phi^m_{p\to q,g} \cdot\ \mathbb{E}_q\left[f\left(R_m(p,q),B(g,h)\right)\right]} \\
&= e^{\ \mathbb{E}_q\left[\log \pi_{q,h}\right]} \cdot \prod_g \left(\ B(g,h)^{R_m(p,q)} \cdot\ \big(\ 1 - B(g,h)\ \big)^{1-R_m(p,q)}\right)^{\phi^m_{p\to q,g}}
\end{aligned}$$

for $g, h = 1, \ldots, K$. These estimates of the parameters underlying the distribution of the nodes' group indicators $\vec{\phi}^m_{p\to q}$ and $\vec{\phi}^m_{p\leftarrow q}$ need be normalized, to make sure $\sum_k \phi^m_{p\to q,k} = \sum_k \phi^m_{p\leftarrow q,k} = 1$.

Isolating terms containing $\gamma_{p,k}$ we obtain $\mathcal{L}_{\gamma_{p,k}}(q, \Theta)$. Setting $\frac{\partial \mathcal{L}_{\gamma_{p,k}}}{\partial \gamma_{p,k}}$ equal to zero and solving for $\gamma_{p,k}$ yields:

$$\hat{\gamma}_{p,k} = \alpha_k + \sum_m \sum_q \phi^m_{p\to q,k} + \sum_m \sum_q \phi^m_{p\leftarrow q,k},$$

for all nodes $p \in \mathcal{P}$ and $k = 1, \ldots, K$.

The t-th iteration of the variational E step is carried out for fixed values of $\Theta^{(t-1)} = (\vec{\alpha}^{(t-1)}, B^{(t-1)})$, and finds the optimal approximate lower bound for the likelihood $\mathcal{L}_{\Delta^*}(q, \Theta^{(t-1)})$.

Variational M Step The optimal lower bound $\mathcal{L}_{\Delta^*}(q^{(t-1)}, \Theta)$ provides a tractable surrogate for the likelihood at the t-th iteration of the variational M step. We derive empirical Bayes estimates for the hyper-parameters Θ that are based upon it.[4] That is, we maximize $\mathcal{L}_{\Delta^*}(q^{(t-1)}, \Theta)$ with respect to Θ, given expected sufficient statistics computed using $\mathcal{L}_{\Delta^*}(q^{(t-1)}, \Theta^{(t-1)})$.

Isolating terms containing $\vec{\alpha}$ we obtain $\mathcal{L}_{\vec{\alpha}}(q, \Theta)$. Unfortunately, a closed form solution for the approximate maximum likelihood estimate of $\vec{\alpha}$ does not exist (Blei et al., 2003). We can produce a Newton-Raphson method that is linear in time, where the gradient and Hessian for the bound $\mathcal{L}_{\vec{\alpha}}$ are

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\vec{\alpha}}}{\partial \alpha_k} &= N \left(\psi \left(\sum_k \alpha_k \right) - \psi(\alpha_k) \right) + \sum_p \left(\psi(\gamma_{p,k}) - \psi \left(\sum_k \gamma_{p,k} \right) \right), \\
\frac{\partial \mathcal{L}_{\vec{\alpha}}}{\partial \alpha_{k_1} \alpha_{k_2}} &= N \left(\mathbb{I}_{(k_1=k_2)} \cdot \psi'(\alpha_{k_1}) - \psi' \left(\sum_k \alpha_k \right) \right).
\end{aligned}$$

Isolating terms containing B we obtain $\mathcal{L}_B$, whose approximate maximum is

$$\hat{B}(g,h) = \frac{1}{M} \sum_m \left(\frac{\sum_{p,q} R_m(p,q) \cdot \phi^m_{p \to qg} \, \phi^m_{p \leftarrow qh}}{\sum_{p,q} \phi^m_{p \to qg} \, \phi^m_{p \leftarrow qh}} \right),$$

for every index pair $(g,h) \in [1,K] \times [1,K]$.

In Section 3.1.2 we introduced an extra parameter, ρ, to control the relative importance of presence and absence of interactions in likelihood, i.e., the score that informs inference and estimation. Isolating terms containing ρ we obtain $\mathcal{L}_\rho$. We may then estimate the sparsity parameter ρ by

$$\hat{\rho} = \frac{1}{M} \sum_m \left(\frac{\sum_{p,q} \left(1 - R_m(p,q) \right) \cdot \left(\sum_{g,h} \phi^m_{p \to qg} \, \phi^m_{p \leftarrow qh} \right)}{\sum_{p,q} \sum_{g,h} \phi^m_{p \to qg} \, \phi^m_{p \leftarrow qh}} \right).$$

Alternatively, we can fix ρ prior to the analysis; the density of the interaction matrix is estimated

[4] We could term these estimates *pseudo* empirical Bayes estimates, since they maximize an approximate lower bound for the likelihood, $\mathcal{L}_{\Delta^*}$.

with $\hat{d} = \sum_{m,p,q} R_m(p,q)/(N^2 M)$, and the sparsity parameter is set to $\tilde{\rho} = (1 - \hat{d})$. This latter estimator attributes all the information in the non-interactions to the point mass, i.e., to latent sources other than the block model B or the mixed membership vectors $\vec{\pi}_{1:N}$. It does however provide a quick recipe to reduce the computational burden during exploratory analyses.[5]

Smoothing In problems where the number of clusters is deemed to be likely large a-priori, we can smooth the (consequently large number of) cluster-to-cluster relation probabilities encoded in the block model B by positing that all the elements $B(g,h)$ of the block model are non-observable samples from a common (prior) distribution. In the admixture of latent blocks model we posit that $p(B|\vec{\lambda})$ is a collection non-symmetric beta distributions, with a pair of hyper-parameters $\vec{\lambda}$ common to all elements of B.

Example 17 (Continued) Sampson (1968) surveyed 18 novice monks in a monastery and asked them to rank the other novices in terms of four *sociometric relations*: like/dislike, esteem, personal influence, and alignment with the monastic credo. Sampson's original analysis strongly suggests the existence of tight factions among the novices, and the events that took place during his stay at the monastery support his observations; briefly, novices of one faction left the monastery or were expelled over religious differences. The factions identified by Sampson provide a credible gold standard, to which the results are compared.

I consider Breiger's collation of Sampson's data (Breiger et al., 1975). Briefly, for each of the four sociometric relations above, only the top three choices of each novice were recorded as positive relations—the edges in the graph. The union of all positive relations, disregarding multiplicity as in Handcock et al. (2007), is the starting point of our analysis. To assess model fit,

[5]Note that $\tilde{\rho} = \hat{\rho}$ in the case of single membership. In fact, that implies $\phi^m_{p \to qg} = \phi^m_{p \leftarrow qh} = 1$ for some (g,h) pair, for any (p,q) pair.

I use an approximation to BIC:

$$BIC = 2 \cdot \log p(R) \approx 2 \cdot \log p(R|\widehat{\vec{\pi}}, \widehat{Z}, \widehat{\vec{\alpha}}, \widehat{B}) - |\vec{\alpha}, B| \cdot \log |R|,$$

where $|\vec{\alpha}, B|$ is the number of hyper-parameters in the model, and $|R|$ is the number of positive relations observed—following arguments in Handcock et al. (2007). The approximate BIC value suggests that the relations among monks in the monastery studied by Sampson are best explained by a model with three factions, independently of the number of hyper-parameters in the fitted ALB models. In the left panel of Figure 3.2 I show the approximate BIC for a model with a single hyper-parameter, α scalar. Hence I fixed $\hat{K} = 3$ in subsequent analyses, which involved ALB models with increasing degree of complexity. The right panel of Figure 3.2 shows the estimated faction-to-faction block model, $\widehat{B}$, corresponds to a full model (i.e., no constraints on B). This estimate suggest that the Outcasts are an isolated faction, whereas Young Turks *like* members of the Loyal Opposition, although the sentiment is not reciprocated. Figure 3.3 investigates the the posterior means of the mixed membership scores, $\mathbb{E}[\vec{\pi}|R]$, for the 18 monks in the monastery ($\alpha = 0.058$ scalar, $B := \mathbb{I}_3$). There is a panel for each monk, and the subscripts associated with the names of

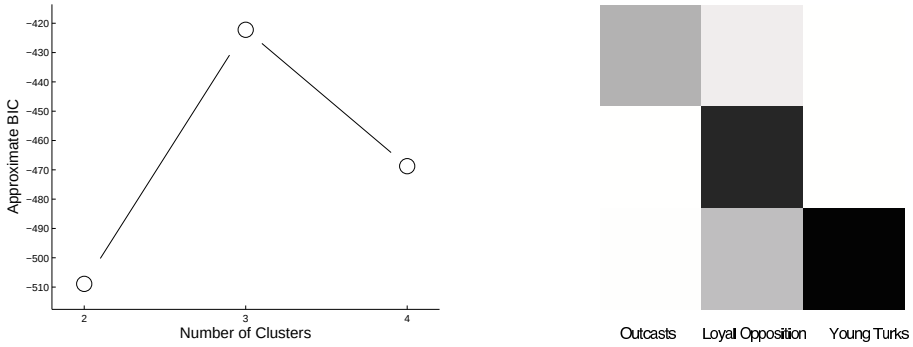

Figure 3.2: The approximate BIC (left panel) suggests the relations among monks are best explained by a model with three factions. The faction-to-faction estimated relational patterns (right panel) suggest that the Outcasts are an isolated faction, whereas Young Turks *like* members of the Loyal Opposition, although the sentiment is not reciprocated.

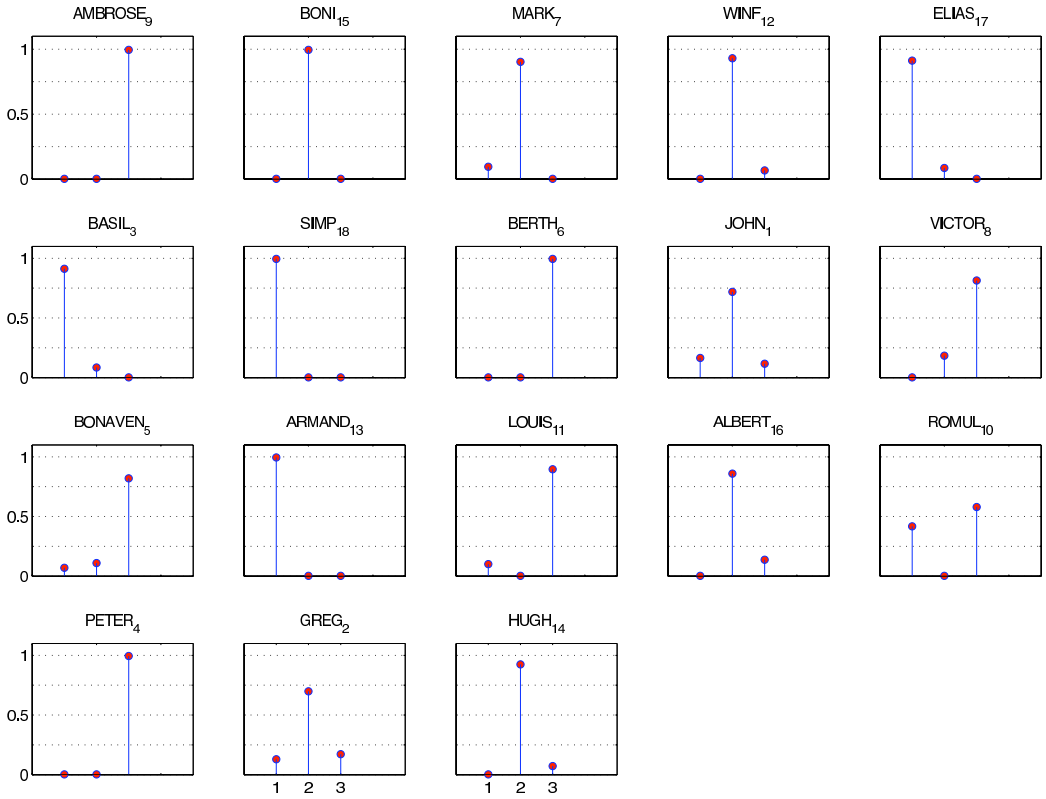

Figure 3.3: The posterior mixed membership scores, $\vec{\pi}$, for the 18 monks in the monastery. Each panel correspond to a monk; the Y axis measures the grade of membership, corresponding to the Outcast (left bar), to the Young Turks (center bar), and to the Loyal Opposition (right bar), on the X axis. The subscripts associated with the names of the monks specify the order according to which they left the monastery.

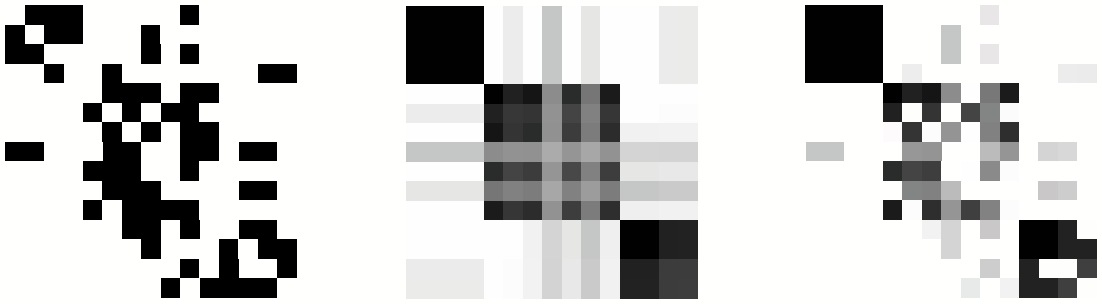

Figure 3.4: Original matrix of sociometric relations (left), and estimated relations obtained by thresholding the posterior expectations $\vec{\pi}_p{}' B\, \vec{\pi}_q | R$ (center), and $\vec{\phi}_p{}' B\, \vec{\phi}_q | R$ (right).

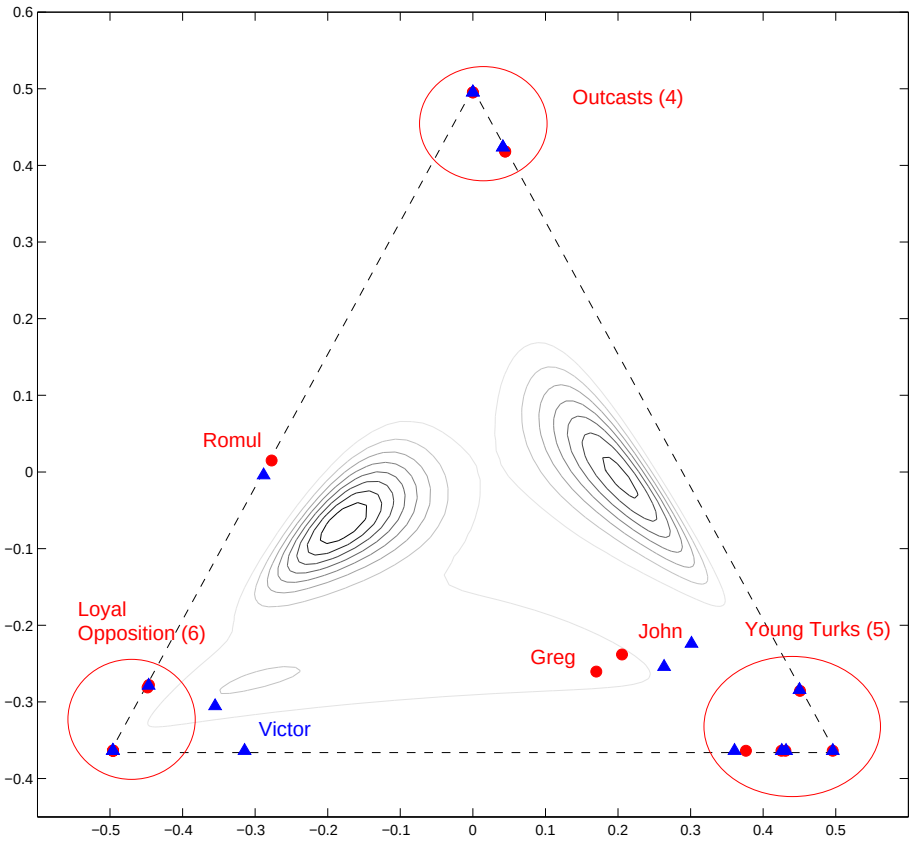

Figure 3.5: Mixed membership vectors, $\vec{\pi}_{1:18}$, plotted in the reference simplex. Marks correspond to individual monks; the red circle marks correspond to an ALB model with $(B = I_3, \alpha = 0.01)$, whereas the blue triangle marks correspond to an ALB model with $(B := I_3, \hat{\alpha} = 0.058)$; where I_K is the K-dimensional identity matrix.

the monks specify the order according to which they left the monastery, e.g., John left first. The three factions on the X axis are the Outcast, the Young Turks , and the Loyal Opposition (from left to right); and the Y axis measures the degree of membership of monks to factions. From these panels, the centrality of the role played by John and Greg, first to leave the monastery, as well as the uncertain affiliations of Romul, and Victor to a minor extent, unequivocally emerge. The mixed membership vectors, $\vec{\pi}_{1:18}$, provide us with low-dimensional representations of monks. Figure 3.5 plots them in their natural space, that is, the(3-dimensional) simplex. Dots correspond

to monks; the red circles were obtained by fixing $B = \mathbb{I}_3$ and $\alpha = 0.01$, whereas the blue triangles correspond to fixing $B := \mathbb{I}_3$, but estimating $\hat{\alpha} = 0.058$. To compare the latent representation of the monks obtained with ALB with the one presented in (Handcock et al., 2007, Table 1), I mapped the contour levels for their the estimated mixture of three Gaussians (Handcock et al., 2007, Table 1) in the reference simplex—using the following transformation,

$$T = \begin{bmatrix} -0.5 & 0.5 & 0 \\ -\frac{\sqrt{3}-1}{2} & \frac{\sqrt{3}-1}{2} & 0.5 \end{bmatrix}.$$

The contour levels of such density in Figure 3.5 suggest that our model and the latent space mixture model lead to different structures and somewhat different interpretations.

Example 18 (Continued) The goal of the analysis here is to analyze proteins' diverse functional roles by analyzing their local and global patterns of interaction. The biochemical composition of individual proteins make them suitable for carrying out a specific set of cellular operations, or *functions*. Proteins typically carry out these functions as part of stable protein complexes (Krogan et al., 2006). There are many situations in which proteins are believed to interact (Alberts et al., 2002); the main intuition behind our methodology is that pairs of protein interact because they are part of the same stable protein complex, i.e., co-location, or because they are part of interacting protein complexes as they carry out compatible cellular operations.

The Munich Institute for Protein Sequencing (MIPS) database was created in 1998 based on evidence derived from a variety of experimental techniques, but does not include information from high-throughput data sets (Mewes et al., 2004). It contains about 8000 protein complex associations in yeast. We analyze a subset of this collection containing 871 proteins, the interactions amongst which were hand-curated. The institute also provides a set of functional annotations, alternative to the gene ontology (GO). These annotations are organized in a tree, with 15 general

Table 3.1: General functional categories in the MIPS tree, and their relative popularity. In the table we report the number of proteins that have at least one functional annotation in the general categories in the left column. Counts refer to the subset of 871 proteins in yeast, which are part of the hand-curated MIPS interaction network.

#	Category	Size
1	Metabolism	125
2	Energy	56
3	Cell cycle & DNA processing	162
4	Transcription (tRNA)	258
5	Protein synthesis	220
6	Protein fate	170
7	Cellular transportation	122
8	Cell rescue, defence & virulence	6
9	Interaction w/ cell. environment	18
10	Cellular regulation	37
11	Cellular other	78
12	Control of cell organization	36
13	Sub-cellular activities	789
14	Protein regulators	1
15	Transport facilitation	41

functions at the first level, 72 more specific functions at an intermediate level, and 255 annotations at the the leaf level. In Table 3.1 we map the 871 proteins in our collections to the main functions of the MIPS annotation tree; proteins in our sub-collection have about 2.4 functional annotations on average.[6] By mapping proteins to the 15 general functions, we obtain a 15-dimensional representation for each protein. In Figure 3.6 each panel corresponds to a protein; the 15 functional categories are ordered as in Table 3.1 on the X axis, whereas the presence or absence of the corresponding functional annotation is displayed on the Y axis.

Protein-protein interactions (PPI) form the physical basis for formation of complexes and pathways which carry out different biological processes. A number of high-throughput experimental approaches have been applied to determine the set of interacting proteins on a proteome-wide scale

[6]We note that the relative importance of functional categories in our sub-collection, in terms of the number of proteins involved, is different from the relative importance of functional categories over the entire MIPS collection.

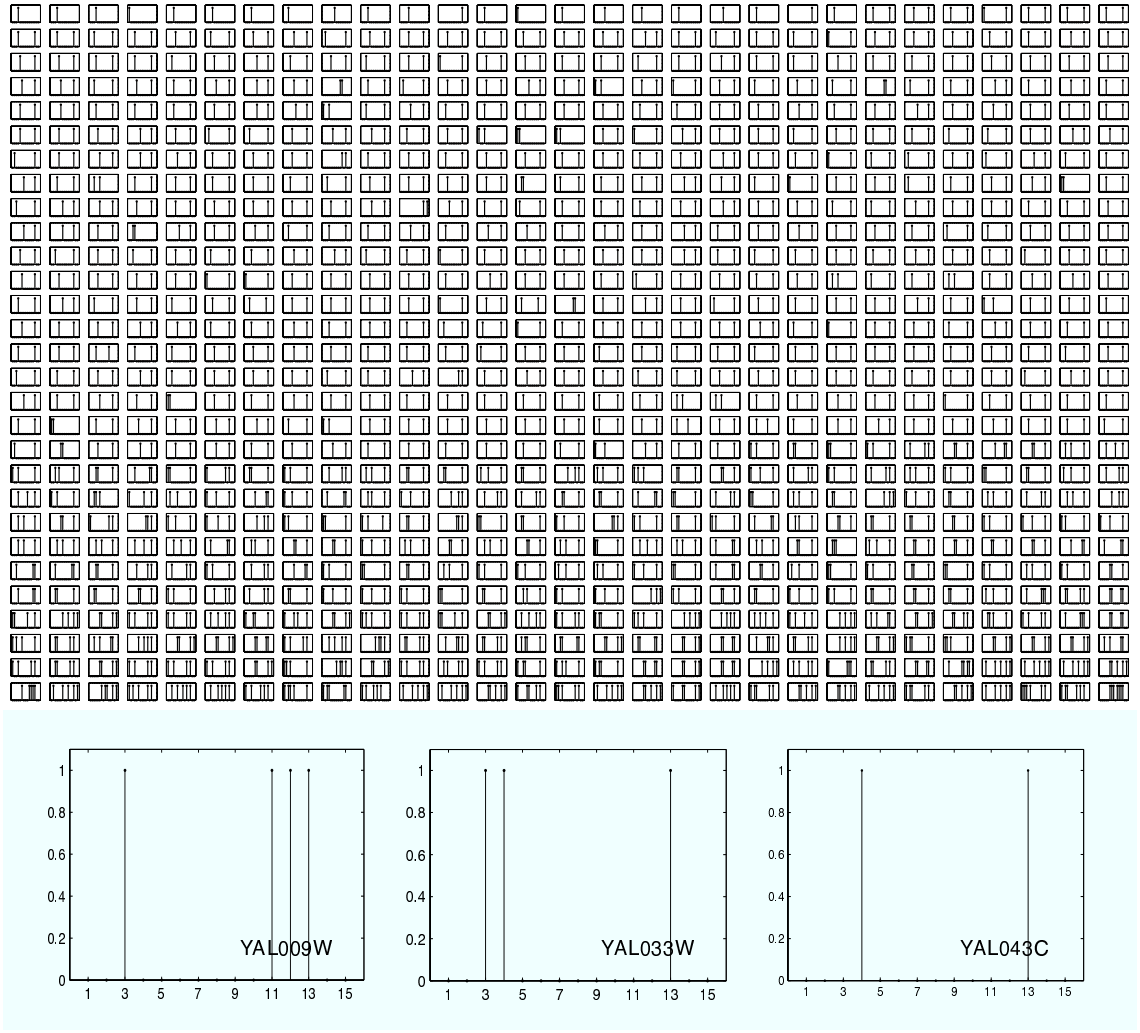

Figure 3.6: By cutting the MIPS annotation tree at the first level we find the 15 general functional categories in Table 3.1. By mapping proteins to the 15 general functions, we obtain a 15-dimensional representation for each protein. In the Figure, each panel corresponds to a protein; the 15 functional categories are displayed on the X axis, whereas the presence or absence of the corresponding functional annotation is displayed on the Y axis. The plots at the bottom zoom into the panels corresponding to three example proteins.

in yeast. These include the two-hybrid (Y2H) screens and mass spectrometry methods. For example, mass spectrometry is used to identify components of protein complexes (Gavin et al., 2002; Ho et al., 2002). High-throughput methods, though, may miss complexes that are not present under the given conditions. For example, tagging may disturb complex formation and weakly associated components may dissociate and escape detection. Statistical models that encode information about functional processes with high precision are then an essential tool for carry out *probabilistic denoising* of biological signals from high-throughput experiments.

In previous work, we established the usefulness of the admixture of latent blocks model for analyzing protein-protein interaction data. For example, we used the ALB for testing functional interaction hypotheses and semi-supervised and unsupervised estimation experiments (Airoldi et al., 2005b). We then attempted to assess whether, and how much, functionally relevant biological signal can be captured in by the ALB model (Airoldi et al., 2005a). In summary, our findings show that ALB identifies protein complexes whose member proteins are tightly interacting with one another. The identifiable protein complexes correlate with the following four categories of Table 3.1: cell cycle & DNA processing, transcription, protein synthesis, and sub-cellular activities. The high correlation of inferred protein complexes can be leveraged for predicting the presence of absence of functional annotations, for example, by using a logistic regression. However, there is not enough signal in the data to independently predict annotations in other functional categories. The empirical Bayes estimates of the hyper-parameters that support these conclusions in the various types of analyses are consistent; $\hat{\alpha} < 1$ and small; and $\hat{B}$ nearly block diagonal with two positive blocks comprising the four identifiable protein complexes. These previous analyses fixed the number of latent protein complexes to 15. Figure 3.7 displays few examples of predicted mixed membership probabilities against the true annotations, given an *estimated mapping* of latent protein complexes to functional categories. The latent protein complexes are not a-priori identifiable in our model. To resolve this, we find mapping between latent complexes and functions by minimizing the diver-

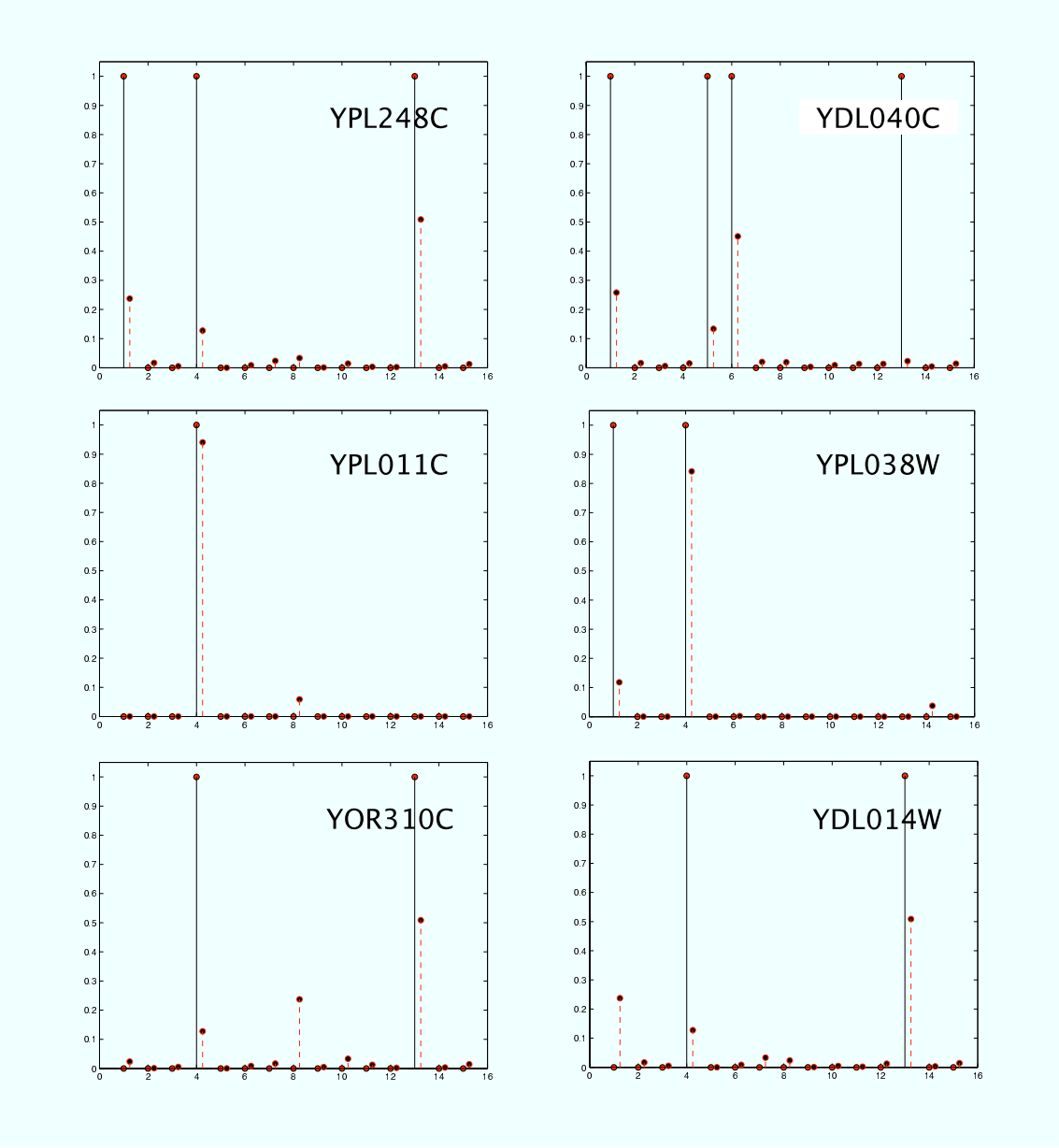

Figure 3.7: Predicted mixed-membership probabilities (dashed, red lines) versus binary manually curated functional annotations (solid, black lines) for 6 example proteins. The identification of latent groups to functions is estimated, and it is discussed in Figure 3.8.

gence between true [7] and predicted marginal frequencies of membership. We used this mapping to compare predicted versus known functional annotations, for all proteins. The best estimated mapping is shown in Figure 3.8.

Following-up on the hypothesis that the size of stable protein complex in Yeast is about 5 proteins on average, and skewed towards bigger complexes (Krogan et al., 2006), we explored a richer space of models with $K = 50, \ldots, 225$. However, using approximate BIC to assess model fit (Handcock et al., 2007) we found that the more parsimonious models $(K = 50)$ provide a better description of the observed interactions. This fact is consistent with previous findings (Airoldi et al., 2005b), and suggest that the interactions in the MIPS collection alone encode a biological signal at a higher aggregation level than that of a specific complexes. In order to explore this hypothesis we considered an alternative annotation scheme to that of the Munich Institute for Protein Sequencing; namely the Saccaromice Cervisiae gene database and gene ontology (GO) (Ashburner et al., 2000). Based on the GO, Myers et al. (2006) recently proposed a solid framework to assess the functional content of biological data. Making use of it, we measure the functional content in the interactions encoded in an ALB model with $K = 50$, fitted using the nested variational EM algorithm detailed in the Appendix. In Figure 3.9, we measure the functional content in the posterior means,

$$\mathbb{E}\left[\, R(p,q) = 1 \,\right] = \widehat{\vec{\pi}}_p{}' \, \widehat{B} \, \widehat{\vec{\pi}}_q \quad \text{and} \quad \mathbb{E}\left[\, R(p,q) = 1 \,\right] = \widehat{\vec{\phi}}_{p \to q}{}' \, \widehat{B} \, \widehat{\vec{\phi}}_{p \leftarrow q},$$

where positive interactions are obtained by thresholding the expectations. Figure 3.9 shows the original MIPS collection as one of the most precise (Y axis) and most extensive (X axis) source of biologically relevant interactions available to date. The posterior means of $(\vec{\pi}_{1:N})$ and the estimates of (α, B) provide a parsimonious representation for the MIPS collection, and lead to precise interaction estimates, however, in moderate amount (the light blue, $-\times$ line). The posterior means

[7] Evaluated on a small fraction of the interactions.

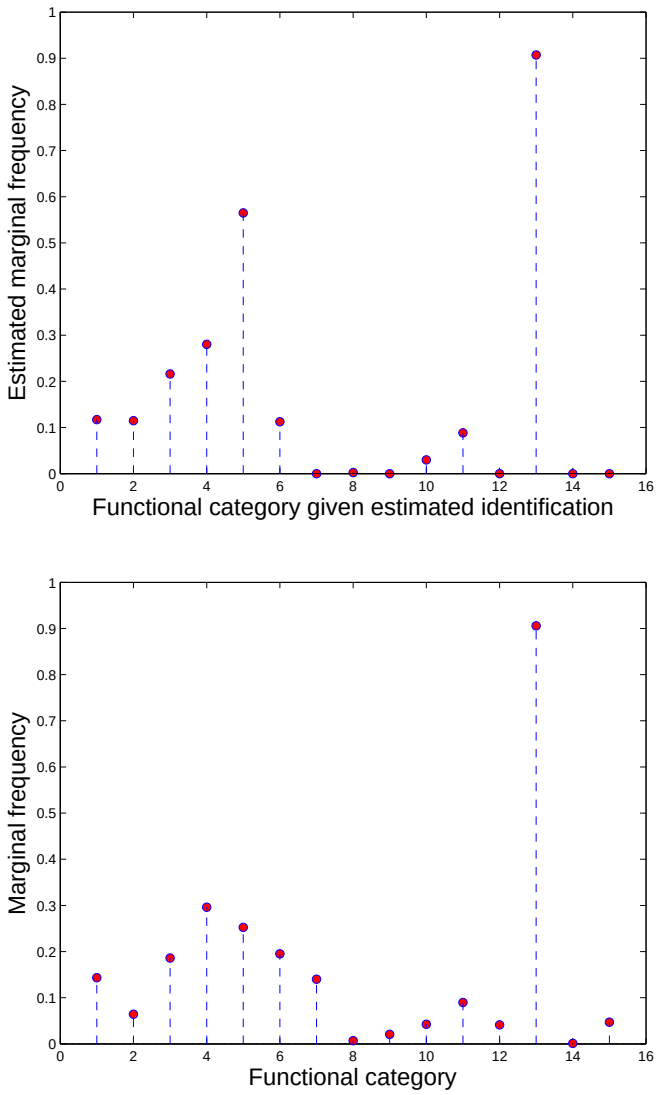

Figure 3.8: We estimate the mapping of latent groups to functions. The two plots show the marginal frequencies of membership of proteins to true functions (bottom) and to identified functions (top), in the cross-validation experiment. The mapping is selected to maximize the accuracy of the predictions on the training set, in the cross-validation experiment, and to minimize the divergence between marginal true and predicted frequencies if no training data is available—see the text.

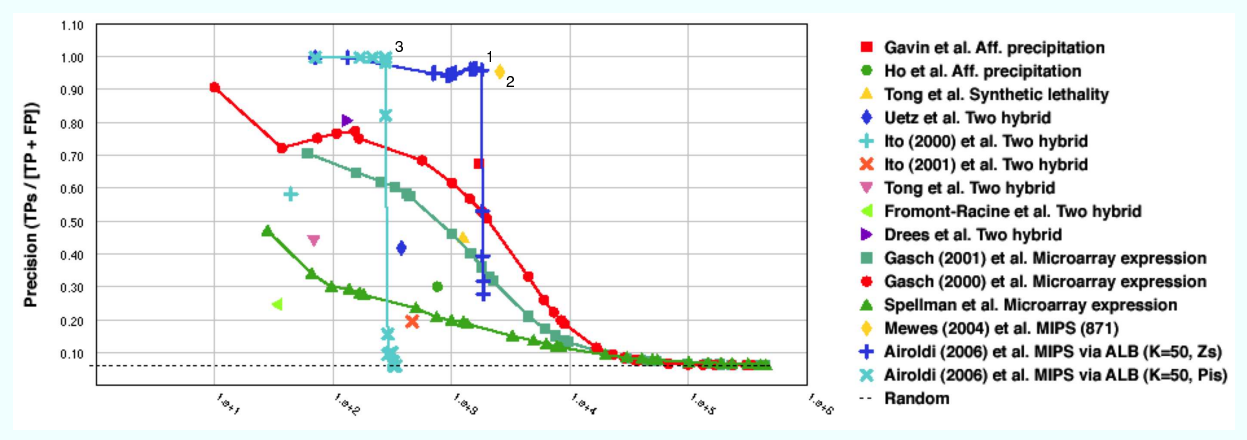

Figure 3.9: In the top panel we measure the functional content of the the MIPS collection of protein interactions (yellow diamond), and compare it against other published collections of interactions and microarray data, and to the posterior estimates of ALB models—computed as described in the text. A breakdown of three estimated interaction networks (the numbered points) into most represented gene ontology categories is detailed in Table 3.2.

of $(Z^{\rightarrow}, Z^{\leftarrow})$ do not provide a parsimonious representation for the data, and describe most of the functional content of the MIPS collection with high precision (the dark blue, $-+$ line). A breakdown of three example interaction networks displayed in Figure 3.9 into most represented gene ontology categories is detailed in Table 3.2. We investigate the correlations between data collections (rows) and a sample of gene ontology categories (columns). The intensity of the square (red is high) measures the area under the precision-recall curve For more detail about these plots see Figures 5–6 in Myers et al. (2006).

* * *

When applied to a sample of measurements on pairs of objects, *Admixture of Latent Blocks* simultaneously extracts information about (i) the mixed membership of objects to latent aspects, and (ii) the connectivity patterns among latent aspects, using a nested variational EM algorithm. I found it useful for revealing group membership in social networks, as well as for describing and summarizing the functional content of a protein interaction network, and I envision its use for de-noising

Table 3.2: Breakdown of three example interaction networks into most represented gene ontology categories. The digit in the first column refers to the numbered points in Figure 3.9. The last two columns quote the number of predicted, and possible pairs for each GO term.

#	GO Term	Description	Pred.	Tot.
1	GO:0043285	Biopolymer catabolism	561	17020
1	GO:0006366	Transcription from RNA polymerase II promoter	341	36046
1	GO:0006412	Protein biosynthesis	281	299925
1	GO:0006260	DNA replication	196	5253
1	GO:0006461	Protein complex assembly	191	11175
1	GO:0016568	Chromatin modification	172	15400
1	GO:0006473	Protein amino acid acetylation	91	666
1	GO:0006360	Transcription from RNA polymerase I promoter	78	378
1	GO:0042592	Homeostasis	78	5778
2	GO:0043285	Biopolymer catabolism	631	17020
2	GO:0006366	Transcription from RNA polymerase II promoter	414	36046
2	GO:0016568	Chromatin modification	229	15400
2	GO:0006260	DNA replication	226	5253
2	GO:0006412	Protein biosynthesis	225	299925
2	GO:0045045	Secretory pathway	151	18915
2	GO:0006793	Phosphorus metabolism	134	17391
2	GO:0048193	Golgi vesicle transport	128	9180
2	GO:0006352	Transcription initiation	121	1540
3	GO:0006412	Protein biosynthesis	277	299925
3	GO:0006461	Protein complex assembly	190	11175
3	GO:0009889	Regulation of biosynthesis	28	990
3	GO:0051246	Regulation of protein metabolism	28	903
3	GO:0007046	Ribosome biogenesis	10	21528
3	GO:0006512	Ubiquitin cycle	3	2211

new collection of interactions from high-throughput experiments.

A recurring question, which bears relevance to mixed membership models in general, is why one does not necessarily want to integrate out the single membership indicators—$(\vec{z}^{m}_{p\to q}, \vec{z}^{m}_{p\leftarrow q})$ in the specifications above. There are some computational aspects to this but a practical issue that argues against such marginalization is that we would often lose interpretable quantities that are useful for making predictions, for de-noising new measurements, or for performing other tasks.

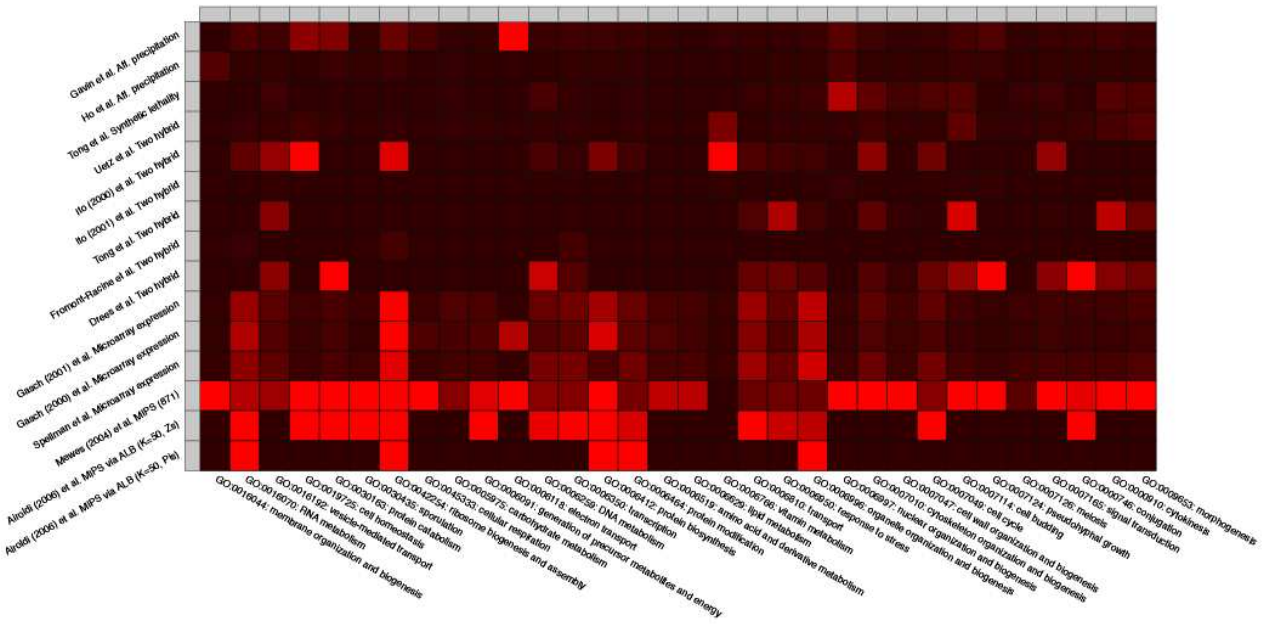

Figure 3.10: We investigate the correlations between data collections (rows) and a sample of gene ontology categories (columns). The intensity of the square (red is high) measures the area under the precision-recall curve.

In fact, the posterior distributions of such quantities typically carry substantive information about elements of the application at hand. In the application to protein interaction networks, for example, they encode the interaction-specific memberships of individual proteins to protein complexes.

There is a tight relationship between ALB and the latent space models in Hoff et al. (2002); Handcock et al. (2007). In the latent space models, the latent vectors are drawn from Gaussian distributions and the interaction data is drawn from a Gaussian with mean $\vec{\pi}_p{}' \mathbb{I} \vec{\pi}_q$. In ALB, the marginal probability of an interaction takes a similar form, $\vec{\pi}_p{}' B \vec{\pi}_q$, where B is the matrix of probabilities of interactions for each pair of latent factions. In contrast to the latent space model, the relations can be modeled by an arbitrary distribution, in our model. With binary relations a collection of Bernoulli parameters can be used; with continuous relations, a collection of Gaussian parameters can be used. While more flexible, ALB does not subsume latent space models; they make different assumptions about the data. See Handcock et al. (2007) with discussion (Blei and Fienberg, 2007;

Airoldi, 2007) for more details.

3.2 Local Diffusion Potentials

Here I briefly situate in the context of this thesis some recent developments in mathematics that bear relevance to statistical network analysis. I wish to thank Ann B. Lee for carrying out the calculations and for generous advice on the material presented here.

The main linkage between the mathematics of diffusion (Lafon and Lee, 2006) and statistical network analysis is a notion of distance that measures the connectivity between nodes through a multiple-step multiple-path diffusion process on the graph. This formulation depends on non-observable, node-specific quantities, which I term *local diffusion potential*. Higher-order connectivity patterns could be naturally incorporated into statistical models of graphs and networks in diffusion space (Coifman et al., 2005a,b).

Example 19. *Consider the diffusion of innovation among physicians studied by Coleman et al. (1957). Doctor A suggests doctor B to try a new drug, who later suggests its use to doctor C. In a sense, the influence of doctor A indirectly extends to doctor C. Such mediated connections may occur through multiple steps and multiple paths. Defining a distance metric that explicitly encodes this enriched notion of connectivity, that is, the diffusion potential specific to a doctor (to a node in a graph), is the focus of this section.*

3.2.1 Goals of the Analysis

An abstract framework to study diffusion has been introduced by Coifman et al. (2005a,b) in the context of high-dimensional data analysis and manifold learning. In such a framework nodes in a graph are represented in terms of their multivariate attributes, $\vec{x}_n \in \mathbb{R}^p$ for each $n \in \mathcal{N}$. A kernel

function $f\left(\frac{||\mathbf{x}_n-\mathbf{y}||}{h}\right)$, with a certain bandwidth h, defines the local neighborhood of $\vec{x}_n$, and it used to compute weights of the edges to be imputed. This procedure *results* in a (fully connected) graph among the nodes. Within this framework, a notion of distance has been defined that controls the influence of a each node on its neighbors, through a multiple-step multiple-path diffusion process on the graph, at a global scale (Lafon et al., 2006).

Rather, for the purposes of this thesis, a graph *is given*. It is possible, however, to measure distance between nodes through a multiple-step multiple-path diffusion process on the graph by defining *local diffusion potentials* in the form of local scales, t_n, specific to nodes $n \in \mathcal{N}$. Distances in the graph, based upon local diffusion potentials, correspond to Euclidean distances on the lower dimensional manifold implicitly defined by the graph.[8]

3.2.2 Technical Preliminaries

Consider a connected graph $G = (V, E)$, where V is a set of N vertices, and E is a set of undirected edges. Edges are mapped to weights w_{ij}, for $i, j \in V$. The weights $W = \{w_{ij}\}$ satisfy the following conditions: (i) symmetry, $W = W^T$, (ii) pointwise positivity, $w_{ij} \geq 0$ for $i, j \in V$ and $w_{ii} > 0$, and (iii) positive semi-definiteness. These conditions can be relaxed, and the methodology extended, to cover the more general case of directed graphs.

Consider the spectral properties of the Markov chain on W. The transition matrix P has a set of left and right eigenvectors according to:

$$\phi_k^T P = \lambda_k \phi_k^T \text{ and } P\psi_k = \lambda_k \psi_k \tag{3.5}$$

where the eigenvalues $\lambda_0 = 1 \geq \lambda_1 \geq ... \geq \lambda_{N-1} \geq 0$. Furthermore, the left and right eigenvectors satisfy the biorthogonality relation $\phi_k^T \psi_l = \delta_{kl}$, where δ_{kl} is Dirac's delta function. For

[8] Explicitly defined by $\vec{x}_{1:N} \in \mathbb{R}^p$ in the formulation of Coifman et al. (2005a,b).

convenience, the eigenvectors are normalized with respect to $1/\phi_0$ and ϕ_0, respectively, so that:

$$\begin{aligned} \|\phi_k\|^2_{1/\phi_0} &= \textstyle\sum_i \frac{\phi_k^2(i)}{\phi_0(i)} &&= 1 \\ \|\psi_k\|^2_{\phi_0} &= \textstyle\sum_i \psi_k^2(i)\phi_0(i) &&= 1. \end{aligned} \tag{3.6}$$

It can be verified that $\lambda_0 = 1$, $\psi_0 \equiv 1$, and that ϕ_0 is defined as in Eq. 3.11. Furthermore, it follows that

$$\psi_k(i) = \frac{\phi_k(i)}{\phi_0(i)} \tag{3.7}$$

for $k = 0, 1, \ldots, N-1$ and $i \in V$. Rewrite the transition probabilities $p_t(i,j)$ and the diffusion metric in terms of these eigenvectors and eigenvalues. By inserting the biorthogonal spectral decomposition

$$p_t(i,j) = \sum_{k\geq 0} \lambda_k^t \psi_k(i)\phi_k(j), \tag{3.8}$$

into Eq. 3.10, and using orthonormality $\sum_j \frac{\phi_k(j)\phi_l(j)}{\phi_0(j)} = \delta_{kl}$, it follows

$$\begin{aligned} \mathcal{D}^2(n,m;t_n,t_m) &= \sum_{k>0} \left(\lambda_k^{t_n}\psi_k(n) - \lambda_k^{t_m}\psi_k(m)\right)^2 \\ &\simeq \sum_{k=1}^{K} \left(\lambda_k^{t_n}\psi_k(n) - \lambda_k^{t_m}\psi_k(m)\right)^2. \end{aligned} \tag{3.9}$$

Note that the $k = 0$ term does not appear in the sum as $\lambda_0 = 1$ and $\psi \equiv 1$.

3.2.3 The Main Result

The calculations above lead to a generalized diffusion distance between nodes n and m according to

$$\begin{aligned}\mathcal{D}^2(n,m;t_n,t_m) &= \|p_{t_n}(n,\cdot) - p_{t_m}(m,\cdot)\|^2_{1/\phi_0} \\ &= \sum_{j\in V} \frac{(p_{t_n}(n,j) - p_{t_m}(m,j))^2}{\phi_0(j)},\end{aligned} \tag{3.10}$$

where the scale parameters t_n and t_m determine the local influence of nodes n and m on their neighbors, and the function

$$\phi_0(j) = \frac{d_j}{\sum_{k\in V} d_k} \tag{3.11}$$

is the stationary distribution of the Markov chain, i.e. $\lim_{t\to+\infty} p_t(i,j) = \phi_0(j)$. According to this metric, nodes n and m will be close, *if they interact with the same nodes in the graph* [9]. For an undirected graph, such a situation occurs when there are many paths connecting the two nodes.

The main points are the following:

- From Eq. 3.9, it is clear that *the diffusion metric is a distance on the graph induced by a one-parametric family of eigenmaps*

$$\Psi_t : n \longmapsto \begin{pmatrix} \lambda_1^t\psi_1(n) \\ \lambda_2^t\psi_2(n) \\ \vdots \\ \lambda_K^t\psi_K(n) \end{pmatrix} \tag{3.12}$$

  for $n \in V$.

[9] The weights $1/\phi_0(j)$ penalize discrepancies on nodes of lower degree more than differences on neighboring nodes of higher degree.

- Effectively, we only need to keep the first K terms, where $K \ll N$, in the identity for D. The accuracy of the approximation depends on the value of K, the speed of the decay of the eigenvalues $1 > \lambda_i \geq 0$ (for $i = 1, 2, \ldots, N-1$), and the exponent t. For a fixed accuracy, a larger t implies fewer terms in the sum.

In other words, Eq. 3.10 ca be expressed as a Euclidean distance

$$\mathcal{D}^2(n, m; t_n, t_m) \simeq \|\Psi_{t_n}(n) - \Psi_{t_m}(m)\|^2 \tag{3.13}$$

in a low-dimensional "diffusion space". The coordinates of nodes n and m in this space are given by a diffusion map at time scales t_n and t_m, respectively.

* * *

In this chapter, I introduced stochastic block models of mixed membership, which extend block models (Holland and Leinhardt, 1975) to include mixed-membership in a hierarchical Bayesian framework. I presented summaries of two successful applications of such models in the context of social and protein interaction networks (Airoldi et al., 2006c,d, 2007b). I discussed similarities and differences between stochastic block models of mixed membership and latent space models (Hoff et al., 2002; Handcock et al., 2007; Airoldi, 2007; Blei and Fienberg, 2007). I concluded by situating in the context of this thesis some recent developments in the mathematics of diffusion that bear relevance to the proposed methodology for statistical network analysis.

Chapter 4

Complexity and Integration

In the previous chapter, I developed generative models of networks where the identity of each node was the only attribute that was observed. In this chapter, I develop Bayesian mixed-membership models of objects' attributes. I then develop models where such objects are the nodes of a graph. The resulting *integrated models* can accommodate measurements on relations and attributes involving objects of different types, along with the corresponding sets of latent variables, in a hierarchical Bayesian framework. I describe a multivariate generalization of models of attributes and relations that is amenable to theoretical analysis—to be pursued in future work. This modeling effort informs a discussion of alternative strategies for integrating complex data. Two flavors of integration strategies emerge that are best suited to support *descriptive* and *predictive* analyses.

4.1 Heavy-Tailed Attributes

In this section, I develop statistical models for estimating latent patterns from attribute data with a heavy-tailed distribution. The notion of *contagion*, i.e., the dependence among multiple occurrences of the same attribute is introduced to express variability profiles induced by heavy tails.

Furthermore, contagion is a convenient analytical formalism to characterize semantic themes such as *biological context.* Model variants tailored to different properties of the data are explored, and a general scheme for approximate posterior inference is presented, which is based on variational methods.

Example 20. *A fundamental problem in the serial analysis of gene expression (SAGE) data is that of identifying temporal patterns of gene expression (i.e., latent distributions over a predetermined sequence of epochs) that can help explain a biological process from a large pool of observed temporal gene expression profiles (Blackshaw et al., 2004; Cai et al., 2004). The set of latent expression patterns can then be used for suggesting hypotheses and further analyses, or for making predictions.*

Example 21. *A recent problem in text and natural language processing is that of identifying topics, i.e., latent distributions over words in the vocabulary, that best explain a collection of documents (Minka and Lafferty, 2002; Blei et al., 2003; Erosheva et al., 2004; Blei and Lafferty, 2006). The set of topics provides a low-dimensional representation of each document and can be used for organizing and browsing the collection of documents efficiently.*

The description of the methodology in this section exploits the intuition developed in the biological context of Example 20.

From a methodological perspective, the task of identifying latent temporal patterns is essentially an allocation problem; observed gene expression profiles need be allocated to latent temporal patterns. The goal is to make inference on: (i) the number of latent patterns, (ii) a numerical description of the patterns themselves, and (iii) the mixed membership of the observed gene expression profiles to latent patterns. This is an instance of the more general problem of allocating observed sequences, i.e., longitudinal representations of objects in terms of an attribute, to latent sequential patterns, where each observation is allowed to be the measurable manifestation of more than one pattern. In the context of serial analysis of gene expression (SAGE), Cai et al. (2004)

introduce a variant of K-means algorithm that minimizes a non-standard scoring function, which combines the Chi-square statistic (to measure the strength of co-expression) with the Poisson distribution (to measure the likelihood of the expression level of genes at each epoch). Approaches based on clustering methods, however, constrain the expression level of a gene at each epoch to follow the expression profile typical of a single pattern. In other words, such approaches entail *unique* membership of observations to patterns, rather than *mixed* membership.

Models of mixed membership have been successfully applied in the context of different problems (e.g., Pritchard et al., 2000; Rosenberg et al., 2002; Xing et al., 2003a; Minka and Lafferty, 2002; Blei et al., 2003; Griffiths and Steyvers, 2004; Buntine and Jakulin, 2004; Blei and Lafferty, 2006). Such models, however, fall short of accommodating the marginal variability profiles of observed attributes, jeopardizing the accuracy and the interpretability of the inferences. Existing models appear to be unsuitable for the biological application to the SAGE data in part because of the assumption of *independence*, as discussed in Section 4.1.1. Below, I shall refer to popular models based on such an assumption as *independence models*.

4.1.1 The Data and Goals of the Analysis

Serial analysis of gene expression (SAGE) is a technology that quantitatively measure the copy numbers of mRNA transcripts, simultaneously for a large number of genes in a biological sample, such as a cell population or a tissue (Vesculescu et al., 1995). This technology is used to aid the discovery of gene expression profiles that characterize functional processes of interest, and to compare and catalog new genes.

A SAGE experiment begins by sampling a total of B transcripts at random from a biological sample under some specific condition (e.g., a cell cycle stage), and then use N gene-specific tags to probe the existence of possible genes in each of the B transcripts. Let $X_b = (X_{b1}, X_{b2}, \ldots, X_{bN})^T$,

such that $X_{bn} \in \{0, 1\}$ and $\sum_n X_{bn} = 1$, be a *unit-base* indicator vector recording the probing results for transcript b (i.e., $X_{bn} = 1$ indicates that gene n is present on transcript b). The number of mRNA copies of a gene n, denoted by Y_n, and the vector of copy counts for all genes (i.e., an expression profile), $Y = (Y_1, Y_2, \ldots, Y_N)^T$, can then be simply expressed as:

$$Y_n = \sum_{b=1}^{B} X_{bn}, \qquad Y = \sum_{b=1}^{B} X_b. \tag{4.1}$$

Note that Y_n's are each binomial distributed, controlled by gene-specific parameters $p_{1:N}$ each captures the probability of occurrence of gene on a random transcript, and a common sample size parameter B. When multiple cellular conditions are of interest, e.g., stage sequences in a cell cycle, an additional index will denote the specific conditions, e.g., Y^t, for measurements obtained at time t.

The main random quantities of interest are: the observed *gene expression levels* Y_n^t's, for the n-th gene at the t-th epoch; the observed *gene expression profiles* $Y_n^{1:T}$'s, for the n-th gene; and the latent *gene expression patterns*, e.g., $p_k^{1:T}$ or $\lambda_k^{1:T}$, for the k-th theme, as defined in Pritchard et al. (2000) and in the basic model of Sections 4.1.2, respectively. Technically, the latent gene expression patterns are multivariate emission probabilities for the gene expression levels, conditionally on the *active* membership of that gene. The notation I adopt puts forward the set of parameters underlying a specific distribution, e.g., $\lambda_k^{1:T}$ is a vector of Poisson rates, which control the expression levels of those genes that are expressed according to the k-th pattern. For example, whenever the n-th gene is expressed according to the k-th pattern I shall write

$$Y_n^{1:T} \sim \big[\ Pois(\lambda_k^1), \ldots, Pois(\lambda_k^T)\ \big]\ .$$

Analytical Justifications of Contagion Occurrences of the same gene under single and multiple conditions are not independent of one another, because they are sampled from a cell population or

a tissue that provides a specific *biological context*. Contagion processes provide a useful analytical mechanism to capture this notion. The two proposed generative models for analyzing temporal gene expression profiles $\{Y_n^{1:T}\}_{n=1}^N$, that instantiate the contagion process, are based on the Poisson and the negative-binomial distributions of integer counts, at multiple levels. For a review of various parameterizations of the negative-binomial and the corresponding estimators refer to Airoldi et al. (2005c), Johnson et al. (1992) and Kadane et al. (2006).

These choices were motivated by few main considerations. The Poisson distribution offers a computational advantage over the binomial distribution. It can be safely assumed that the gene-specific probabilities of occurrence $p_{1:N}$ are very small, given that there is a large amount of transcripts present in a specific biological sample. Consequently, it is reasonable as well as computationally efficient to approximate the binomial probabilities with Poisson probabilities. The sampling algorithms underlying both the Poisson and negative-binomial distributions lead to marginal and conditional[1] distributions for the gene expression levels with desirable properties. Assuming Poisson or negative-binomial conditional emission probabilities relaxes the assumption that, in the (sequential) sampling process described in Section 4.1.1, subsequent observed instances of the same gene tag are independent. In fact, such independence leads to binomial conditional emission probabilities (Pritchard et al., 2000). The dependence among different observations of the same gene tag at the conditional level is one aspect of the notion of contagion. Another aspect of contagion is found at the marginal level. Recall that ideally patterns can be interpreted as *biological or functional contexts*. Following the intuition that each gene may be expressed under multiple biological contexts to a different degree, the probability of observed gene expression levels, Y_n^t, is modeled as a mixture of conditional emission probabilities, where the gene-specific mixture weights given by the mixed membership vectors, θ_n, are constant over time (or across experimental conditions). The mixing leads to marginal distributions that are more skewed than the corresponding conditional distributions and this is the *contagion effect* one is most likely to

[1] Conditionally on the *active* membership.

encounter in the literature (e.g., see Simon, 1955). For example, in the case where the conditional probabilities are Poisson, their mixing would increase the variability of the expression levels. A formal model of contagion that encodes this intuition is the negative-binomial model, which arises as an infinite Gamma mixture of Poisson distributions. These arguments support our distributional choices. Furthermore, the marginal distributions that encode contagion fit well the observed expression levels.

To summarize, contagion processes are the result of latent regularities present in structured data, such as the SAGE profiles under study. The fact that genes may be expressed in multiple biological contexts implies a hierarchical mixture of emission probabilities, which ultimately leads to the over-dispersion of gene expression levels. Although this second characteristic of contagion processes is more common in the literature, there is an subtle point to notice in latent aspect models that feature independence of subsequent observed instances of the same gene tag (Pritchard et al., 2000; Minka and Lafferty, 2002; Blei et al., 2003). Specifically, if themes are modeled as multinomial distributions, then Dirichlet distributed mixing weights will not alter the mean-to-variance ratio of the marginal distribution, which is still multinomial. Rather, the main effect of mixing is an increased variability.

Empirical Evidence The data set that motivates this modeling effort is the set of mouse retinal SAGE libraries analyzed in Cai et al. (2004). The raw mouse retinal data consists of 10 SAGE libraries (38,818 unique genes that appeared more than twice in the sample) from developing retina taken at 2-day intervals, ranging from embryonic day to postnatal day, and adult, for total of 10 epochs (Blackshaw et al., 2004). Of the 38,818 genes, 1,467 that appeared more than 20 times in at least one of the 10 libraries were selected. These 1,467 genes were purported as the potentially most biologically relevant because of their high frequency of occurrence. The data analyzed in this paper consists of the pool of observed expression profiles $(Y_n^1, Y_n^2, \ldots, Y_n^{10})$ for the 1,467 selected genes, measured at ten epochs during the development period. Before fitting the models, I tested

Table 4.1: Methods-of-Moments estimates of negative-binomial parameters for gene expression levels in mouse retinal cells of at 10 different stages of development Cai et al. (2004). A discussion of the estimators is given in Airoldi et al. (2005c).

Epoch	mean	var.	$\frac{\sqrt{\text{var.}}}{\text{mean}}$	σ	ξ
1	30.1172	150.8648	2.2381	11.1733 ± 0.3655	4.3000 ± 0.2155
2	26.5542	163.8892	2.4843	9.8514 ± 0.4075	6.1021 ± 0.3304
3	28.1718	155.4820	2.3493	10.4516 ± 0.2936	2.9376 ± 0.1448
4	31.5446	204.2503	2.5446	11.7029 ± 0.3267	3.2591 ± 0.1588
5	26.0307	94.4013	1.9043	9.6572 ± 0.4154	6.4720 ± 0.3562
6	26.6489	82.0171	1.7543	9.8866 ± 0.2118	1.5748 ± 0.0795
7	27.3122	82.0405	1.7331	10.1327 ± 0.2491	2.1565 ± 0.1066
8	25.1990	53.6102	1.4586	9.3487 ± 0.2637	2.6407 ± 0.1319
9	27.1513	89.7169	1.8178	10.0730 ± 0.4472	7.2014 ± 0.4008
10	20.8160	81.2509	1.9757	7.7226 ± 0.5975	16.8959 ± 1.3156

the distributional assumptions discussed in Section 4.1.1 on the SAGE data at hand.

Table 4.1 reports summary statistics and estimates for the negative-binomial parameters described in Airoldi et al. (2005c). The exploratory data analysis confirms the expected over-dispersion of the gene counts, entailed by the *mixture of Poisson distributions* assumption. Moreover, the estimates of the extra-Poissonness parameter δ are all positive[2] with very high probability, as indicated by a quick inspection of the corresponding standard deviations. Lastly, I note that the log transformation $\zeta = \log(1 + \delta)$ is effective in reducing the heavy tail of the distribution of δ. Thus, it is preferable to work on the ζ scale, where a simple prior is sensible.

In conclusion, the SAGE data analyzed here are over-dispersed, i.e., variance $>$ mean. Thus models that treat the random variables $\{X_n^{1:B}\}$ as Bernoulli processes (e.g., Pritchard et al., 2000; Rosenberg et al., 2002) are not appropriate for the SAGE data at hand. Such an assumption leads to clustering models based on Multinomial latent patterns and binomial emission probabilities for feature counts (Blei et al., 2003; Griffiths and Steyvers, 2004; Buntine and Jakulin, 2004; Blei and Lafferty, 2006), which are not warranted in this context.

[2]Recall that as $\delta \to 0$ the negative-binomial density degenerates into a Poisson density.

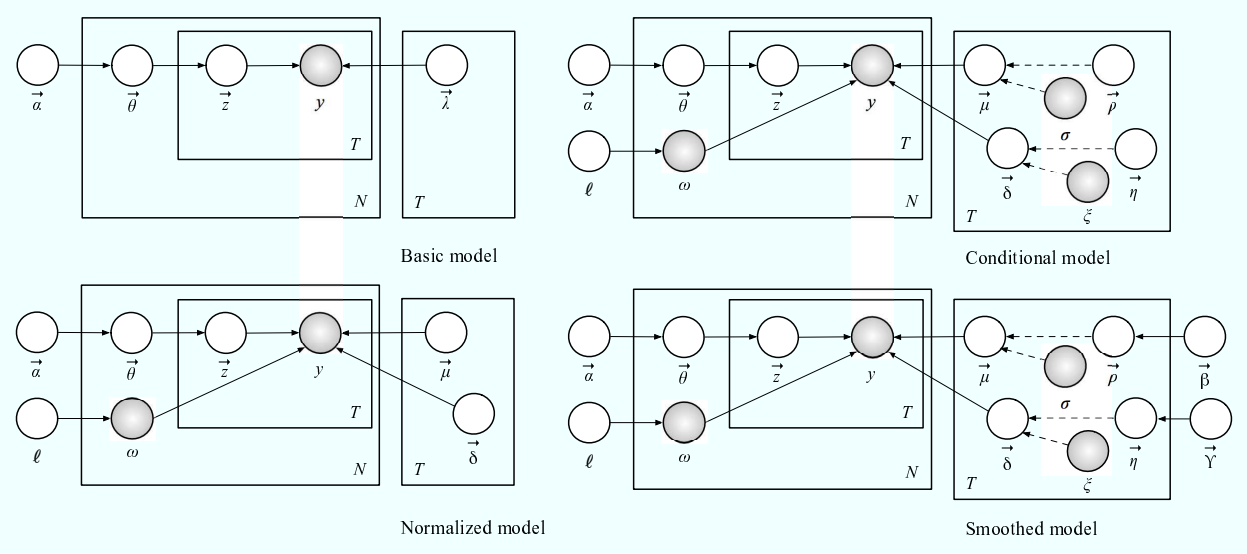

Figure 4.1: Graphical representation of the generative processes of contagion based on the Poisson (top left) and negative-binomial sampling schemes. The representation for the processes of contagion based on the Poisson sampling scheme for the non-basic models are easily obtained, by removing the part of the graphical models depending on δ. In fact, recall that δ is the extra-Poissonness parameter, and as $\delta \to 0$ the negative-binomial density converges to the corresponding Poisson limit. See Johnson et al. (1992) for more details.

4.1.2 Model Specifications

In this section, I fully specify the two hierarchical Bayesian generative processes for allocating SAGE profiles to temporal expression patterns in an unsupervised fashion. These models capture *biological context* through the notion of contagion. Recall that the observations consist of sequences of counts $(Y_n^1, Y_n^2, \ldots, Y_n^T)$ that measure the abundances of the n-th gene in the target cell or tissue across epochs 1 though T. The models introduced below need the following two assumptions: (i) a fixed number, K, of latent expression profiles exists; (ii) genes are expressed under different profiles to different degrees (mixed membership).

Poisson Generative Process The first generative process is based on the Dirichlet and Poisson distributions. There are four flavors of the Dirichlet-Poisson generative process: basic (bDiP), normalized (nDiP), conditional (cDiP), and smoothed (sDiP).

The *basic* model explicitly posits the *mixed-membership* of genes to latent patterns by associating each gene with a Dirichlet vector of probabilities, $\boldsymbol{\theta}_n$. The observed expression profile $Y_n^{1:T}$ of the n-th gene, assuming K latent expression profiles, is generated as follows.

1. Sample $\boldsymbol{\theta}_n \sim Dirichlet_K(\boldsymbol{\alpha})$
2. For each epoch $t = 1, \ldots, T$
   2.1. Sample $\boldsymbol{z}_n^t \sim Multinomial(\boldsymbol{\theta}_n, 1)$
   2.2. Sample $y_n^t \sim Poisson(\lambda_{tk} | z_{nk}^t = 1)$.

The genes are the sampling units in SAGE experiments, and the total volume of their expressions often vary over time. Recovering *calibrated* expression profiles that do not depend on the total expression volume is desirable. To this extent, I posit the *normalized* model, which rescales the samples (i.e., the genes) according to their different sizes (the total expression volumes), and ultimately improves the estimates. In the basic model, the matrix $\boldsymbol{\lambda} \equiv \{\lambda_{tk}\}$ contains the rates that govern the expression level of genes at T different epochs for each of the K different latent profiles. In the normalized model, the expected expression level of the n-th gene at time t for profile k is written as follows,

$$\lambda_{tk} = \omega_n \cdot \mu_{tk}, \tag{4.2}$$

where ω_n is scalar and observed, and denotes the total expression level of the n-th gene as a multiple of a fixed total expression level β used as a reference expression level. This new parameter β may a fixed pre-determined value, estimated via, e.g., empirical Bayes (Carlin and Louis, 2005), or given a distribution as part of a full Bayesian analysis (Airoldi et al., 2006a).

Note 1. *In both the basic and the normalized models above, the rows of the parameter matrices* $\boldsymbol{\lambda}$ *and* $\boldsymbol{\mu}$ *control the rates at which genes are expressed. In particular,* λ_{tk} *and* μ_{tk} *encode the expected expression level of genes at time* t *for profile* k*. Since profiles are by definition not observable, none of these parameters can be estimated directly from the data.*

Rows of the normalized rate matrix $\vec{\mu}$ are reparameterized with the sum/ratio parameterization, i.e., for every epoch t the following transformation is applied

$$(\mu_{t1}, \mu_{t2}, \ldots, \mu_{tK}) \longrightarrow (\sigma_t, \rho_{t1}, \rho_{t2}, \ldots, \rho_{tK}), \tag{4.3}$$

where the sum parameter $\sigma_t := \sum_{k=1}^{K} \mu_{tk}$, the ratio parameters $\rho_{tk} := \frac{\mu_{tk}}{\sigma_t}$, and the constraint that $\sum_{k=1}^{K} \rho_{tk} = 1$ makes the ratio parameter ρ_{tK} redundant for each t. This reparameterization leads to the *conditional* model, where the sum parameters $(\sigma_1, \sigma_2, \ldots, \sigma_T)$ are directly estimable from the data, and inference can be can carried out conditionally on them. This is possible since the parameters $\sigma_{1:T}$ encode the total normalized expression levels at time t (sum of the expression levels over the K latent patterns), which is an observable quantity as it does not depend on the latent profiles. Conditioning on the MLEs for the total expression parameters, σ_t, leads to a new allocation problem where the differential expression levels of genes under the K profiles needs be inferred. In other words, the total expression level at each time t needs be allocated among the latent patterns, given a constraint on their sum and a direct estimate of σ_t.

Lastly, I introduce the *smoothed* model, which posits a prior for the differential expression rate parameters to smooth their estimates. In the smoothed model I assume that the differential expression levels are sampled

$$\boldsymbol{\rho}_{t\,\cdot} \sim Dirichlet_{\,K}\,(\boldsymbol{\beta})$$

for each epoch $t = 1, 2, \ldots, T$. See Figure 4.1. In principle, it is possible to posit a prior distribution on the total expression rate parameters as well. A brief analysis of the observed total rates suggests that it is appropriate to apply a logarithmic transformation on them to stabilize the variability, and one can introduce a Gaussian prior on the transformed rates; however, an inspection of the total rates σ_t over time (see Table 4.1) suggests that some other phenomenon is possibly going on, which leads to a decreasing occurrence of the genes in the SAGE libraries. Therefore

the observed total rates are used to inform our inferences directly, as in the conditional model. Smoothing the overall rates $\{\sigma_{tk}\}$ would impose a model on data that cannot be justified, since it is not clear why the overall rates are declining. This would cast some doubts on the interpretability of the inferences such a model would lead to.

Summarizing, the Dirichlet-Poisson generative process possesses a few advantages: (i) the sampling scheme encodes contagion in the sense that multiple occurrences of the same gene tag at the same epoch depend on one another, given their active memberships to a specific latent expression pattern; (ii) the sampling scheme arises naturally in SAGE biological experiments discussed in Section 4.1.1; (iii) computing Poisson probabilities is more efficient than computing binomial probabilities, since binomial coefficients need not be evaluated.

Negative-Binomial Generative Process The generative process of contagion based on the negative-binomial sampling scheme is similar in spirit to the previous one based on the Poisson sampling scheme. A formal treatment of the models is given in Airoldi et al. (2005c). Intuitively, the negative-binomial distribution has two parameters that control mean and variance; furthermore, its variance is always greater than its mean—a useful property that replicates the observed over-dispersion of gene expression levels. The negative-binomial density can be written as a Poisson density with an extra parameter δ that controls the amount of extra-Poisson variability. Thus,

$$NB\left(\, y_n^t \;\middle|\; \omega_n\mu_t, \omega_n\delta_t\right) = \frac{\Gamma(y_n^t + \kappa_t)}{y_n^t!\Gamma(\kappa_t)} \frac{(\omega_n\delta_t)^{y_n^t}}{(1+\omega_n\delta_t)^{(y_n^t+\kappa_t)}},$$

where $\kappa_t := \frac{\mu_t}{\delta_t}$ for convenience of notation. In the normalized model, $\{\mu_{tk}\}$ are the profile-specific Poisson rates and $\{\delta_{tk}\}$ are profile-specific extra-Poissonness parameters. The conditional model then follows from the application of the sum/ratio parameterization (see Equation 4.3) to both sets

of parameters

$$(\mu_{t1}, \mu_{t2}, \ldots, \mu_{tK}) \longrightarrow (\sigma_t, \rho_{t1}, \rho_{t2}, \ldots, \rho_{tK})$$
$$(\delta_{t1}, \delta_{t2}, \ldots, \delta_{tK}) \longrightarrow (\xi_t, \eta_{t1}, \eta_{t2}, \ldots, \eta_{tK}).$$

Lastly, the smoothed model imposes probabilistic constraints on both the differential expression levels and the differential extra-Poissonness parameters by assuming that they are independent samples from two Dirichlet distributions with distinct sets of underlying constants,

$$\boldsymbol{\rho}_{t\,.} \sim Dirichlet_{\,K}\,(\boldsymbol{\beta}) \qquad \text{and} \qquad \boldsymbol{\eta}_{t\,.} \sim Dirichlet_{\,K}\,(\boldsymbol{\gamma}),$$

for each epoch $t = 1, 2, \ldots, T$. See Figure 4.1.

4.1.3 Estimation and Inference

In order to obtain the posterior for the latent variables,

$$p\left(\{\theta_n, z_n^{1:T}\}_{n=1}^N \,\middle|\, \{y_n^{1:T}\}_{n=1}^N, \alpha, \{\lambda_k^{1:T}\}_{k=1}^K\right) = \\ = \frac{p\left(\{\theta_n, z_n^{1:T}\}_{n=1}^N, \{y_n^{1:T}\}_{n=1}^N \,\middle|\, \alpha, \{\lambda_k^{1:T}\}_{k=1}^K\right)}{p\left(\{y_n^{1:T}\}_{n=1}^N \,\middle|\, \alpha, \{\lambda_k^{1:T}\}_{k=1}^K\right)}, \tag{4.4}$$

one needs to evaluate the likelihood in Expression 4.4, which is given by an integral with no closed form solution—the denominator. Thus I develop a mean-field approximation to the posterior, which involves the substitution of an integrable lower bound for the likelihood. The mean-field approximation involves positing a simple distribution, q, over the latent variables, which depends upon an extra set of (variational) free parameters, $\{\nu_n, \phi_n^{1:T}\}_{n=1}^N$ in this case. The free parameters are then set to minimize the Kullback-Leibler divergence between the true and approximate posteriors. This is equivalent to maximizing a lower bound for the likelihood within each E-step,

over the free parameters, and then compute pseudo-expectations for the latent variables using the maximized lower bound. The overall inference algorithm is a variational EM scheme. At each iteration, the EM algorithm employs the mean-field approximation to carry out the E-step (just discussed above) and then employs a regular M-step, where the maximum likelihood estimates of the model parameters, e.g., $(\alpha, \{\lambda_k^{1:T}\}_{k=1}^K)$ for the basic model, are updated by maximizing the lower bound for the likelihood, over such parameters. These two steps are iterated till convergence of the lower-bound for the likelihood.

The variational EM scheme just described practically translates into a coordinate ascent algorithm, where parameters are naturally organized into batches with similar semantics. The parameter updates corresponding to the model variants considered above are summarized in Table 4.2.

A General Bayesian Formalism for Latent Aspects Analysis The variational inference scheme developed for the two models of counts is actually quite general. In fact, the free parameter updates (that are used to maximize the lower bound for the likelihood within each E-step) take a generic form applicable to all different conditional emission probability functions considered above, e.g., Table 4.2. Furthermore, for generic conditional emission probabilities $p(y_n^t|\beta_k^t)$ for all (n, t, k), with parameter set $\{\beta_k^{1:T}\}_{k=1}^K$, the following general free parameter updates can be used

$$\phi_{ntk}^* \propto \Upsilon \cdot p\left(y_n^t \mid \beta_k^t \right),$$

where $\Upsilon := e^{\mathbb{E}_q[\log \theta_{nk}]}$ as in Table 4.2. The updates for $\nu_{nk}^* = \alpha_k + \sum_t \phi_{ntk}$ remain unchanged.

The generality of the approximate E-step in latent aspects analysis that feature one latent group indicator, z_n^t, for each gene-epoch pair (n, t) is due the specific hierarchical formulation of our models. Such a formulation posits exchangeable measurements on features, e.g., gene expression levels at each epoch. Different conditional emission probabilities only lead to different estimators for the corresponding parameters, $\{\beta_k^{1:T}\}_{k=1}^K$, in the M-step.

Table 4.2: The table quotes the parsimonious mean-field approximation for the various models. The parsimonious mean-field approximation posits one latent expression profile indicator z for each (gene,epoch) pair. Note that $\Upsilon := e^{\mathbb{E}_q[\log \theta_{nk}]}$, and Po, NB, are short for $Poisson$, and $Negative\text{-}Binomial$, respectvely. ** Alternatively use the Method of Moments described in Airoldi et al. (2005c) pretending to observe pseudo counts $\{\phi^t_{nk} \cdot y^t_n\}$ as the expression levels of the $n\text{-}th$ gene according to the $k\text{-}th$ latent theme.

	Poisson	Negative-Binomial
Basic	$\nu^*_{nk} = \alpha_k + \sum_t \phi_{ntk}$ $\phi^*_{ntk} \propto \Upsilon \cdot Po\left(y^t_n \mid \lambda_{tk} \right)$ $\lambda^*_{tk} = \frac{\sum_n \phi_{ntk} y^t_n}{\sum_n \phi_{ntk}}$ α^*_k with Newton-Raphson	
Norm.	$\nu^*_{nk} = \alpha_k + \sum_t \phi_{ntk}$ $\phi^*_{ntk} \propto \Upsilon \cdot Po\left(y^t_n \mid \omega_n \mu_{tk} \right)$ $\mu^*_{tk} = \frac{\sum_n \phi_{ntk} y^t_n}{\sum_n \phi_{ntk} \omega_n}$ α^*_k with Newton-Raphson	$\nu^*_{nk} = \alpha_k + \sum_t \phi_{ntk}$ $\phi^*_{ntk} \propto \Upsilon \cdot NB\left(y^t_n \mid \omega_n \mu_{tk} \right)$ $\mu^*_{tk} = \frac{\sum_n \phi_{ntk} y^t_n}{\sum_n \phi_{ntk} \omega_n}$ $\delta^*_{tk} = L\text{-}BFGS$ ** α^*_k with Newton-Raphson
Cond.	$\nu^*_{nk} = \alpha_k + \sum_t \phi_{ntk}$ $\phi^*_{ntk} \propto \Upsilon \cdot Po\left(y^t_n \mid \omega_n \sigma_t \rho_{tk} \right)$ $\rho^*_{tk} = \frac{\sum_n \phi_{ntk} y^t_n}{\sum_n \phi_{ntk} \omega_n \sigma_t}$ α^*_k with Newton-Raphson	$\nu^*_{nk} = \alpha_k + \sum_t \phi_{ntk}$ $\phi^*_{ntk} \propto \Upsilon \cdot NB\left(y^t_n \mid \omega_n \sigma_t \rho_{tk} \right)$ $\rho^*_{tk} = \frac{\sum_n \phi_{ntk} y^t_n}{\sum_n \phi_{ntk} \omega_n \sigma_t}$ $\eta^*_{tk} = L\text{-}BFGS$ ** α^*_k with Newton-Raphson

Related Work There is a simple connection between the algorithms developed here and the PoissonC and PoissonL algorithms introduced by Cai et al. (2004). In the problem at hand the goal is to allocate the observed temporal expression profiles $\{Y^{1:T}_n\}_{n=1}^N$ into, say, K patterns or clusters. Recall that the K-means unsupervised clustering algorithm searches for K means $m_{1:K}$ that minimize

$$MSE = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} \mathbb{I}\left(y^{1:T}_n \in k \right) \left\| y^{1:T}_n - m_k \right\|^2 .$$

That is, the means $m_{1:K}$ are the centers of K clusters in the sense of Euclidean norm. The PoissonC and PoissonL algorithms introduced by Cai et al. (2004) substitute the Euclidean norm in the

equation with the chi-squared score,

$$\chi^2(n,k) = \sum_{t=1}^{T} \frac{\left(y_n^t - \hat{\mu}_{tk}\, \hat{\omega}_n \right)^2}{\hat{\mu}_{tk}\, \hat{\omega}_n},$$

and the negative log-likelihood,

$$\ell(n,k) = -\sum_{t=1}^{T} \log \left(\frac{e^{-(\hat{\mu}_{tk}\ \hat{\omega}_n)}\ (\hat{\mu}_{tk}\ \hat{\omega}_n)^{y_n^t}}{y_n^t!} \right),$$

respectively. The normalized model based on the Poisson distribution is an extension of the PoissonL algorithm, where Dirichlet distributed mixed-membership vectors are introduced, θ_n, not known in advance. In the PoissonL algorithm the mixed-membership vectors θ_n are known, i.e., for the n-th gene it follows that

$$\theta_{nk} = \begin{cases} 1 & \text{if } k = j_n \\ 0 & \text{otherwise,} \end{cases}$$

where $j_n = \arg\min \left\{ L(n,k) : k \in [1,K] \right\}$. This extension is similar in spirit to that introduced by Gaussian mixture to regular K-means (Blei and Fienberg, 2007). In fact,

$$\theta_{nk} = Pr\left(cluster = k \,\middle|\, data, parameters \right).$$

Note that introducing latent Dirichlet distributed mixed-membership vectors, θ_n, ties together all the data in the inference task. This has the beneficial effect of reducing the variability of pattern-specific parameters since all the gene counts are used (independently of which pattern they express the most) in estimating each such parameters. Such an improvement in the estimates is expected James and Stein (1961). Our basic Poisson model is similar to that of Canny (2004). For a technical survey of related latent aspects models in the context of text data analysis see Buntine and Jakulin

(2006).

Example 4.1 (Continued) Contagion induces a non-trivial difference in the generative process with respect to the *independence model* (Pritchard et al., 2000; Minka and Lafferty, 2002; Blei et al., 2003) that has far reaching implications for the analysis of data. For example, models of contagion provide a better fit for data in biological applications such as SAGE by providing a realistic mean-to-variance marginal ratio. A better fit helps recovering more precise mixed memberships of genes to patterns, as well as finding cleaner temporal expression patterns when compared to those found by independence models. This general issue is explored further in Section 6.2.1.

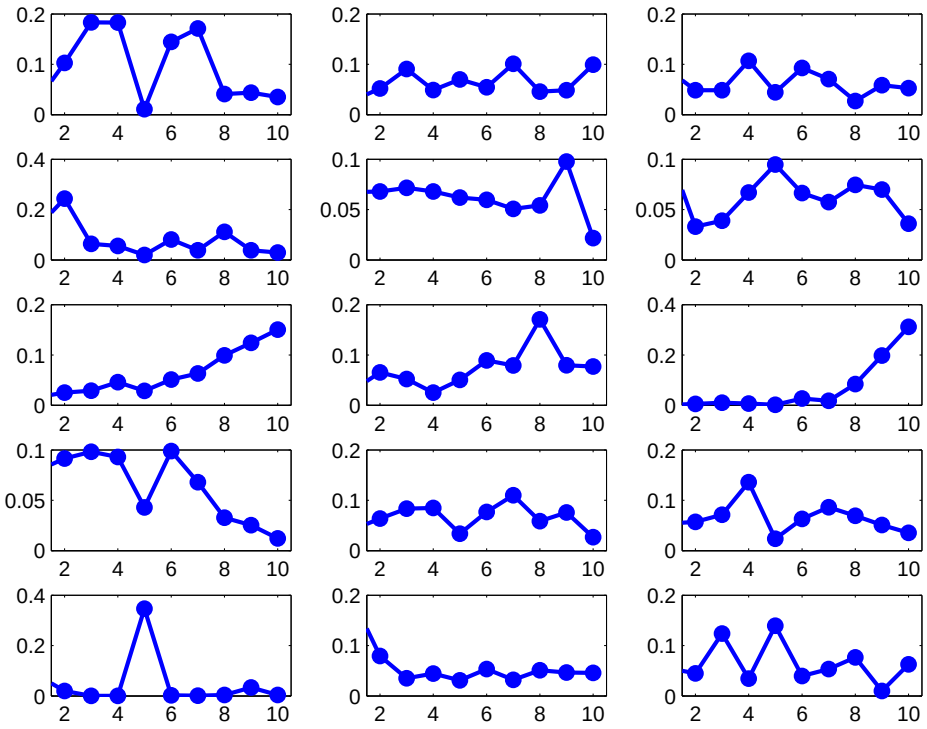

Figure 4.2: Gene expression themes learned from mouse retinal SAGE using conditional DiP.

Recall that the mouse retinal SAGE libraries analyzed in Cai et al. (2004) contain 38,818 unique genes for total of 10 epochs. At first, I perform model selection by means of a five-fold cross validation scheme, to estimate the plausible number of latent themes that best explain the data. The held-out likelihood peaked at 15 themes for cDiP, and 10 for the independence model. Figure 4.2 shows the 15 gene expression patterns inferred using the conditional Poisson model (cDIP). The variance of each pattern is not shown—the variances are so small that the variance-bars are masked by the dot symbol in our plots. Notably, the magnitude of the held-out likelihood for cDiP is about ten times larger (on the log scale) than that for the independence model, suggesting a better overall fit of cDiP to the data. Furthermore, the corresponding mixed-membership estimates $\{\theta_n\}$ are more sharply peaked; this result holds in general for over-dispersed data sets. See Figure 6.2 for an example. The result is indirectly supported by the estimates of Dirichlet hyper-parameter as well, $\hat{\alpha}_{Indep} = 1.355$ versus $\hat{\alpha}_{DiP} = 0.066$. The patterns (or clusters) shown in Figure 4.2 indeed lead to reasonable predictions of mouse retinal gene functions. For example, a preliminary biological validation of the patterns inferred using cDiP based on the GO annotation shows correlation between the latent patterns and gene functions such as photoreceptors and rhodospin, i.e., genes with similar functional annotations tend to fall into the same pattern. An in-depth analysis of the biological significance of the inferred patterns is given elsewhere (Airoldi et al., 2006f).

Modeling Choices and Inference In problems where attributes co-occur frequently (e.g., a pair of genes can be present on many transcripts), the computational gains sought after by positing models that rely on unrealistic assumptions are seldom achieved. Applications to problems that arise in computational biology, e.g., SAGE and microarray data, are one such case. Probabilistic models that replicate salient features of the data typically lead to better inferences on latent quantities of interest, e.g., the latent temporal patterns of Example 20. In the models introduced in this section, the salient features of interest are the *marginal variability* and the notion of *contagion*. The inference suggests that the inferred latent patterns can be interpreted as temporal expression

patterns that are typical of fairly distinct *functional biological contexts*—the desired outcome. This contrasts the poorly interpretable results obtained with the independence model, and makes a good case for modeling choices that "let the data tell their story." Following these thoughts, model variants tailored to different properties of biological data have been introduced, and the general inference scheme for posterior inference has been derived.

Concluding, the estimates the proposed models provide are sharper than those entailed by existent methods based on stronger independence assumptions, in the context of the SAGE analysis. This demonstrates the feasibility of a promising hierarchical Bayesian formalism for soft clustering and latent aspect analysis.

4.2 Multivariate Model Specifications

Here I present a multivariate generalization of the hierarchical models of mixed membership for attributes and relations.

4.2.1 Attributes: Hierarchical Bayesian Models of Mixed Membership

There are a number of earlier instances of mixed-membership models that have appeared in the scientific literature (e.g., see Erosheva and Fienberg, 2005). A general formulation characterizes the models of mixed-membership in terms of assumptions at four levels (Erosheva et al., 2004).

Assumption 1 (Population Level)**.** *There are K classes or sub-populations in the population of interest and J distinct characteristics. Denote by $f(x_{nj}|\beta_{jk})$ the probability distribution of j-th response variable in the k-th sub-population for the n-th subject, where β_{jk} is a vector of relevant parameters, $n \in [1, N]$, $j \in [1, J]$, and $k \in [1, K]$. Within a subpopulation, the observed responses are assumed to be independent across subjects and characteristics.*

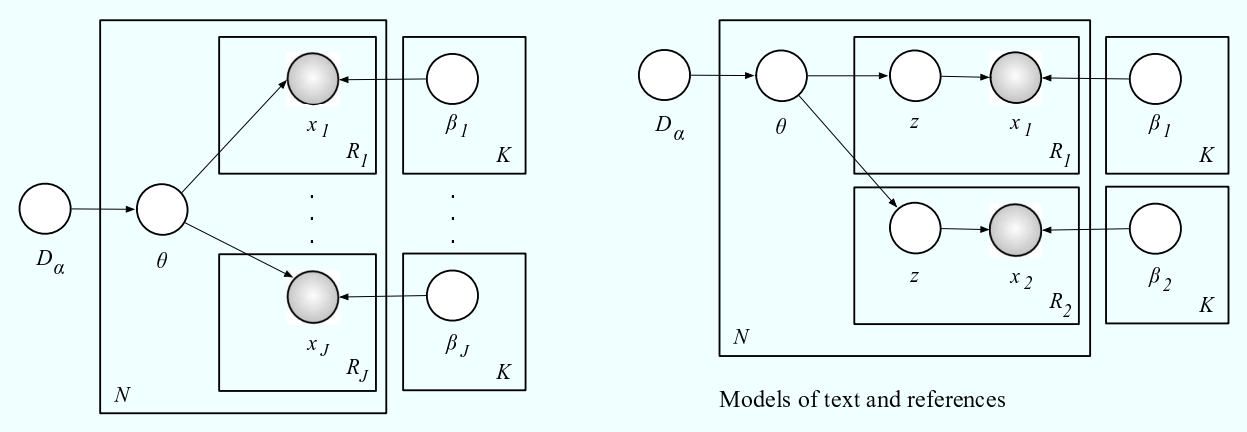

Figure 4.3: Left: A graphical representation of hierarchical Bayesian models of mixed-membership. Right: Models of text and references used in this paper. Specifically, replicates of variables $\{x_1^r, x_2^r\}$ are paired with latent variables $\{z_1^r, z_2^r\}$ that indicate which latent aspects informs the parameters underlying each individual replicate. The parametric and non-parametric version of the error models for the label discussed in the text refer to the specification of D_α—a Dirichlet distribution versus a Dirichlet process, respectively.

Assumption 2 (Subject Level)**.** *The components of the mixed-membership vector* $\theta_n = (\theta_{n[1]}, \ldots, \theta_{n[K]})'$ *represent the membership of the* n-th *subject to the various sub-populations.*[3] *The distribution of the observed response* x_{nj} *given the individual membership scores* θ_n*, is then*

$$Pr\,(x_{nj}|\theta_n) = \sum_{k=1}^{K} \theta_{n[k]} f(x_{nj}|\beta_{jk}). \tag{4.5}$$

Conditional on the mixed-membership scores, the response variables x_{nj} *are independent of one another, and independent across subjects.*

Assumption 3 (Latent Variable Level)**.** *The mixed-membership vectors,* $\theta_{1:N}$*, are independent realizations of a latent random quantity with distribution* D_α*, parameterized by vector of underlying constants* α*. The probability of observing* x_{nj}*, given the parameters, is then*

$$Pr\,(x_{nj}|\alpha, \beta) = \int \left(\sum_{k=1}^{K} \theta_{n[k]} f(x_{nj}|\beta_{jk}) \right) D_\alpha(d\theta). \tag{4.6}$$

[3] I denote components of a vector v_n with $v_{n[i]}$, and the entries of a matrix m_n with $m_{n[ij]}$.

Assumption 4 (Sampling Scheme Level)**.** *The R independent replications of the J distinct response variables corresponding to the n-th subject are independent of one another. The probability of observing $\{x^r_{n1}, \ldots, x^r_{nJ}\}_{r=1}^R$, given the parameters, is then*

$$Pr\left(\{x^r_{n1}, \ldots, x^r_{nJ}\}_{r=1}^R | \alpha, \beta\right) = \int \left(\prod_{j=1}^{J} \prod_{r=1}^{R} \sum_{k=1}^{K} \theta_{n[k]} f(x^r_{nj} | \beta_{jk}) \right) D_\alpha(d\theta). \tag{4.7}$$

The number of observed response variables is not necessarily the same across subjects, i.e., $J = J_n$. Likewise, the number of replications is not necessarily the same across subjects and response variables, i.e., $R = R_{nj}$.

Example 22 (Latent Dirichlet Allocation)**.** *The general formulation encompasses popular data mining models such as the latent Dirichlet allocation model (LDA) for use in the analysis of scientific publications (Minka and Lafferty, 2002; Blei et al., 2003). Consider a collection of documents; sub-populations correspond to latent topics, indexed by k; subjects correspond to "documents," indexed by n; $J = 1$, i.e., there is only one response variable that encodes which "word" in the vocabulary is chosen to fill a position in a text of known length, so that j is omitted; positions in the text correspond to replicates, and we have a different number of them for each document, i.e. we observe R_n positions filled with words in the n-th document. The model assumes that each position in a document is filled with a word that expresses a specific topic, so that distinct instances of the same word may be expression of different topics. In order to do so, an explicit indicator variables z^r_n is introduced for each observed position in each document, which indicates the topic that expresses the word in such position. The function $f(x^r_n|\beta_k)$ is given by the probability $Pr\left(x^r_n = 1 | z^r_n = k\right)$, which is specified as $Multinomial\left(\beta_k, 1\right)$, where β_k is a random vector the size of the vocabulary, say V, and $\sum_{v=1}^{V} \beta_{k[v]} = 1$. A mixed-membership vector θ_n is associated to the n-th document, which encode the topic proportions that ultimately inform the choice of words in that document, and it is distributed according to a Dirichlet distribution, which specifies D_α. Equation 4.8 is obtained by integrating out the topic indicator variable z^r_n at the word level—the*

latent indicators z_n^r are distributed according to a $Multinomial\ (\theta_n, 1)$.

Example 23 (Grade of Membership)**.** *The Grade of Membership model (GoM) is another specific model that can be cast in terms of mixed-membership. This model was first introduced by Woodbury in the 1970s in the context of medical diagnosis Woodbury et al. (1978) and was developed further and elaborated upon in a series of papers and in Manton et al. (1994). Erosheva (2002) reformulated the GoM model according to the specifications of Section 4.2.1. Consider disability survey data collected for the National Long Term Care Survey; there are no replications, i.e., $R_n = 1$, but several attributes of each american senior are recorded, i.e., $J = 16$ daily activities. Furthermore, the scalar parameter β_{jk} is the probability of being disabled on the activity j for a member of latent pattern k, that is, $\beta_{jk} = P(x_j = 1|\theta_k = 1)$. Dealing with binary data (individuals are either disabled or healthy), the probability distribution $f(x_j|\beta_{jk})$ is specified by a Bernoulli distribution with parameter β_{jk}. Therefore, a member n of latent profile k is disabled on the activity j, i.e., $x_{nj} = 1$, with probability β_{jk}. In other words, introducing a profile indicator variable z_{nj}, we have $P(x_{nj} = 1|z_{nj} = k) = \beta_{jk}$. Each individual n is characterized by a vector of membership scores $\theta_n = (\theta_{n1}, \ldots, \theta_{nK})$. In this model the membership scores θ_n follow the distribution D_α (for example a Dirichlet distribution with parameter $\alpha = (\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K)$. Note that the ratio $\alpha_k / \sum_k \alpha_k$ represents the proportion of the population that "belongs" to the k-th latent pattern.*

Note 2 (Related Work)**.** *It is possible to situate this formulation in a familiar landscape by discussing similarities with other unsupervised data mining methods. Recall that the problem is to group observations about N subjects $\{x_n^{1:R_n}\}_{n=1}^N$ into, say, K groups. K-means clustering, for example, searches for K centroids $m_{1:K}$ that minimize*

$$MSE = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} \mathbb{I}\left(x_n^{1:R_n} \in k \right) \left\| x_n^{1:R_n} - m_k \right\|^2,$$

where the centroids $m_{1:K}$ are centers of respective clusters in the sense of Euclidean norm. Subjects

have single group membership in K-means. In the mixture of Gaussians model, a popular model that extends K-means, the MSE scoring criterion is substituted by the likelihood $\sum_{n,k} \ell(n,k)$. The unknown mixed-membership vectors θ_n relax the single membership implicit in K-means. The connection is given by the fact that the mixed-membership vectors θ_n, i.e., the class abundances, have a specific form in K-means, i.e., for the n-th subject it follows that

$$\theta_{n[k]} = \begin{cases} 1 & \text{if } k = j_n \\ 0 & \text{otherwise}, \end{cases}$$

where $j_n = \arg\min \{ \ell(n,k) : k \in [1,K] \}$. In general, the unknown mixed-membership vectors θ_n are independent samples from D_α. Furthermore, in the general formulation of Section 4.2.1 it is possible to have more complicated likelihood structures.

4.2.2 Relations: Stochastic Block Models of Mixed Membership

The class of stochastic block models of mixed-membership is a rich class of models that is instrumental for thinking about the scientific problems outlined in Section 3.1 and amenable to theoretical analysis. A general formulation characterizes stochastic block models of mixed-membership in terms of assumptions at four levels, as follows (Airoldi et al., 2006d).

Assumption 5 (Population Level)**.** *There are K classes or sub-populations in the population of interest. Denote by $f(y_{jnm}|\eta_{gh})$ the probability distribution of the j-th response graph at the pair of nodes (n,m), where the n-th node is in the h-th sub-population, the m-th node is in the k-th sub-population, and η_{gh} contains the relevant parameters. The indices n,m run in $\mathcal{N}$, and the indices g,h run in $[1,K]$. Within sub-population pairs, the observed paired responses are assumed independent.*

Assumption 6 (Node Level)**.** *The components of the membership vector $\theta_n = (\theta_{n1},\ldots,\theta_{nK})'$*

encodes the mixed-membership of the n-th node to the various sub-populations. The distribution of the observed response y_{jnm} given the relevant, node-specific membership scores, (θ_n, θ_m), is then

$$Pr\ (y_{jnm}|\theta_n, \theta_m, \eta) = \sum_{g,h=1}^{K} \theta_{ng} f(y_{jnm}|\eta_{gh})\theta_{mh}. \tag{4.8}$$

Conditional on the mixed-membership scores, the response edges y_{jnm} are independent of one another, both across distinct graphs and pairs of nodes.

Assumption 7 (Latent Variable Level)**.** *The mixed-membership vectors, $\theta_{1:N}$, are independent realizations of a latent random quantity with distribution D_α, parameterized by a vector of underlying constants α. The probability of observing $_{jnm}$, given the parameters, is then*

$$Pr\ (y_{jnm}|\alpha, \eta) = \int \left(\sum_{g,h=1}^{K} \theta_{ng} f(y_{jnm}|\eta_{gh})\theta_{mh} \right) D_\alpha(d\theta). \tag{4.9}$$

Assumption 8 (Sampling Scheme Level)**.** *The R independent replications of the J distinct response graphs are independent of one another. The probability of observing the whole collection of graphs, $\{y_{jrnm}\}$, given the parameters, is then given by the following equation.*

$$Pr\ (\ \{y_{jrnm}\}\ |\ \alpha, \eta\ ) = \int \left(\prod_{j=1}^{J} \prod_{r=1}^{R} \prod_{n,m=1}^{N} \sum_{g,h=1}^{K} \theta_{ng} f(y_{jrnm}|\eta_{gh})\theta_{mh} \right) D_\alpha(d\theta). \tag{4.10}$$

The number of replications is not necessarily the same across different response graphs, i.e., $R = R_j$. Likewise, the block model can be response specific, i.e., $\eta = \eta_j$. More variations along these lines are possible.

A graphical representation of models in this family is given in Figure 4.4. Full model specifications immediately adapt to the different kinds of data, e.g., multiple data types through the choice of f, or parametric or semi-parametric specifications of the prior on the number of clusters through the choice of D_α.

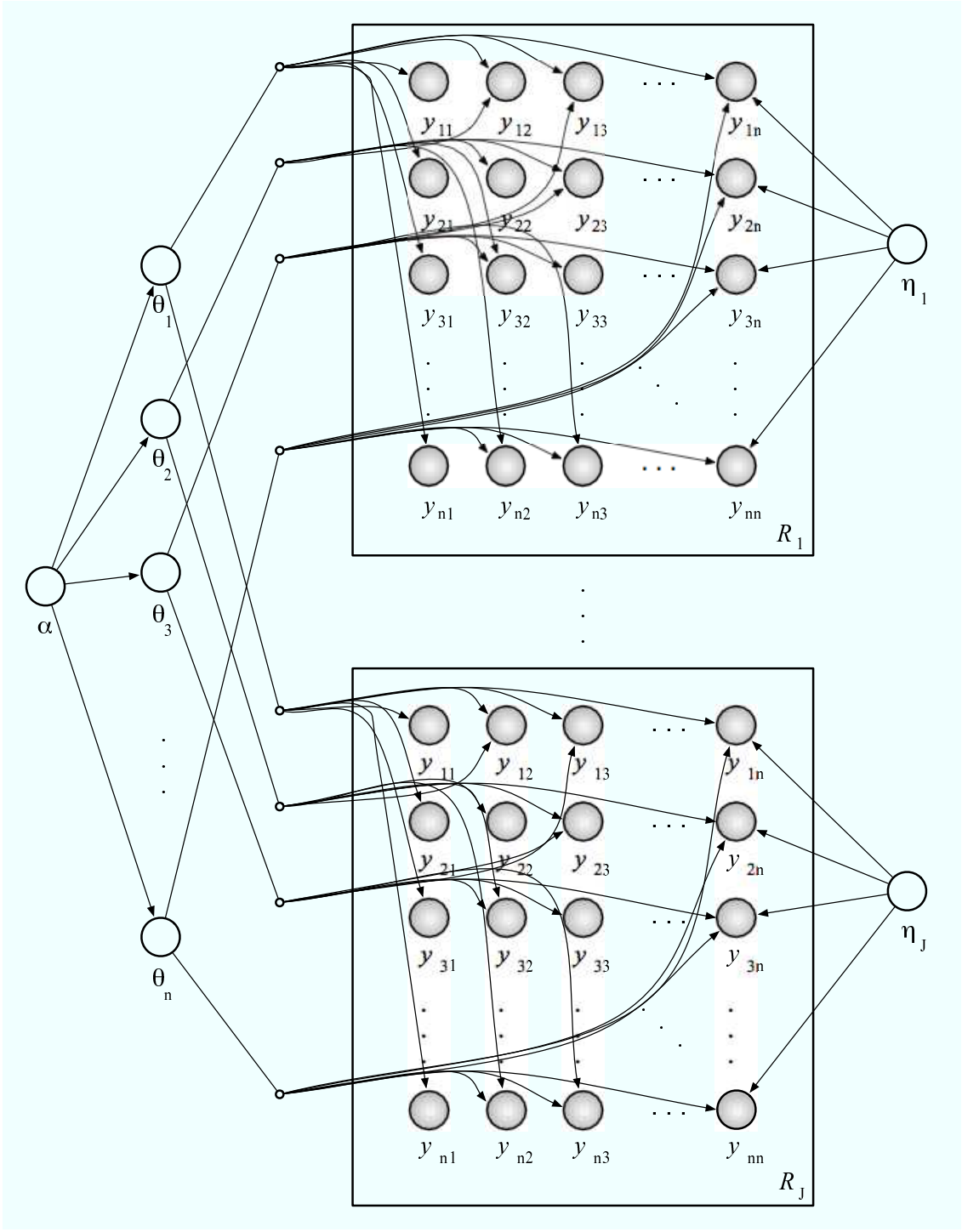

Figure 4.4: The graphical representation of stochastic block models of mixed membership using plates. For clarity, few arrows out of the block models $\eta_{1:J}$ are shown, however, all interactions y_{jrnm} depend on the corresponding block model.

Example 24 (Admixture of Latent Blocks)**.** *Airoldi et al. (2006c, 2007b) introduced the Admixture of Latent Blocks model to analyze a collection of protein-protein interactions. This model is defined by the simplest set of model specifications for a stochastic block model of mixed membership, and it was used to analyze the most basic kind of relational data. Given a single undirected unipartite graph with binary edges, the Admixture of Latent Blocks model recovers membership of nodes to clusters (i.e., the mixed membership vectors $\theta_{1:N}$) and cluster-to-cluster interaction probabilities (i.e., the block model η), under the assumption that K non-observable clusters exist. Using this model on protein-protein interaction data: sub-populations correspond to non-observable "stable protein complexes", indexed by k; nodes correspond to "proteins", indexed by n; there is only one response variable that encodes whether a pair of proteins interacts or not, so that j is omitted;*

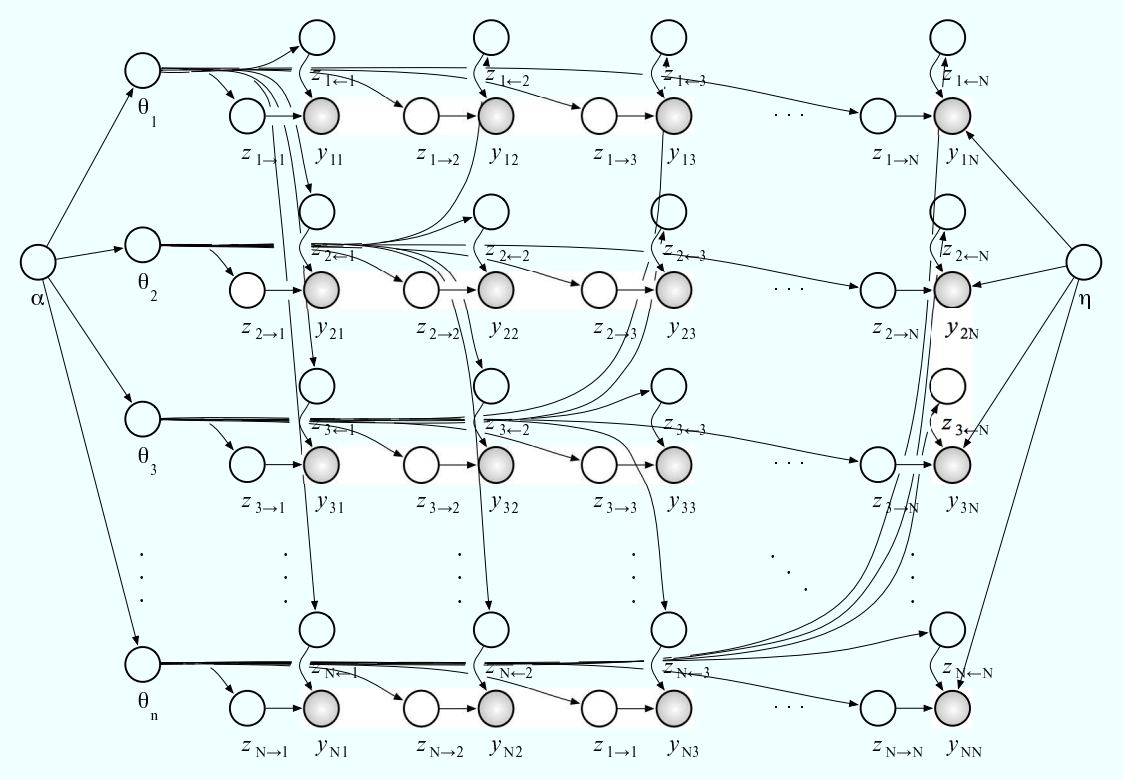

Figure 4.5: The graphical representation of the admixture of latent blocks introduced by Airoldi et al. (2006c) using plates. Note that only few arrows out of the block model η have been drawn, for clarity, however all the interactions y_{nm} depend on it.

there is only one replicate, since the interactions have been measured with an experimental procedure such as "Yeast Two Hybrid" under a single experimental condition. The model assumes that each interaction in the collection is either present or absent given the memberships to specific protein complexes of the pair of single proteins involved. That is, each protein participates in the various interactions as a member of possibly different protein complexes. In order to simplify the inference, an explicit pair of indicator variables $(z^{\rightarrow}_{nm}, z^{\leftarrow}_{nm})$ *is introduced for each interaction in the observed collection, which indicates the protein complexes that the two proteins are members of as they interact. The function* $f(y_{nm}|\eta_{gh}) = Pr\ (y_{nm} = 1|z^{\rightarrow}_{nm} = g, z^{\leftarrow}_{nm} = h) = Bernoulli\ (\eta_{gh})$*, where* η_{gh} *is the probability that a protein in complex* g *interacts with a protein in complex* h*. A mixed-membership vectors* $\theta_{1:N}$ *encode the expected protein complex proportions. They are distributed according to* D_α*, i.e., a Dirichlet distribution. Equation 4.8 is obtained by integrating out the protein complex indicator variables* $(z^{\rightarrow}_{nm}, z^{\leftarrow}_{nm})$ *at the interactions level—the latent indicators* $z^{\rightarrow}_{nm}$ *are distributed according to a* $Multinomial\ (1, \theta_n)$*, whereas the latent indicators* $z^{\leftarrow}_{nm}$ *are distributed according to a* $Multinomial\ (1, \theta_m)$*. A graphical representation of this specific model is given in Figure 4.5.*

4.3 Strategies for Integrating Complex Data

Integration of the measurements on relations and attributes involving objects of different types can take many forms. For the purposes of this thesis, it will suffice to distinguish two types of integration, one relates to descriptive versus predictive analyses, and the other relates to the integration of labels.

4.3.1 Descriptive Analyses

In a descriptive analysis, non-observables always contribute equally to the data generation, and, in turn, observables always inform equally the inference process about non-observables. This is what happens, for example, to the multivariate relations and multivariate attributes in the previous Section; at the sampling scheme level relations and attributes of different types are assumed to be independent.

A layer of complication may be introduced. Consider a data set with N objects and, for simplicity, assume a fixed number of latent patterns, K. Consider measurements on J attributes for each on each object.[4] The mixed membership vectors in the model are object-specifc $\vec{\pi}_{1:N}$; should they be attribute specific as well? In the general formulation in Section 4.2.1 the answer is no, but it need not be so. Introducing mixed membership vectors that are specific to object-attribute pairs allows for more flexibility in the *description* of the data. However, the description inferred from the data may not be optimal when the goal of the analysis is to predict one attribute given the rest (Barnard et al., 2003).

4.3.2 Predictive Analyses

In a predictive analysis, one set of non observables always contributes to the data generation conditionally on the values assumed by a second set of observables, and, in turn, the two sets of observables inform the inference process about non-observables unequally—namely, the information the latter set contributes to the inference process is used to describe *residual variability*, which cannot be explained by information contributed by the former set of observables. This is what happens to the labels Z in Example 11.

[4] The discussion applies to a set of J relations, unchanged.

A Lengthy Example Consider observations consisting of T sets of edges, $Y_{1:T}$, among a common set of nodes, $\mathcal{N}$. The data generating process is as follows.

1. For each node $n = 1, \ldots, N$

   1.1. Sample the mixed-membership vector $\vec{\pi}_n \sim Dirichlet\ \left(\ \vec{\alpha}\ \right)$

   1.2. Sample the component indicator $\vec{z}_n \sim Multinomial\ _K\ \left(\ \vec{\pi}_n, 1\ \right)$

   1.3. Sample the latent representation $\vec{x}_n \sim \prod_{k=1}^{K} Gaussian\ _2\ \left(\ \vec{\mu}_k, \Sigma_k\ \right)^{z_{nk}}$

2. For each pair of nodes $(n, m) \in [1, N] \times [1, N]$

   2.1. Sample the value of interactions from a generalized linear model

$$y_{nm} \sim Generalized\ Linear\ Model\ \left(\ link = g^{-1}\ \right)$$

where the link function g maps the support of the average response function $\mu_{nm} = \mathbb{E}\ \left[\ y_{nm}\ \right]$ onto $\mathbb{R}$, that is, the support of the linear model η_{nm}. The linear model $\eta_{nm} = \eta_{nm}\ \left(\ \beta, \vec{x}_n, \vec{x}_m\ \right)$ involves latent, node-specific covariates, $\vec{x}_n \in \mathcal{X}$, and a global drift, β, shared by all nodes. A graphical representation of the DGP for the parametric case, using plates, is shown in Figure 4.6.

The data generating process posits that representations of nodes in a low dimensional latent space, $x^t_{1:N} \in \mathcal{X}$, are sampled independently for each graph, G_t, from a finite mixture of K Gaussians with parameters $(\vec{\mu}_{1:K}, \Sigma_{1:K})$, which encode the group centroids in the latent space $\mathcal{X}$ for al graphs[5]. At the top of the hierarchy, the mixed membership vectors, $\vec{\pi}^t_{1:N}$, are independent and identically distributed samples from a Dirichlet distribution over the K-dimensional simplex with hyper-parameter vector $\vec{\alpha}$. They provide the mixture weights. The edge weights are then generated through a "generalized linear model" that makes use of the low dimensional, latent representations

[5]For example, if we take the low dimensional space $\mathcal{X}$ to be $\mathbb{R}^2$, then each one of the K components of the mixture of Gaussians is a two-dimensional Gaussian.

of nodes, $x_{1:N}$, as covariates, along with an extra parameter β. In particular, each edge weight, y^t_{nm}, may be generated starting from the relevant pair of node representations, (x_n, x_m), through a distance model.

Following the formalism in McCullagh and Nelder (1989) we specify the generalized linear model that generates the observed edge weights y_{nm} (see step 2.1 of the data generating process) in terms of three elements.

1. The error model, $p(y_{nm})$, i.e., the model for the observed edge weights with mean $\mu_{nm} = \mathbb{E}[y_{nm}]$.

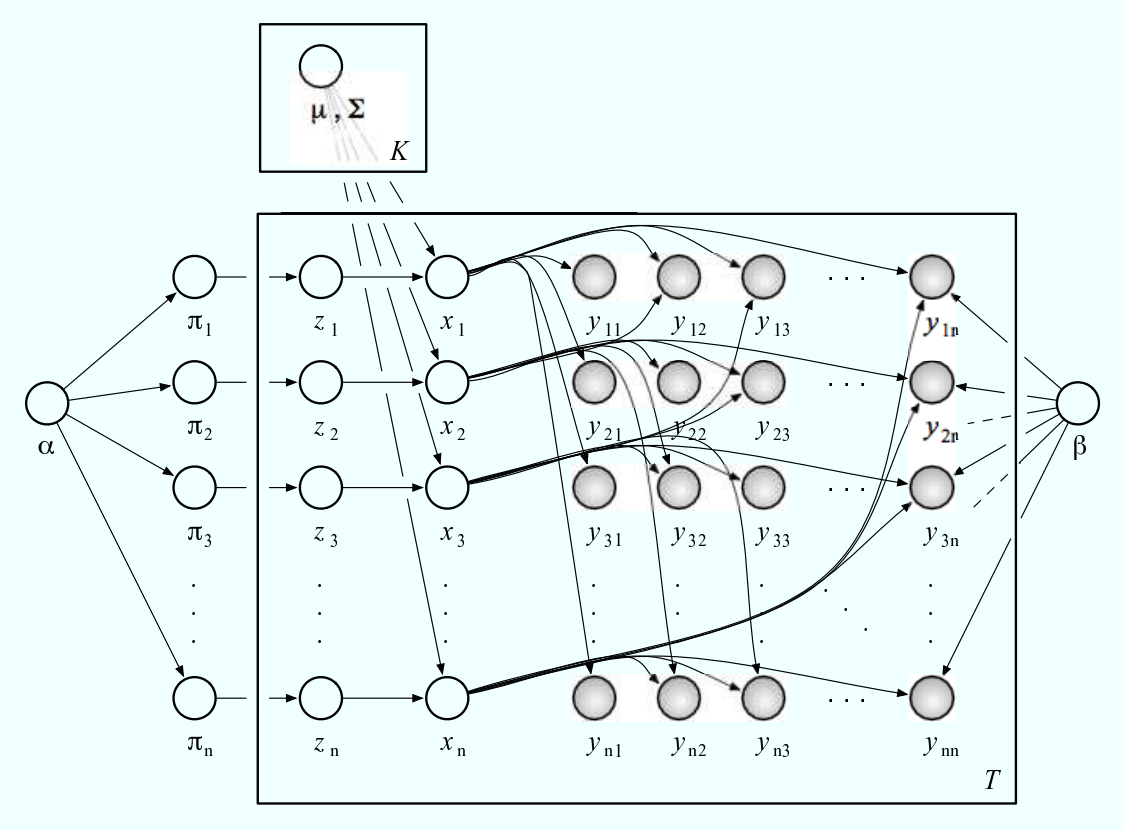

Figure 4.6: The graphical representation of the parametric model using plates, for a set of T matrices. Note that we did not draw all the arrows out of γ, for clarity, since all the interactions y_{nm} depend on it.

2. The linear model, $\eta_{nm} = \eta_{nm}(\beta, \vec{x}_n, \vec{x}_m) = \eta_{nm}\left(\beta, d(\vec{x}_n, \vec{x}_m)\right)$, for any explicit distance model d.

3. The link function[6], $g(\mu_{nm}) = \eta_{nm}$, which maps the support of μ_{nm} to that of η_{nm}—typically $\mathbb{R}$.

In particular, the linear model η_{nm} includes an explicit distance model, d, in the latent space, $\mathcal{X}$. Using the models proposed in Hoff et al. (2002),

$$\begin{aligned} \eta_{nm} &= \eta_{nm}\left(\beta, \vec{x}_n, \vec{x}_m\right) \\ &= \eta_{nm}\left(\beta, d(\vec{x}_n, \vec{x}_m)\right) \\ &= \begin{cases} \beta - |\vec{x}_n - \vec{x}_m| & \text{distance model} \\ \beta + \frac{\vec{x}_n{}^\top \vec{x}_m}{|\vec{x}_m|} & \text{projection model.} \end{cases} \end{aligned} \tag{4.11}$$

Intuitively, edges are more likely to be generated between paris of nodes whose corresponding representations in the latent space are close.

Note 3. *In a binary graph we can posit* $p(y_{nm}) = Bernoulli\ (\mu_{nm})$*, where* $\mu_{nm} \in [0, 1]$ *for all node pairs* $(n, m) \in \mathcal{N}$*. The linear model is* $\eta_{nm} = \beta + d(\vec{x}_n, \vec{x}_m)$*. The link function is* $g(\mu_{nm}) = \log\left(\frac{\mu_{nm}}{1-\mu_{nm}}\right)$*, and its inverse is* $\mu_{nm} = \frac{1}{1+\exp(-\eta_{nm})}$*. In a graph with non-negative, integer edge weights we can posit* $p(y_{nm}) = Poisson\ (\mu_{nm})$*, where* $\mu_{nm} \in \mathbb{R}_+$ *for all node pairs* $(n, m) \in \mathcal{N}$*. The linear model is* $\eta_{nm} = \left(\beta, d(\vec{x}_n, \vec{x}_m)\right)$*, as in Equation 4.11. The link function is* $g(\mu_{nm}) = \log(\mu_{nm})$*, and its inverse is* $\mu_{nm} = e^{\eta_{nm}}$*.*

This model follows closely the models in Hoff et al. (2002) and Handcock et al. (2007), with the novelty that it depends on the set of mixed membership vectors $\vec{\pi}_{1:N}$. In a sense, it is predictive because a model for the joint probability of latent variables and data is missing, $p(Y, \vec{\pi}_{1:N}, \vec{x}_{1:N})$. However, in order to make it predictive in the sense intended here a little more work is needed.

[6] Here I consider *canonical* link functions (McCullagh and Nelder, 1989).

First, we need to introduce a second source of information, for example, multivariate attributes on the nodes, $U_{1:N}$, where the quantity $u_n(m)$ encodes the value of the m-th attribute measured on the n-th node. Then we posit a data generating process for the attributes, e.g., along the lines of the models in Section 4.1. Finally, and here is where the *predictive* is used in the sense I intend, we need to make a decision about how to link the two models; for interactions and attributes. If the goal of the analysis is that of predicting, or de-noising, interactions from attributes (Airoldi et al., 2006c) then we want to condition the interactions on the attributes in the generating process. There are several ways of doing this; a possibility is that of generating node-specific mixture component indicator $\vec{z}_n$, in the model for the interactions Y, from the node-specific mixture component indicators $\{\vec{z}_n^m : m = 1, \ldots, M\}$ already samples for the attribute—see Barnard et al. (2003) for another example.

Going back to the multivariate models of attributes and relations of Section 4.2, I need to specify a generative link between the attributes and/or relations at the *sampling scheme level*; univariate measurements are no longer independent. Figure 4.7 show the relevant portion of the graphical model structure that is common to models of both multivariate attributes and relations. By positing structural dependencies in the model predictive analyses can be supported; that is, latent patterns associated with a set of measurements will be inferred that are useful in predicting a different set of measurements.

Modeling Text and References I conclude this chapter with an application of data integration in a larger context: models of data integration are instrumental to resolve a substantive issue about model choice.

Example 25 (Proceedings of the National Academy of Sciences, PNAS)**.** *Erosheva et al. (2004) and Griffiths and Steyvers (2004) report on their estimates about the number of latent topics, and find evidence that supports a small number of topics (e.g., as few as 8 but perhaps a few dozen)*

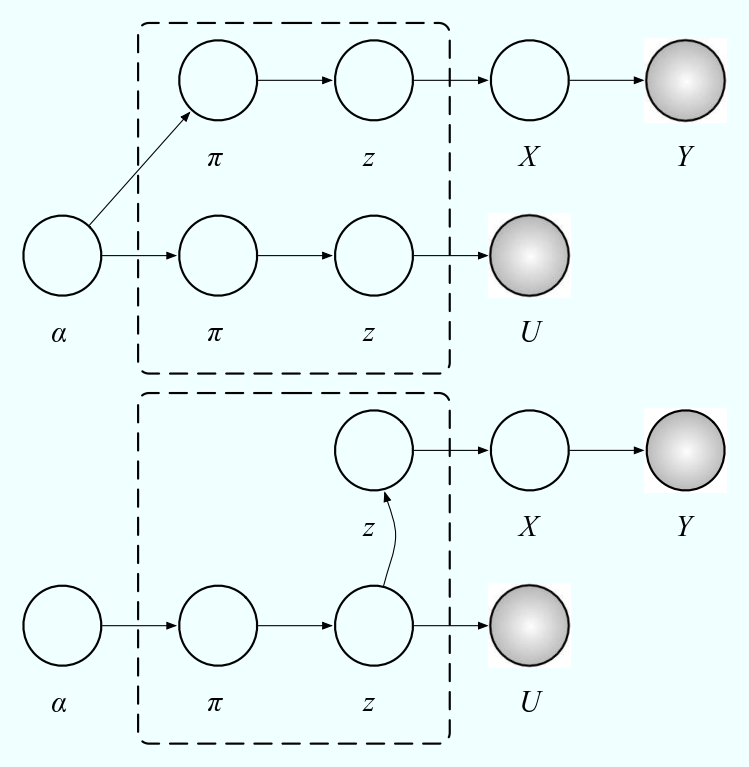

Figure 4.7: The structural dependencies among the (latent) variables in the dashed box distinguish the type of analysis. The absence of dependencies (top panel) leads to models that support descriptive analyses, whereas the presence of dependencies (bottom panel) leads to models that support predictive analyses.

or as many as 300 latent topics, respectively. There are a number of differences between the two analyses: the collections of papers were only partially overlapping (both in time coverage and in subject matter), the authors structured their dictionary of words differently, one model could be thought of as a special case of the other but the fitting and inference approaches had some distinct and non-overlapping features. The most remarkable and surprising difference come in the estimates for the numbers of latent topics: Erosheva et al. focus on values like 8 and 10 but admit that a careful study would likely produce somewhat higher values, while Griffiths & Steyvers present analyses they claim support on the order of 300 topics! Should we want or believe that there are only a dozen or so topics capturing the breadth of papers in PNAS or is the number of topics so large that almost every paper can have its own topic? A touchstone comes from the

journal itself. PNAS, in its information for authors (updated as recently as June 2002), states that it classifies publications in biological sciences according to 19 topics. When submitting manuscripts to PNAS, authors select a major and a minor category from a predefined list list of 19 biological science topics (and possibly those from the physical and/or social sciences).

Below, I summarize an alternative set of analyses (Airoldi et al., 2006e) using the version of the PNAS data on biological science papers analyzed in (Erosheva et al., 2004). Said analyses employ both parametric and non-parametric strategies for model choice, and make use of both text and references of the papers in the collection, in order to resolve this issue. This case study gives us a basis to discuss and assess the merit of the various strategies. In the process I explore how to perform the model selection for Bayesian models of mixed-membership. After choosing an *optimal* value for the number of topics, K^*, and its associated words and references usage patterns, I also examine the extent to which they correlate with the *actual* topic categories specified by the authors.

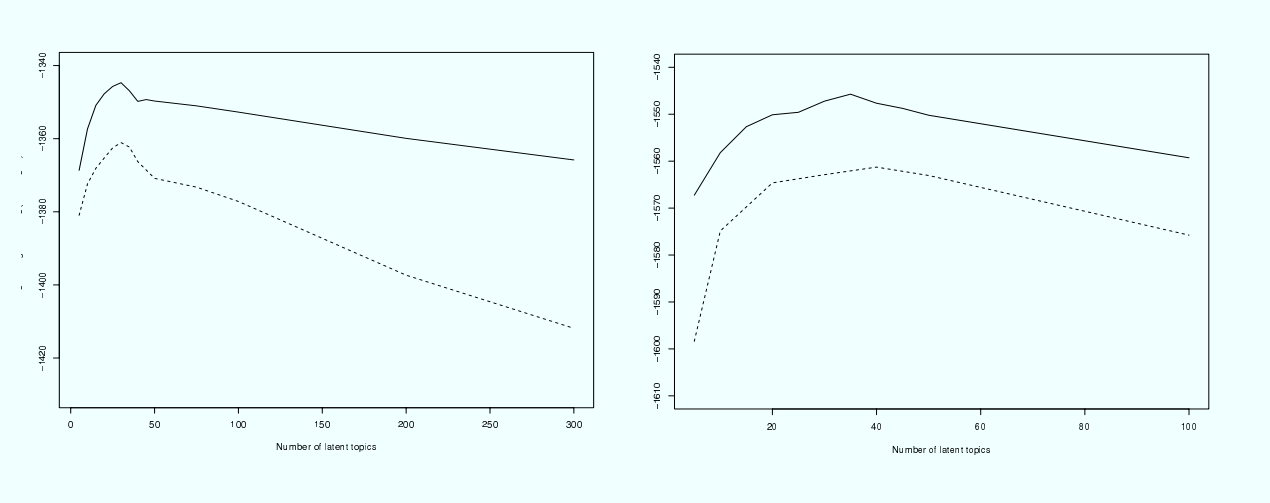

Figure 4.8: Left Panel: Log-likelihood (5 fold cv) for $K = 5, \ldots, 50, 75, 100, 200, 300$ topics. We plot: text only, α fitted (solid line); text only, α fixed (dashed line). Right Panel: Log-likelihood (5 fold cv) for $K = 5, \ldots, 50, 100$ topics. We plot: text and references, α fitted (solid line); text and references, α fixed (dotted line).

Six Bayesian mixed membership models were fitted to infer the topics underlying the PNAS

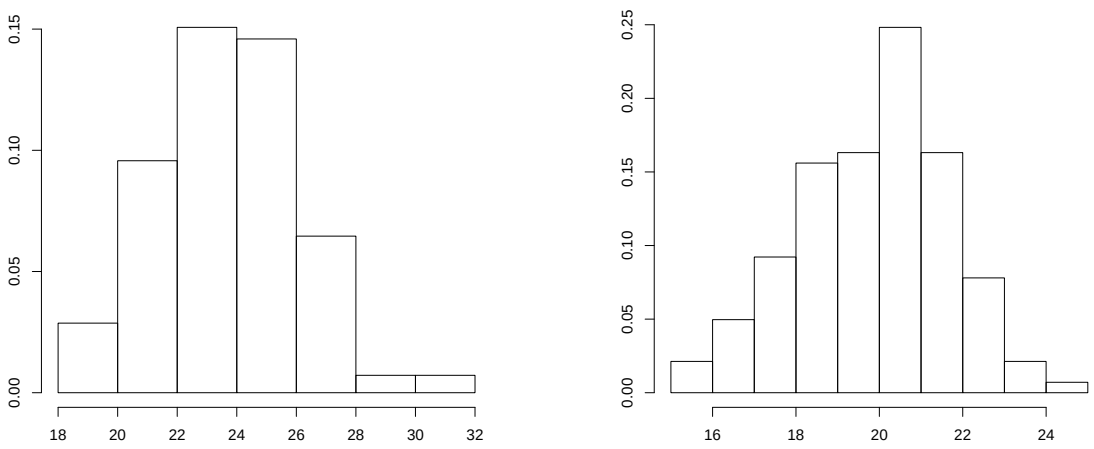

Figure 4.9: Posterior distribution of K for the PNAS scientific collection corresponding to the infinite mixture models of text (left panel) and of text and references (right panel).

dataset: words alone or both words and references were modeled with parametric and semi-parametric mixed model specifications, and for fully parametric specifications the Dirichlet hyper-parameter α was either fitted using an empirical Bayes strategy or fixed with an ad-hoc strategy inspired by the one used in the analysis of PNAS data by Griffiths and Steyvers (2004). Full details about model specifications and posterior inference algorithms, using both variational methods and MCMC, are given in Airoldi et al. (2006e). See the right panel of Figure 4.3 for a graphical representation of the models of text and references.

The plots of the log likelihood in Figure 4.8 suggest a number of topics between 20 and 40 whether words or words and references are used. The semi-parametric model generates a posterior distribution for the number of topics, K, given the data. Figure 4.9 shows the posterior distribution ranges from 23 to 33 profiles. We can expect that the semi-parametric model will require more topics than the parametric model, since it leads to a hard clustering of documents—into topics. By choosing $K = 20$ topics, a meaningfully interpretation all of the word and reference usage patterns can be found. A parametric model with 20 topics was fitted to the data, both words and references,

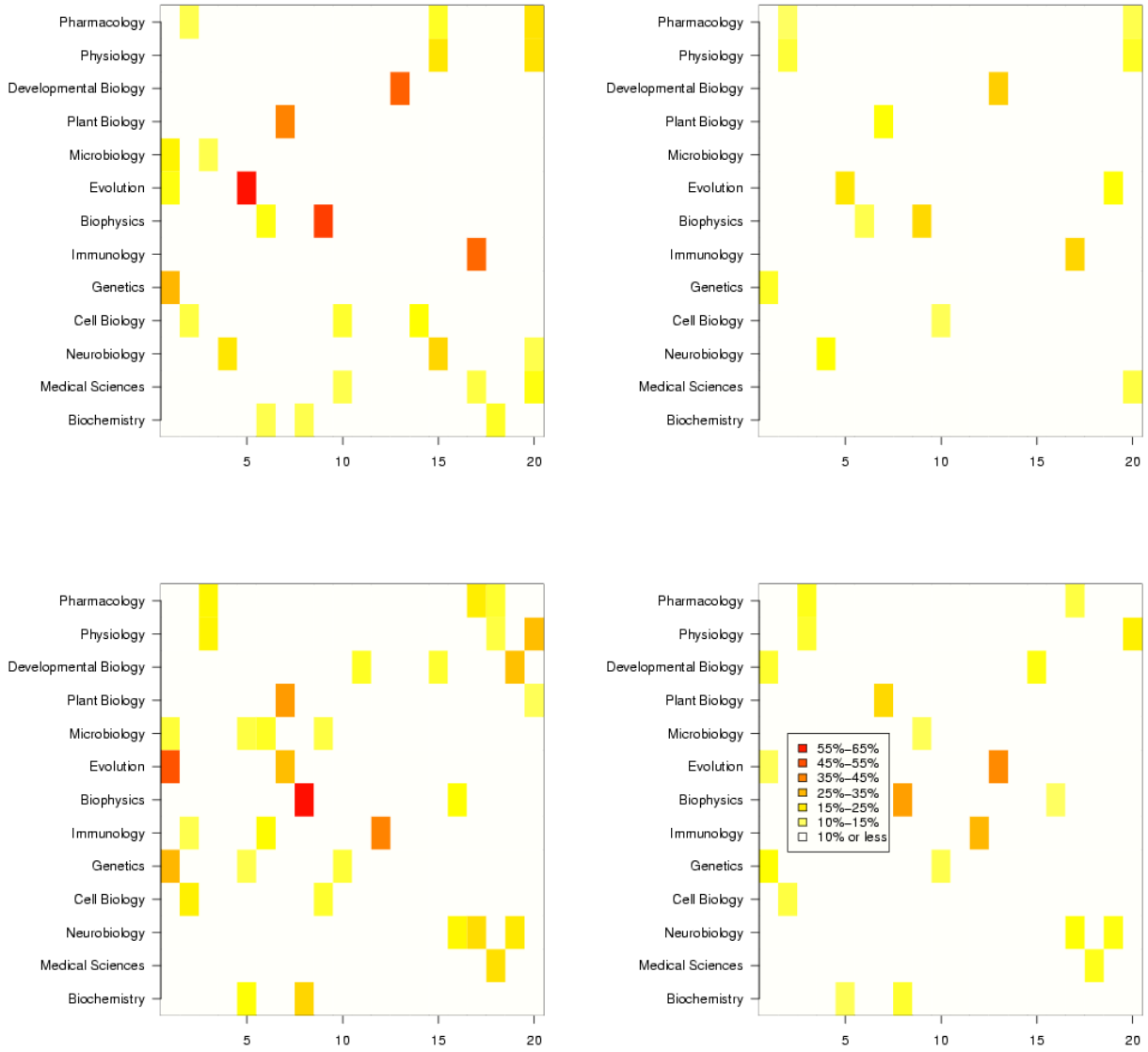

Figure 4.10: The average membership in the 20 latent topics (columns) for articles in thirteen of the PNAS editorial categories (rows). Darker shading indicates higher membership of articles submitted to a specific PNAS editorial category in the given latent topic and white space indicates average membership of less than 10%. Note that the rows sum to 100% and therefore darker topics show concentration of membership and imply sparser membership in the remaining topics. These 20 latent topics were created using the four finite mixture models with words only (1^{st}, 2^{nd}) or words and references (3^{rd}, 4^{th}) and α estimated (1^{st}, 3^{rd}) or fixed (2^{nd}, 4^{th}).

Table 4.3: Word usage patterns corresponding to the model of text & references, with $K = 20$ topics.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
gene	kinase	cells	cortex	species
genes	activation	virus	brain	evolution
sequence	receptor	gene	visual	population
chromosome	protein	expression	neurons	populations
analysis	signaling	human	memory	genetic
genome	alpha	viral	activity	selection
sequences	phosphorylation	infection	cortical	data
expression	beta	cell	learning	different
human	activated	infected	functional	evolutionary
dna	tyrosine	vector	retinal	number
number	activity	protein	response	variation
identified	signal	vectors	results	phylogenetic
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
enzyeme	plants	protein	protein	cells
reaction	plant	rna	model	cell
ph	acid	proteins	folding	tumor
activity	gene	yeast	state	apoptosis
site	expression	mrna	energy	cancer
transfer	arabidopsis	activity	time	p53
mu	activity	trna	structure	growth
state	levels	translation	single	human
rate	cox	vitro	molecules	tumors
active	mutant	splicing	fluorescence	death
oxygen	light	complex	force	induced
electron	biosynthesis	gene	cdata	expression
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
transcription	dna	cells	protein	ca2+
gene	rna	cell	membrane	channel
expression	repair	expression	proteins	channels
promoter	strand	development	atp	receptor
binding	base	expressed	complex	alpha
beta	polymerase	gene	binding	cells
transcriptional	recombination	differentiation	cell	neurons
factor	replication	growth	actin	receptors
protein	single	embryonic	beta	synaptic
dna	site	genes	transport	calcium
genes	stranded	drosophila	cells	release
activation	cdata	embryos	nuclear	cell
Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
peptide	cells	domain	mice	beta
binding	cell	protein	type	levels
peptides	il	binding	wild	increased
protein	hiv	terminal	mutant	insulin
amino	antigen	structure	gene	receptor
site	immune	proteins	deficient	expression
acid	specific	domains	alpha	induced
proteins	gamma	residues	normal	mice
affinity	cd4	amino	mutation	rats
specific	class	beta	mutations	treatment
activity	mice	sequence	mouse	brain
active	response	region	transgenic	effects

to focus on the interpretation of the 20 topics. Table 4.3, lists 12 high-probability words from the estimated 20 topics after filtering out the stop words. Table 4.4 shows the 5 references with the highest probability for 6 of the topics.

Using both tables, here is a possible interpretation of the 20 topics:

- Topics 1 and 12 focus on nuclear activity (genetic) and (repair/replication).
- Topic 2 concerns protein regulation and signal transduction.

Table 4.4: References usage patterns for 6 of the 20 topics corresponding to the model of text & references, with $K = 20$ topics.

Author	**Journal**
Topic 2	
THOMPSON,CB	SCIENCE, 1995
XIA,ZG	SCIENCE, 1995
DARNELL,JE	SCIENCE, 1994
ZOU,H	CELL, 1997
MUZIO,M	CELL, 1996
Topic 5	
SAMBROOK,J	MOL. CLONING. LAB. MANU., 1989
ALTSCHUL,SF	J. MOL. BIOL., 1990
EISEN,MB	P. NATL. ACAD. SCI. USA, 1998
ALTSCHUL,SF	NUCLEIC. ACIDS. RES, 1997
THOMPSON,JD	NUCLEIC. ACIDS. RES, 1994
Topic 7	
SAMBROOK,J	MOL. CLONING. LAB. MANU,1989
THOMPSON,JD	NUCLEIC. ACIDS. RES,1994
ALTSCHUL,SF	J. MOL. BIOL,1990
SAITOU,N	MOL. BIOL. EVOL,1987
ALTSCHUL,SF	NUCLEIC. ACIDS. RES,1997
Topic 8	
SAMBROOK,J	MOL. CLONING. LAB. MANU,1989
KIM,NW	SCIENCE, 1994
BODNAR,AG	SCIENCE, 1998
BRADFORD,MM	ANAL. BIOCHEM., 1976
FISCHER,U	CELL, 1995
Topic 17	
SHERRINGTON,R	NATURE,1995
HO,DD	NATURE,1995
SCHEUNER,D	NAT. MED.,1996
THINAKARAN,G	NEURON,1996
WEI,X	NATURE,1995
Topic 20	
CHOMCZYNSKI,P	ANAL. BIOCHEM., 1987
BRADFORD,MM	ANAL. BIOCHEM., 1976
KUIPER,GGJM	P. NATL. ACAD. SCI. USA, 1996
MONCADA,S	PHARMACOLREV, 1991
KUIPER,GG	ENDOCRINOLOGY, 1998

- Two topics are associated with the study of HIV and immune responses: topic 3 is related to virus treatment and topic 17 concerns HIV progression.

- Two topics relate to the study of the brain and neurons: topic 4 (behavioral) and topic 15

(electrical excitability of neuronal membranes).

- Topic 5 is about population genetics and phylogenetics.
- Topic 7 is related to plant biology.
- Two topics deal with human medicine: topic 10 with cancer and topic 20 with diabetes and heart disease.
- Topic 13 relates to developmental biology.
- Topic 14 concerns cell biology.
- Topic 19 focus on experiments on transgenic or inbred mutant mice.
- Several topics are related to protein studies, e.g., topic 9 (protein structure and folding), topic 11 (protein regulation by transcription binding factors), and topic 18 (protein conservation comparisons).
- Topics 6, 8, and 16 relate to biochemistry.

These labels for the topics are primarily convenience, but they do highlight some of the overlap between the PNAS sections (Plant Biology and Developmental Biology) and the latent topics (7 and 13). However, many plant biologists may do molecular biology in their current work. By examining the topics ones can see that small sections such as Anthropology do not emerge as topics and broad sections such as Medical Science and Biochemistry have distinct subtopics within them. This also suggests special treatment for general sections such as Applied Biology and cutting-edge interdisciplinary papers when evaluating the classification effectiveness of a model.

To summarize the distribution of latent aspects over distributions, a graphical representations of the distribution of latent topics for each of the PNAS topics is provided in Figure 4.10. The third figure represents the model used for Tables 4.3 and 4.4. The two figures on the right represent

models where the α parameter of the Dirichlet prior over topics is fixed. These two models are less sparse than the corresponding models with α fit to the data. For twenty latent topics, the hyper-parameter α was fixed at $50/20 = 2.5 > 1$ and this means each latent topic is expected to be present in each document and a priori we expect equal membership in each topic. By contrast the fitted values of α are less than one lead to models that expect articles to have high membership in a small number of topics. The PNAS topics tend to have a few latent topics highly represented when α is fit and low to moderate representation in all topics when α is fixed (as seen by white/light colored rows). For additional discussion of further consequences of these assumptions see the simulation at the end of Section 6.2.2.

Further examining Figure 4.10, note that topic 1, identified with genetic activity in the nucleus, was highly represented in articles from Genetics, Evolution, and Microbiology. Also note that nearly all of the PNAS classifications are represented by several word and reference usage patterns in all of the models. This highlights the distinction between the PNAS topics and the discovered latent topics. The assigned topics used in PNAS follow the structure of the historical development of Biological Sciences and the divisions/departmental structures of many medical schools and universities. The latent topics, however, show the greater ideas of interest within the field. Topic 9, which concerns the structure and topology of proteins, is highly represented in theoretical papers in Evolution, Genetics, Cell and Developmental Biology as well as in applied papers in Ecology, Pharmacology, and Applied Biological Sciences. These latent topics, however, are structured around the current interest of Biological Sciences. Figure 4.10 also shows that there is a lot of hope for collaboration and interest between separate fields which are researching the same ideas.

The held-out log likelihood plot corresponding to five-fold cross validation in Figure 4.8 suggests a number between 20 and 40 topics for the parametric model. Further analyses with parametric mixed membership models of words and references supports support values towards the

lower end of this range, i.e., $K = 20$, more than other choices. This is also true in the posterior distribution of K for the semi-parametric mixed membership model. To conclude, the hyper-parameter α was fixed to $50/K$, following the choice in Griffiths and Steyvers (2004), as well as estimated using empirical Bayes. Both sets of analyses produced a similar conclusion. While Griffiths and Steyvers (2004) found posterior evidence for nearly 300 topics, a number on the order of 20 or 30 provides a far better fit to the data, assessed robustly by multiple criteria and model specifications that integrate different types of data. Moreover, a lower number of latent topics appears to be simpler and more interpretable in a meaningful way; this is not possible with 300 topics.

* * *

The generative models for attributes presented here differ much from published alternatives (e.g., Pritchard et al., 2000; Blei et al., 2003) in terms of the way the data inform the allocation of objects to patterns. For example, they are models of counts that lead to variability in the totals, and such variability has influence on the allocation—see Section 6.2.1 for a discussion. I derived a multivariate characterization of both models of attributes, in Section 4.1, and relations, in Section 3.1. Fast posterior inference is available for the general formulations as well (Airoldi et al., 2006d,e). I described alternative strategies for integrating multiple sources of data in such models, depending on whether the goals of the analysis is descriptive or predictive.

Integrating heterogeneous data types under a unified model is a challenge to the analysis of *complex data*, which are simultaneously described by intrinsically different types of characteristics, such as features in attribute space and links in relational space. This chapter suggests that Bayesian models of mixed membership provide us with a solution to modeling and algorithmic issues that arise in (what I term) *integrated-learning* problems that involve *complex data* by combining modules specific to multivariate attribute and relations within a hierarchical framework.

Research along this line is still very limited, especially work based on well-founded statistical principles. My methodology supports robust inter-modal inference, latent mechanism discovery, and information retrieval. The strategies for integrating complex data presented here enable modular and distributable engineering solutions for organization and prediction problems. Feasible engineering approaches to such problems in essence require realistic statistical models, accompanied by scalable computational methods.

Chapter 5

Dynamics and Evolution

In this chapter I describe how the models introduced in previous chapters can be extended to take temporal evolution into account. To this extent, several models of dynamic behavior are present in the classical statistical literature, which can be used to model the evolution of latent patterns for a finite number of epochs, T. The basic idea is to chose which sets of variables evolves over time, e.g., $\Theta^{1:T}$, and posit a model for the transition, e.g., $P_{\mathcal{A}}(\Theta^t|\Theta^{t-1})$.

For instance, recall the state-space model of Example 10, which extends the factor analysis model of Example 7 by evolving the latent factors from one epoch to the next. The data generating process for the observations $X^{(1:T)}$ is as follows,

1. At epoch $t = 0$
    1.1. For each object $n \in \mathcal{N}_1$
        1.1.1. Sample the latent factors $\vec{\phi}_n \sim Normal\ _K\ (0, I)$
        1.1.2. Sample the error $\vec{\epsilon}_n^{(0)} \sim Normal\ _M\ (0, \Psi)$
        1.1.3. Define the multivariate attribute $\vec{x}_n^{(0)} = \Lambda\vec{\phi}_n + \vec{\epsilon}_n^{(0)}$,
2. At epoch $0 < t < T$

2.1. For each object $n \in \mathcal{N}_1$

2.2.1. Evolve the latent factors $\vec{\phi}_n^{(t)} = F\vec{\phi}_n^{(t-1)}$,

2.2.2. Sample the error $\vec{\epsilon}_n^{(t)} \sim Normal_{\ M}\ (0, \Psi)$

2.2.3. Define the multivariate attribute $\vec{x}_n^{(t)} = \Lambda\vec{\phi}_n^{(t)} + \vec{\epsilon}_n^{(t)}$,

where F is a $(K \times K)$ matrix that encodes the dynamics of the latent factors. As before, the algorithm suggests a hierarchical decomposition of the joint probability distribution of the attributes, $X^{(1:T)} = \vec{x}_{1:N}^{(1:T)}$, and the latent factors, $\Theta^{(1:T)} = (\vec{\phi}_{1:N}^{(1:T)}, \vec{\epsilon}_{1:N}^{(1:T)})$, given a set of underlying constants, $\mathcal{A} = (F, \Lambda, \Psi)$ that does not change over time.[1] The likelihood is then,

$$\begin{aligned} \ell(X^{(1:T)}|\mathcal{A}) &= \int P_1(\Theta^{(0)}|\mathcal{A})\ P_2(X^{(0)}|\Theta^{(0)}, \mathcal{A}) \times \\ &\times \left(\prod_{t=1}^{T} P_0(\Theta^{(t)}|\Theta^{(t-1)}, \mathcal{A})\ P_2(X^{(t)}|\Theta^{(t)}, \mathcal{A}) \right) d\Theta^{(1:T)}, \end{aligned} \tag{5.1}$$

where P_1 and P_2 are K- and M-dimensional Gaussian densities, respectively, and P_0 is the deterministic transformation in Step 2.2.1. of the data generating process. A graphical representation of FA and SSM is given in Figure 5.1, which highlights the simple connection between the two models.

Model specifications vary depending on the applications, e.g., ARIMA, possibly multivariate, linear versus non-linear, deterministic versus stochastic transition, Gaussian versus non-Gaussian errors, or Markovian versus complex (Wasserman, 1980; Rabiner, 1989; Brockwell and Davis, 1991; Karr, 1991; West and Harrison, 1997; Doucet et al., 2001).

Example 26. *The admixture of latent blocks model of Section 3.1 is a model for a network Y^t. Denote by $P(Y^t|\vec{\alpha}, B)$ the model for the network at time t, given the hyper-parameter $\vec{\alpha}$, which governs the distribution of the mixed membership vectors $\vec{\pi}_{1:N}$, and the stochastic block model B.*

[1] The dynamic matrix F may be easily modeled as time dependent and/or stochastic, as the problem requires (Airoldi and Faloutsos, 2004; Airoldi et al., 2005d).

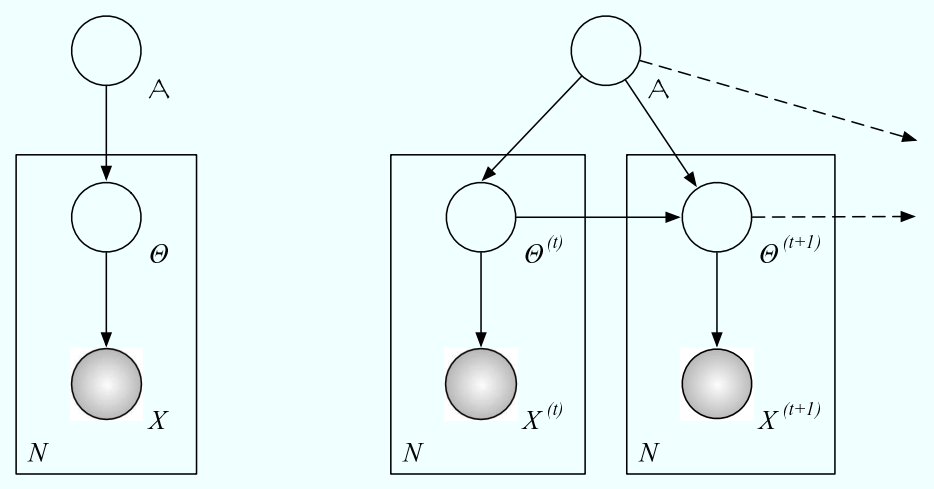

Figure 5.1: Graphical representations of a factor analysis model (left) and of a state-space model for observations at two consecutive epochs (right). White nodes denote non-observables, whereas shadowed nodes denote observables.

The model can be extended to account for time by evolving the hyper-parameter α as follows.

1. *At epoch $0 < t < T$*

   1.1. *Sample the error $\vec{\epsilon}^{(t)} \sim Normal_K\,(0, \sigma^2 I)$*

   1.2. *Evolve the latent position $\vec{\alpha}^{(t)} = I\,\vec{\alpha}^{(t-1)} + \exp\{\vec{\epsilon}^{(t)}\}$,*

This extra step specifies a linear transition model, $P_\sigma(\vec{\alpha}^t|\vec{\alpha}^{t-1})$. In the same spirit it is possible to evolve the stochastic block model B.

Example 27. *The latent space model of mixed membership introduced in Section 4.3.2 is a model for a set of networks $Y^{1:T}$. Denote by $P(Y^t|\vec{\mu}_{1:K}, \Sigma_{1:K})$ the model for the network at time t, given a parametric description of the clusters in the latent space. This model can be extended to account for time, at epoch $0 < t < T$, by evolving the latent cluster positions as follows.*

1. *For each cluster $k = 1, \ldots, K$*

   1.1. *Sample the error $\vec{\epsilon}_k^{(t)} \sim Normal_2\,(0, \Psi)$*

1.2. Evolve the latent position $\vec{\mu}_k^{(t)} = F\ \vec{\mu}_k^{(t-1)} + \vec{\epsilon}_k^{(t)}$,

This extra step in the process specifies a set of K *independent linear transition models,* $P_{\mathcal{A}_k}(\vec{\mu}_k^t \mid \vec{\mu}_k^{t-1})$, *where the parameter sets* $\mathcal{A}_k = \Psi$, *for all* k.

Alternatively, it is possible to specify temporal patterns directly as a part of (Ξ, Θ). Such a modeling strategy allows to consider longitudinal sequences of observations about objects as admixtures of complicated patterns, specified in a parametric or non-parametric fashion, and avoids technical issues that arise when considering the specification of an explicit model of evolution.

5.1 Dynamic Network Tomography

The models discussed above resolve the observed sequence of graphs into simple patterns, which evolve over time with some regularity. Independently of how such patterns are specified, their description is *parsimonious*. In some problems, however, we need to solve the opposite problem; namely, that of resolving the observed sequence of graphs into patterns with an order of complexity higher than that of the observations. In other words, for latent patterns $\Theta \in \mathcal{T}$ and observations $Y \in \mathcal{Y}$, in the models considered so far the dimensionality of $\mathcal{T}$ was lower than that of $\mathcal{Y}$. This is no longer true in the models presented in this section, where the dimensionality of the space $\mathcal{T}$ is higher than that of $\mathcal{Y}$. Problems of this sort, where the solution space is orders of magnitude larger than the space spanned by the data and the constraints, are referred to as *inverse problems* in the literature (Hansen, 1998).

The distinction above is not evident from the graphical representation of the models. The issues are deeper: (i) identifying the space of solutions is often not trivial; (ii) regularization conditions are needed to induce a well-behaved optimization problem. The driving application here is *network tomography*, where the origin-destination (OD) traffic flows need be estimated, e.g., who is com-

municating with whom in a local area network. The direct measurement of the OD traffic is usually difficult and typically unfeasible; instead, the loads on every link can be easily measured, that is, sums of desired OD flows. In a network with N nodes, the problem is then to recover $O(N^2)$ OD flows from $O(N)$ sums. Such problem has been studied by many in the statistical literature (Vanderbei and Iannone, 1994; Vardi, 1996; Tebaldi and West, 1998; Cao et al., 2000; Zhang et al., 2003; Airoldi and Faloutsos, 2004). The model proposed here starts from the Bayesian analysis of Tebaldi and West (1998), and extends it to a dynamic context by: (i) introducing explicit time dependence among the traffic flows; (ii) positing a stochastic multiplicative process for the dynamic; and (iii) positing realistic, non-Gaussian marginals for the traffic flows. The findings echos those of Tebaldi and West (1998) with regard to the need for informative priors in order to mitigate the bias in the estimated traffic flows due to the presence of multiple peaks in the likelihood, and to the presence of ridges in between those peaks, e.g., see Figure 5.8. The solution presented here scales linearly with new observations and is more accurate then alternative solutions, on real network traffic measured at Carnegie Mellon and at AT&T.

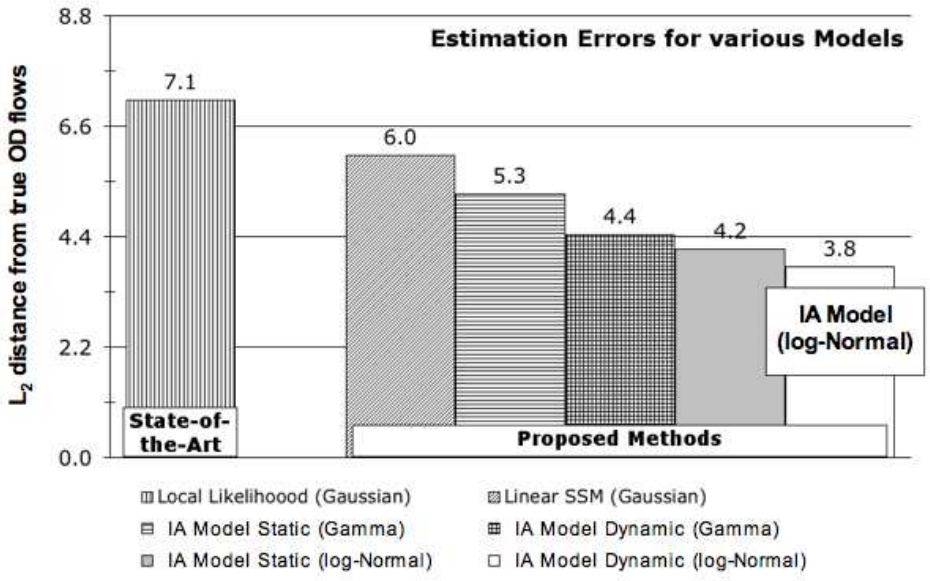

Figure 5.2: Estimation error in ℓ_2 distance.

Table 5.1: Summary of symbols.

Symbol	Description
T	Number of time points.
ℓ	Number of observable link loads.
κ	Number of non-observable OD flows.
$\boldsymbol{A}$	$(\ell \times \kappa)$ fixed routing matrix.
$\boldsymbol{Y}$	$(\ell \times T)$ matrix of link loads.
$\boldsymbol{X}$	$(\kappa \times T)$ matrix of OD traffic flows.
$\boldsymbol{\Lambda}$	$(\kappa \times T)$ matrix of means of $\boldsymbol{X}\|\boldsymbol{\Lambda}, \boldsymbol{\phi}$.
$\boldsymbol{\phi}$	$(1 \times T)$ vector of scale factors for $\boldsymbol{X}\|\boldsymbol{\Lambda}, \boldsymbol{\phi}$.
$\boldsymbol{\theta}$	Generic vector of hyper-parameters.
$\pi(\boldsymbol{\theta})$	Generic prior distribution.

5.1.1 Goals of the Analysis

Knowledge about the origin-destination (OD) traffic matrix allows network engineers and managers to solve problems in design, routing, configuration debugging, monitoring and pricing; in fact the OD traffic matrix provides valuable information about who is communicating with whom in a network, at any given time. Unfortunately the direct measurement of the OD traffic is usually difficult, or even infeasible, in real networks. The direction of current research is to develop methods to infer the OD traffic flows from observed traffic loads on the links of the network, however the methods that have been proposed so far seem not to fully take advantage of two of the main empirically observed features of network traffic; namely its very skewed marginal distribution, and its time dependent nature.

I introduce the *inverse allocation* model (IA henceforth) which improves the models present in the literature by introducing two realistic assumptions: (i) the log-Normal distribution provides a realistic model for the marginal OD traffic flows, (ii) time dependence between successive flows on a same OD route narrows the variability of the estimates. A two-stage estimation procedure is proposed to estimate parameters of the IA model.

The Problem and its Facets In a formulation of the problem we want to solve there are several time series which we would like to estimate, but which we cannot observe, say, a vector of traffic flows $\boldsymbol{x}(t)$ over times $t = 1, ..., T$. However, we are able to observe linear combinations of these traffic flows, the vector of link loads $\boldsymbol{y}(t)$ over times $t = 1, ..., T$, and we know which components of $\boldsymbol{x}(t)$ mix into each of the components $y(i, t)$ at each time t trough the routing matrix $\boldsymbol{A}$, that does not change over time There are two modeling aspects to this problem.

Problem 1 (Inverse Problem)**.** *Given the matrix of link loads $\boldsymbol{Y}_{(\ell \times T)}$ and a routing matrix $\boldsymbol{A}_{(\ell \times \kappa)}$, we want to find the matrix of non-observable OD traffic flows $\boldsymbol{X}_{(\kappa \times T)}$ such that $\boldsymbol{Y} = \boldsymbol{A} \cdot \boldsymbol{X}$. Always $\kappa > \ell$.*

Example 28. *The linear equations that correspond to the routing scheme of the star network in Figure 5.3 below are:*

$$\begin{bmatrix} y(1,t) \\ y(2,t) \\ y(3,t) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x(1,t) \\ x(2,t) \\ x(3,t) \\ x(4,t) \end{bmatrix} \tag{5.2}$$

$y(1, t)$ measures the traffic load on the link from node 1 to the router and captures both the OD flow from node 1 to node 2, $x(2, t)$, and the OD flow from node 1 to itself, $x(1, t)$. $y(3, t)$ measures the

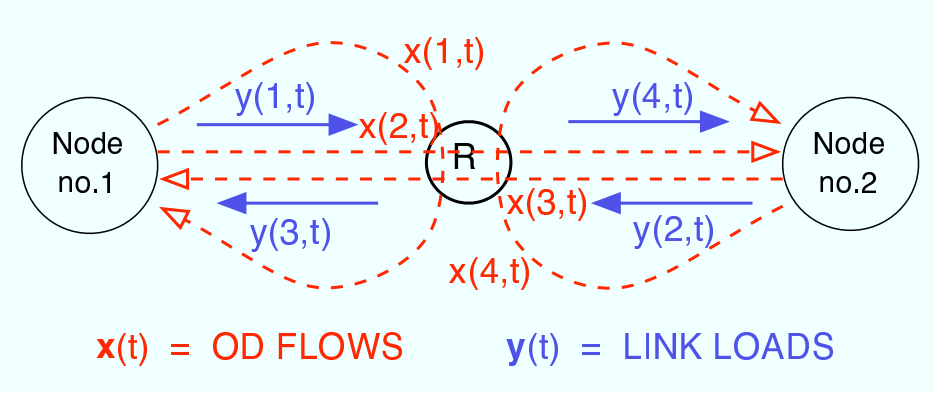

Figure 5.3: Two subnetworks connect to a router. We observe the link loads (solid blue arrows), and want to infer the hidden traffic flows (dashed red arrows).

traffic load on the link from the router to node 1 and captures both the OD flow from node 2 to node 1, $x(3,t)$, and the OD flow from node 1 to itself, $x(1,t)$. We want to estimate four (κ) unobservable quantities starting from three (ℓ) independent observations[2]*. The system is under-specified, $\kappa > \ell$, hence some extra information is needed in order to identify one single solution.*

Problem 2 (Regularization)**.** *Impose a set of additional constraints (or a penalty term) on $\boldsymbol{X}_{(\kappa \times T)}$ in order to induce smoothness on the space of solutions of the inverse problem.*

The likelihood of the data entailed by a statistical model provides us with a natural criterion to discern *likely* solutions from unreasonable ones. Following this idea we model the unobservable quantities $\boldsymbol{x}(t)$ with a joint probability distribution; this induces a probabilistic mapping on the space of the observations $\boldsymbol{y}(t)$ via equation 5.2, so that we can compute the likelihood of the observations, and look for traffic flows that maximize the probability of particular data observations. Unfortunately in time-independent models the likelihood of $\boldsymbol{y}(t)$ is not necessarily unimodal, even as we assume independent components in $\boldsymbol{x}(t)$, and even as we use well-behaved functional forms for their distributions. More information is needed to identify a solution. At this point there are two main ways to introduce the extra information we need. In a purely data-driven approach we would augment the data in some way, whereas in a knowledge-driven approach we would make use of informative priors in a Bayesian setting, with the complication in this latter case of defining what we mean by "informative". Data augmentation can be realized, for example, by raising the likelihood of the data to a power, as in simulated annealing, or by borrowing observations from epochs close in time to the current one to obtain a smoothed average solution. Alternatively, we can build "informative" priors based on partial knowledge about the magnitude of the OD flows, and update using Bayes rule and a "more accurate" data model.

The two-stage estimation procedure for the IA model is suggestive of a nonparametric empirical Bayes learning strategy, where the observations are used to first calibrate informative priors, and

[2] We assume that routers neither generate nor absorb traffic.

then to filter the posterior distributions of the OD flows given the data. The proposed solution: (i) uses realistic models for the OD flows; (ii) takes advantage of the time dependence of the data while using the whole history of observations $\{\boldsymbol{y}(1), ..., \boldsymbol{y}(t)\}$ to estimate $\boldsymbol{x}(t)$ in a proper Bayesian fashion.

5.1.2 Model Specifications

Previous models assume independent OD flows across different epochs. Here I introduce models based on dynamical systems, which naturally extend previous approaches by assuming time dependence *explicitly* (Brockwell and Davis, 1991; West and Harrison, 1997; Doucet et al., 2001).

Definition 1. *A linear Gaussian state-space model is defined by the following set of equations,*

$$\begin{cases} \boldsymbol{x}(t) &= \boldsymbol{F} \cdot \boldsymbol{x}(t-1) + \boldsymbol{e}(t) \\ \boldsymbol{y}(t) &= \boldsymbol{A}_{(\ell \times \kappa)} \cdot \boldsymbol{x}(t)\,, \qquad t \geq 1 \end{cases} \tag{5.3}$$

where $\{\boldsymbol{e}(t)\}$ is an i.i.d. Gaussian process with variance-covariance matrix $\boldsymbol{Q}$, and $\boldsymbol{F}$ is a known matrix. Further $\boldsymbol{x}(0) \sim Normal\,(\boldsymbol{m}, \boldsymbol{V})$ and independent of $\boldsymbol{e}(t)$ for $t \geq 1$.

Classical state-space modeling strategies a la Box and Jenkins would look for the additional constraints needed to solve Problem 2 in a known dynamical behavior suggested by some physical law underlying the specific problem at hand and from known seasonal patterns in the traffic, for example the laws of motion in tracking the trajectories of moving objects, or from the presence of strong cross-correlations among the OD flows. This knowledge would translate into constraints on $\boldsymbol{F}$, and $\boldsymbol{Q}$ in the system 5.3 above, and would serve the critical role of driving the inferences towards one particular solution.

Augmented Gaussian State-Space Model The following Gaussian state-space model with drift is used to obtain the preliminary estimates for the OD flows.

$$
\begin{aligned}
&\begin{cases} \boldsymbol{x}(t) &= \boldsymbol{F}\cdot\boldsymbol{x}(t-1)+\boldsymbol{Q}\cdot\mathbf{1}+\boldsymbol{e}(t) \\ \boldsymbol{y}(t) &= \boldsymbol{A}\cdot\boldsymbol{x}(t)+\boldsymbol{\epsilon}(t) \end{cases} \\
&= \begin{cases} \begin{bmatrix} \boldsymbol{x}(t) \\ \mathbf{1} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{F} & \boldsymbol{Q} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix}\cdot\begin{bmatrix} \boldsymbol{x}(t-1) \\ \mathbf{1} \end{bmatrix}+\begin{bmatrix} \boldsymbol{e}(t) \\ \mathbf{1} \end{bmatrix} \\ \boldsymbol{y}(t) &= [\boldsymbol{A}|\,\mathbf{0}\,]\cdot\begin{bmatrix} \boldsymbol{x}(t) \\ \mathbf{1} \end{bmatrix}+\boldsymbol{\epsilon}(t) \end{cases} \\
&= \begin{cases} \tilde{\boldsymbol{x}}(t) &= \tilde{\boldsymbol{F}}\cdot\tilde{\boldsymbol{x}}(t-1)+\tilde{\boldsymbol{e}}(t) \\ \boldsymbol{y}(t) &= \tilde{\boldsymbol{A}}\cdot\tilde{\boldsymbol{x}}(t)+\boldsymbol{\epsilon}(t) \end{cases}
\end{aligned}
\tag{5.4}
$$

for $t \geq 1$, where $\mathbf{1} = (1, ..., 1)'$ is a constant vector of the length κ, the parameter $\phi(t)$ enters into the variance-covariance matrix of $\boldsymbol{e}(t) \sim N(\mathbf{0}, \phi(t)\cdot\boldsymbol{Q}^{\tau})$, $\boldsymbol{x}(1) \sim N(\mathbf{0}, \boldsymbol{V}(1))$, $\boldsymbol{\epsilon}(t) \sim N(\mathbf{0}, \boldsymbol{R})$, $\boldsymbol{x}(1) \perp \boldsymbol{e}(t)$ and $\boldsymbol{x}(1) \perp \boldsymbol{\epsilon}(t)$ for all $t \geq 1$, and finally $\boldsymbol{Q}$ is a diagonal matrix with elements $(q_1, ..., q_\kappa)$, and τ is a known constant. In the model above, if we set $\boldsymbol{F} = \mathbf{0}$ there is a one-to-one mapping between $(q_1, ..., q_\kappa, \phi(t))'$ and the unique elements in $E(\boldsymbol{y}(t)), V(\boldsymbol{y}(t))$. Further it is straightforward to verify that the following lemma holds.

Note 4. *The linear Gaussian state-space model in equations 5.4 contains the model in Cao et al. (2000) as a special case. Such a model can be obtained by simply setting $\boldsymbol{F} = \mathbf{0}$, hence imposing independence among the origin-destination flows $\boldsymbol{x}(t)$ at different epochs.*

In the experiments on Carnegie Mellon origin-destination traffic, assuming a fixed relationship between $x(i,t)$ and $x(i,t+1)$ is an unrealistic constraint. One possible solution is to assume a relationship between the means of the OD flows $\lambda(i,t)$ and $\lambda(i,t+1)$ instead, and to

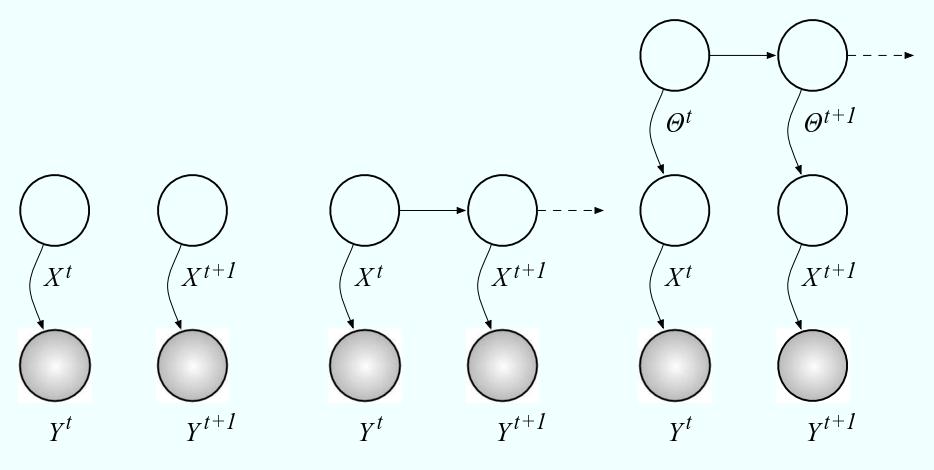

Figure 5.4: Graphical representations: models with no explicit time dependance (left); linear state-space models introduces an explicit dynamical behavior (center); the inverse allocation (IA) model moves the explicit time dependence one layer up in the graphical model, thus allowing for the OD flows to be more diverse (right).

allow for some error. The SSM yields smooth estimates that capture information about this relationship, which we pass to the next estimation stage. In fact, we introduce soft constraints on the average process $\{\boldsymbol{\lambda}(t)\, t \geq 1\}$ in the form of informative priors for the parameters underlying its dynamical behavior. We reduce the number of parameters by merging dynamic and error terms into a stochastic dynamical behavior. The marginal models for the OD traffic flows are independent log-Normals[3]. The main objects of interest are then the posterior distributions $P(\, \boldsymbol{x}(t) \,|\, \boldsymbol{y}(1), ..., \boldsymbol{y}(t)\,)$. In particular the point estimate for the OD traffic vector at time t is given by the mean $\hat{\boldsymbol{x}}(t) = E(\, \boldsymbol{x}(t) \,|\, \boldsymbol{y}(1), ..., \boldsymbol{y}(t)\,)$.

Static Inverse Allocation Model The static version of the IA model considers independent problems at each epoch. Briefly, we are interested in estimating $\hat{\boldsymbol{x}}(t) = E(\, \boldsymbol{x}(t) \,|\, \boldsymbol{y}(t)\,) = E(\, \boldsymbol{x} \,|\, \boldsymbol{y}\,)$.

[3] Airoldi (2003) also considers Gamma models.

To specify the full models at each time t we write:

$$\begin{cases} \boldsymbol{x} \,|\, \boldsymbol{\lambda}, \phi & \sim & p\left(\boldsymbol{\lambda}, \phi\right) \\ \boldsymbol{y} & = & \boldsymbol{A} \cdot \boldsymbol{x}, \end{cases} \tag{5.5}$$

where p is log-Normal, parameterized so that $E(\, x(i) \,|\boldsymbol{\lambda}, \phi\,)$ = $\lambda(i)$, $V(\, x(i) \,|\boldsymbol{\lambda}, \phi\,)$ = $\phi \cdot \lambda(i)^{\tau}$, $Cov(\, x(i), x(j) \,|\boldsymbol{\lambda}, \phi\,) = 0$ for $i = 1, ..., \kappa$ and $i \neq j$. Notice that ϕ is common across OD flows at each epoch, and that τ is a known scalar, which we obtain by inspection of $\boldsymbol{Y}$. The priors for the $\lambda(i)$ are $log\text{-}Normal\,(\theta_1(i), \theta_2(i))$[4], for $i = 1, ..., \kappa$ and independent for $i \neq j$. The prior for ϕ is proportional to a constant, to $1/\phi$ or to $1/\phi^2$.

Dynamic Inverse Allocation Model This dynamic version of the IA model, which yields the best results, implements the following Bayesian dynamical system:

$$\begin{cases} \lambda(i,t) & = & \epsilon(i,t) \cdot \lambda(i, t-1),\ i = 1, ..., \kappa \\ \boldsymbol{x}(t) \,|\, \boldsymbol{\lambda}(t), \phi(t) & \sim & p\left(\boldsymbol{\lambda}(t), \phi(t)\right) \\ \boldsymbol{y}(t) & = & \boldsymbol{A} \cdot \boldsymbol{x}(t), \quad t \geq 1, \end{cases} \tag{5.6}$$

where p is log-Normal, parametrized so that $E(\, x(i,t) \,|\boldsymbol{\lambda}(t), \phi(t)\,)$ = $\lambda(i,t)$, $V(\, x(i,t) \,|\boldsymbol{\lambda}(t), \phi(t)\,)$ = $\phi \cdot \lambda(i,t)^{\tau}$, and $Cov(\, x(i,t), x(j,t) \,|\boldsymbol{\lambda}, \phi\,) = 0$ for $i = 1, ..., \kappa$ and $i \neq j$. Notice that $\phi(t)$ is common across OD flows at time t, and that τ is a known scalar, which we obtain by inspection of $\boldsymbol{Y}$. The priors for $\lambda(i,0)$ are $log\text{-}Normal\,(\theta(i,0), \sigma)^4$, for $i = 1, ..., \kappa$ and independent for $i \neq j$, and for a big number σ that accounts for the uncertainty of the means of OD flows at time zero. The prior for $\phi(t)$ is proportional to a constant, to $1/\phi(t)$ or to $1/\phi(t)^2$. The priors for $\epsilon(i,t)$ are $log\text{-}Normal\,(\theta_1(i,t), \theta_2(i,t))^4$ for $i = 1, ..., \kappa$, and independent for $i \neq j$.

[4] Airoldi (2003) also considers Gamma, Uniform, and truncated Gaussian priors.

Table 5.2: A summary of the models.

Model	Time Dependence	Online Estimation	Skewed Marginals
Local likelihood	No	No	No
Augmented Gaussian SSM	Yes	Yes	No
IA (static)	No	No	Yes
IA (dynamic)	Yes	Yes	Yes

Informative Priors for $\boldsymbol{\lambda}(t)$ The crucial question at this point is: how do we calibrate the hyper-parameters underlying the prior distributions of $\boldsymbol{\lambda}(t)$? First we obtain a preliminary set of estimates $\hat{\boldsymbol{x}}(t)$ with the Gaussian linear SSM. Then, in the case of IA static, $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ at each time are set so that mean and variance of $\boldsymbol{\lambda}$ correspond to those of $\hat{\boldsymbol{x}}(t)$. Variances can be made much larger without significant loss of precision. The intuition is that the preliminary estimates indicate us where OD flows are on average. In the case of IA dynamic the intuition is the same, however it is not possible to set priors for $\boldsymbol{\lambda}(t)$ as the sequence $\{\boldsymbol{\lambda}(1), ..., \boldsymbol{\lambda}(T)\}$ is going to be determined by $\boldsymbol{\lambda}(0)$ alone. The solution is then to extract from $\{\hat{\boldsymbol{x}}(1), ..., \hat{\boldsymbol{x}}(T)\}$ information about their local dynamical behavior and use it to calibrate informative priors for $\{\boldsymbol{\epsilon}(t), t \geq 1\}$. Technically, we set $\epsilon(i,t)$ as independent log-Normals; we use the facts that product convolution of log-Normals is log-Normal (equation 4), and that $log\text{-}Normal(\theta_1(i,t), \theta_2(i,t)) = exp\{\, N(\theta_1(i,t), \theta_2(i,t))\,\}$ to solve the convolution problem exactly for $(\theta_1(i,t), \theta_2(i,t))$, for $i = 1, ..., \kappa$. In other words, values for $(\boldsymbol{\theta}_1(t), \boldsymbol{\theta}_2(t))$ are computed from $(\hat{\boldsymbol{x}}(t), \hat{\boldsymbol{x}}(t-1))$ at each time, and these parameters need not be learned. $\theta(i,0)$ is set to be the average of corresponding OD flow $\{x(i,t),\, t \geq 1\}$.

Notice that every two-stage method that finds preliminary estimates and refines them uses $\{\hat{\boldsymbol{x}}(1), ..., \hat{\boldsymbol{x}}(T)\}$ in the second stage, in some way. It is preferable to translate this information into information about the means of the OD flows $\{\boldsymbol{\lambda}(1), ..., \boldsymbol{\lambda}(T)\}$, according to the intuition that preliminary estimates can identify a smooth version of the OD flows we are looking for, which make reasonable guesses for their underlying average processes.

5.1.3 Estimation and Inference

The estimation strategy involves two stages. In the first stage we find preliminary, smooth estimates for the OD flows, which make a good guess for the averages of the OD traffic. In the second stage we refine these smooth estimates by looking for spikes and bursty periods with one single pass over the data.

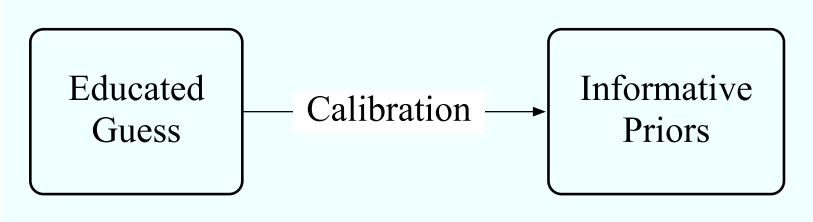

Figure 5.5: A non-parametric empirical Bayes approach to the filtering problem is at the core of the inverse allocation dynamic model.

The IA dynamic model is a Bayesian dynamical systems; EM and particle filter can be used for estimation and inference. The implementation includes skewed models as Gamma and log-Normal, a wide selection of priors as Uniform, Normal, Gamma and log-Normal, and several resampling schemes to further validate the results on top of the main particle filter. Ghahramani and Hinton (1996) show how to learn all the parameters in the linear Gaussian system 5.3, in our case $\boldsymbol{F}$, $\boldsymbol{Q}$, $\boldsymbol{m}$, and $\boldsymbol{V}$, by means of the EM algorithm. Higuchi (2001) shows how a self-organizing system can be built from non-linear non-Gaussian systems, so that all the relevant parameters are learned during the filtering process. Gilks and Berzuini (2001) propose a particle filter that keeps particles diverse. More specifically, we use the linear Gaussian SSM and related EM steps proposed in Airoldi (2003), which includes the model in Cao et al. (2000) as a special case, to obtain smooth estimates of the OD traffic, and we then use these estimates to calibrate informative priors for the parameters underlying the dynamic of a non-Gaussian system, in non-parametric empirical Bayes fashion. Eventually the particle filter makes good use of these priors and of the skewed models, and finds a sequence of better posterior distributions for the traffic flow

on each OD route; we pick their means as point estimates.

In order to filter the posterior distributions of the origin-destination flows and estimate the parameters of the models, I used a variation of the sample-resample-move algorithm of Gilks and Berzuini (2001), briefly outlined below. For simplicity define $\boldsymbol{v}(t)$ to be the vector of all parameters in the model at time t, $\boldsymbol{v}(t) := \big(\boldsymbol{x}(t), \boldsymbol{\lambda}(t), \boldsymbol{\epsilon}(t), \phi(t)\big)$, and $\boldsymbol{v}(0) := \big(\boldsymbol{\lambda}(0)\big)$. The *enhanced* particle filter algorithm is as follows. At $t = 0$ generate N particles $\{\, \tilde{\boldsymbol{\lambda}}_{(i)}(0) \,\}_{i=1}^{N}$ using $\boldsymbol{\theta}(0), \sigma$. Then iterate,

1. Set $t = t + 1$. Move each particle like so: (a) generate $\boldsymbol{\epsilon}_{(i)}(t)$ using $(\boldsymbol{\theta}_{1,(i)}(t), \boldsymbol{\theta}_{1,(i)}(t))$, and $\phi_{(i)}(t)$; (b) compute $\boldsymbol{\lambda}_{(i)}(t)$ using the equation 4; (c) generate $\boldsymbol{x}_{(i)}(t)$ by sampling from equation 5.

2. Resample N new particles from $\{\, \tilde{\boldsymbol{v}}_{(i)}(t) \,\}_{i=1}^{N}$ according to the likelihood, $P(\, \boldsymbol{y}(t) \,|\, \tilde{\boldsymbol{v}}_{(i)}(t) \,)$, they entail.

3. Move the new set of particles according to a MCMC for "several steps" to improve their diversity. Go to 1.

For details about the MCMC see Airoldi (2003).

Scalability and Irreducibility A recent result in network tomography (Cao et al., 2001) states that it is possible to reformulate filtering problems corresponding to large networks as a sequence of problems corresponding to small networks. As a consequence of it, the following result is true.

Lemma 1. *The complexity of the learning algorithm for the dynamic IA model is* $O(\, \kappa \cdot T \,)$.

Proof. The result in Cao et al. (2001) implies that a tomography problem corresponding to a network with κ origin-destination flows is equivalent to $O(\kappa)$ tomography problems, which correspond to disjoint sub-sets of, say, one to four OD traffic flows in the original problem. This fact

along with the fact that our solution is linear in the number of time points for which the OD traffic need be filtered, yields a total complexity of $O(\kappa \cdot T)$ for the learning algorithm of the dynamic IA model. □

Lemma 1 implies if we solve the inverse problem for small size networks, we immediately solve it for arbitrary size networks with comparable estimation errors. Further the following result holds.

Lemma 2. *The inference strategy is based on an irreducible MCMC.*

Proof. See Appendix A. □

Lemma 2 implies that the proposed inference strategy is able to explore the support of the whole joint posterior distribution of the OD flows. Note that, as hinted in the introduction to the problem, this fact cannot be taken for granted in inverse problems; being able to identify and explore the space of solutions is an issue that needs be addresses, problem by problem. Furthermore, the MCMC uses a Gibbs sampler with Metropolis steps.

Discussion of Experimental Evidence The methods were tested on two data sets; both included validation data.

- Carnegie Mellon traffic: the first data set, which we used to choose the appropriate model, contained about 12100 origin-destination traffic flows measured every 5 minutes over slightly less than two days at Carnegie-Mellon university (CMU). We measured an average traffic of 14GB every 5 minutes.

- AT&T traffic: the second data set, which we used to test and compare the filtered traffic obtained with different methodologies, contained 16 origin-destination flows measured every 5 minutes over a one-day period at AT&T, courtesy of Dr. Jin Cao at Bell Labs.

The analysis of Carnegie Mellon origin-destination traffic flows supports the hypothesis of a very skewed distribution. In figure 6 we plotted the logarithms of the observed flows versus the logarithms of the number of times measurements of such a size appear (aka. log-log plot), after discarding the measurements smaller than a standard packet (53 bytes = 424 bytes). The log-log plot indicates a log-Normal distribution may be appropriate. A histogram of the logarithms of the flows indicated that a logarithmic transformation is actually too mild to remove all the skewness, and a double logarithmic transformation would be more appropriate. The AT&T data set is much smaller, and contains traffic flows generated on a smaller network; they are less skewed overall, and a logarithmic transformation is enough to yield a symmetric histogram for the truncated flows. The CMU data set was used to inform model development. The AT&T data set was then used as an independent model validation data set.

The full story about the data sets is presented elsewhere (Airoldi, 2003; Airoldi and Faloutsos, 2004); here I will focus on findings that bear relevance to the methodological issues. In particular, few discussion points emerge that are shared by dynamic hierarchical models in applications to inverse problems.

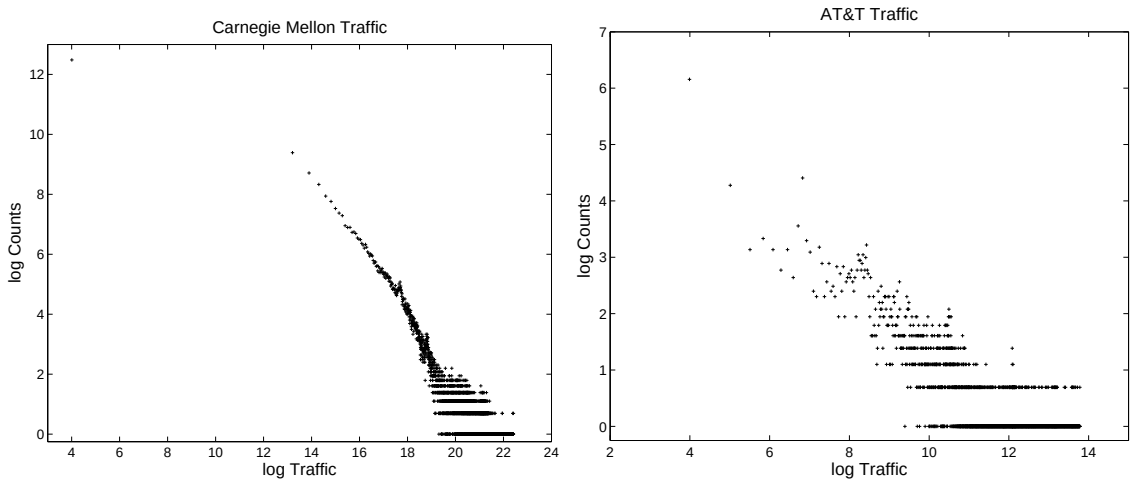

Figure 5.6: Log-log plots of the 12100 traffic flows measured at Carnegie-Mellon (left panel) and of the 16 traffic flows measured at AT&T (right panel).

1. *Skewed Marginals:* what is the impact of skewed model on the accuracy of the estimates? And what is the best model for the OD traffic?

2. *Time Dependence:* what is the impact of explicit time dependence on the accuracy of the estimates?

3. *Informative Priors:* what constraints should we impose to solve the regularization problem? How do they impact the accuracy of our estimates?

The inferences obtained with different methods were compared by computing the ℓ_2 distance between the true OD flows in the validation set and the estimates. The best results were obtained with log-Normal distribution for the flows and Gaussian vague priors.

To isolate the effect of realistic distributions for the OD flows, we compared the estimates obtained with IA where no time dependence was assumed, for Gamma and log-Normal models, and a variety of non-conjugate priors (Uniform, Gaussian, Gamma,l log-Normal) and different pa-rametrizations, with the estimates obtained by local likelihood. Introducing realistic model reduced the error between 25.4% and 40.8%. To isolate the effect of explicit time dependence, we compared the estimates we obtained with the augmented gaussian model that uses independent AR(1) processes for the OD flows, with the estimates obtained with local likelihood. Introducing time dependence reduced the error by 15.5% on average; the reduction ranged between 8.5% and 31.0%. Using the static IA model in 60% of the time points uninformative priors yield flat or multi-modal posteriors, whereas in the remaining 40% of the time points flat priors yield wide uni-modal posteriors. The main effect of the data at $\boldsymbol{y}(t)$ on the posterior $P(\boldsymbol{x}(t)|\boldsymbol{y}(t))$ is on its range; impossible configurations receive zero posterior probability. Informative priors with wide variance all yield uni-modal distributions. The dynamic IA model with informative priors has the advantage of requiring fewer particles than the version based on flat priors; knowing where to sample may introduce bias, but the thick tails of the log-Normal distribution of both $\boldsymbol{x}(t)|\boldsymbol{\lambda}(t), \phi(t)$

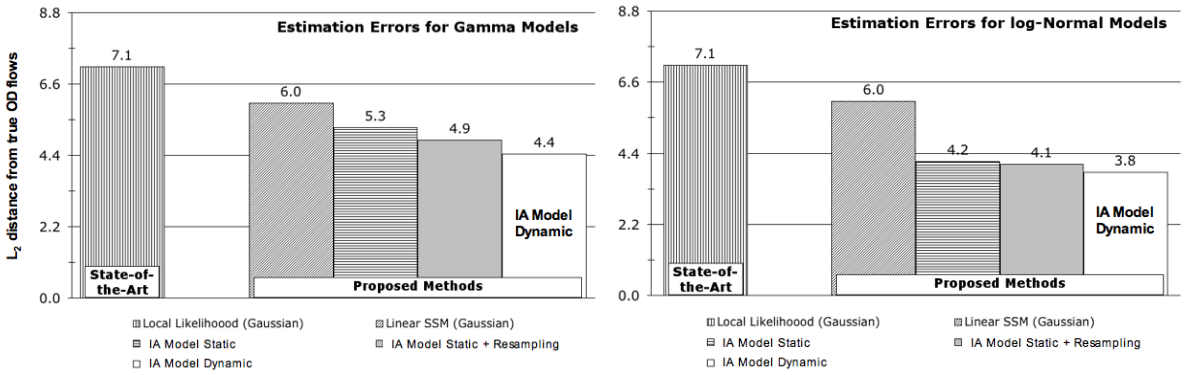

Figure 5.7: The bars represent the average estimation error in a validation set. Specifically we plot the ℓ_2 distance between the true OD flows and the corresponding estimates obtained with the local likelihood approach and IA models in its various flavors. IA based on the Bayesian dynamical system is a clear winner. In both panels we include the estimates obtained with the augmented Gaussian state-space model. Error bars in the left panel correspond to IA models based on Gamma, whereas error bars on the right panel correspond to IA models based on log-Normal.

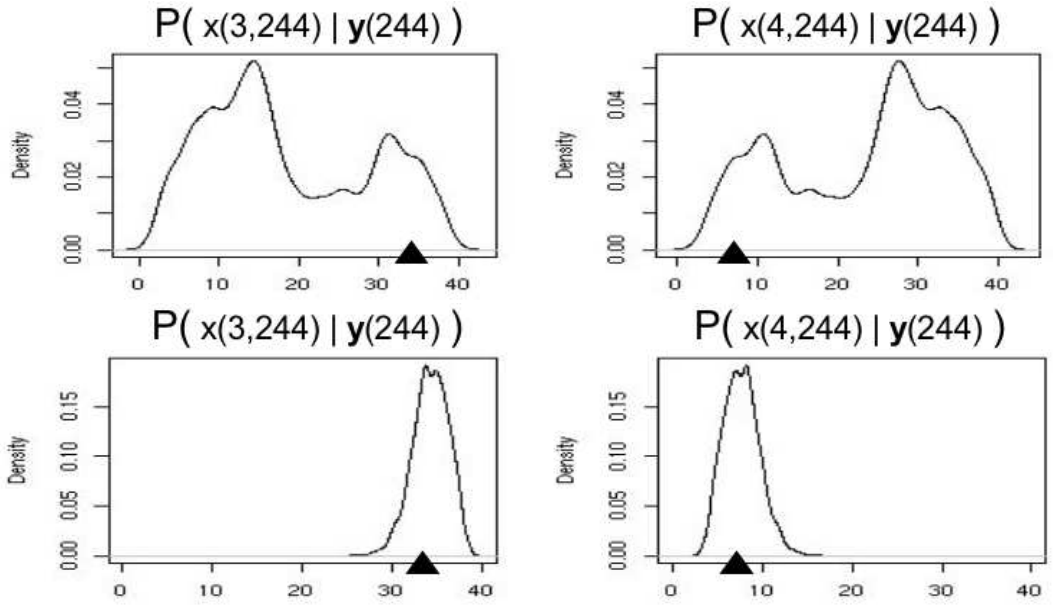

Figure 5.8: Example posterior distributions for the OD flows $x(3, 244)$ and $x(4, 244)$. The traffic on the X axes is measured in Kbytes, and the figures show the posterior distributions we obtained with non-informative priors (top panel) and with informative priors (bottom panel) calibrated using our Gaussian linear SSM. The solid triangles represent the true hidden OD Flows, whereas our point estimates would be the means of the posterior distributions. Making the posteriors *more unimodal* improves the estimates by reducing the bias entailed by extra modes.

and $\boldsymbol{\lambda}(t)|\boldsymbol{\theta}_1(t), \boldsymbol{\theta}_2(t)$ mitigate the problem, and IA captures several of the hidden spikes.

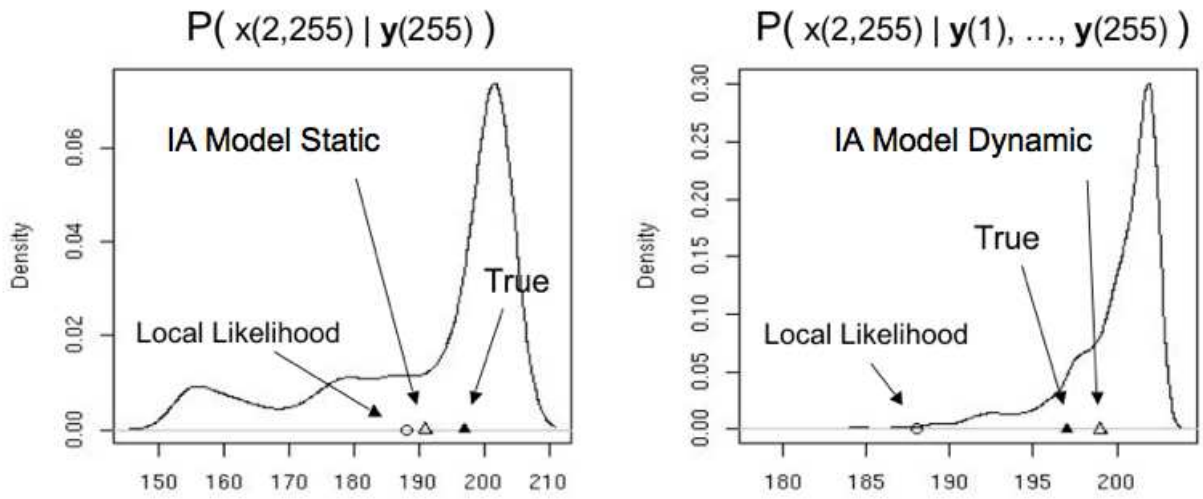

Figure 5.9: Example posterior distributions for the OD flow $x(2, 255)$. The traffic on the X axes is measured in Kbytes, and the figures show the posterior distribution we obtained with IA static (left panel) versus the one we obtained with IA dynamic (right panel). The solid triangles represent the true hidden OD Flow, whereas the empty triangles are our point estimates, which correspond to the means of the posterior distributions. Making use of all the observations $\{ \boldsymbol{y}(1), ..., \boldsymbol{y}(255) \}$ in computing the posterior distribution in the right panel reduced its variability — notice the different ranges — thus improving the inferences.

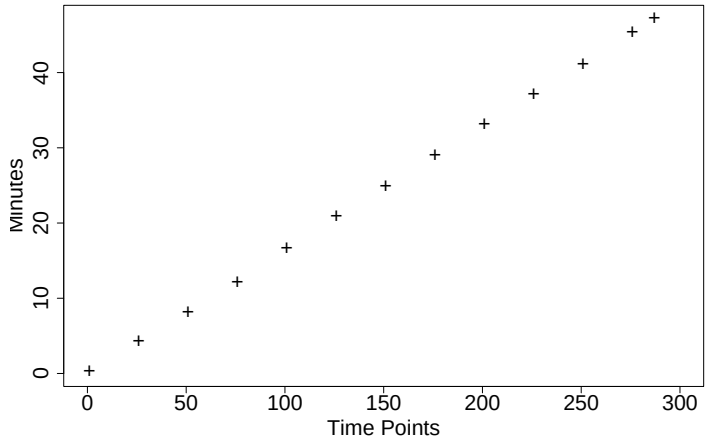

Figure 5.10: The learning algorithm for IA models scales linearly with the problem size (number of time ticks).

Briefly, we recover a smooth version of the OD flows, we calibrate informative priors for some crucial parameters, and eventually we use a dynamical Bayesian system to refine the estimates and

capture bursty traffic. This methodology allows us to combines the three simple ideas above: a realistic model for the data, the use of a filtering scheme which takes advantage of time, probabilistic constraints to overcome the under-determinacy of the problem. In the first stage we use the Gaussian linear SMM proposed in Airoldi (2003), and we calibrate informative priors for $\boldsymbol{\lambda}(t)$ using these estimates. These priors incorporate information about the magnitude and the dynamical behavior of the first stage smooth estimates, and softly constrain the location of the average processes $\{\boldsymbol{\lambda}(t),\, t \geq 1\}$. Other methods proposed in the literature make use of preliminary estimates, but they only retain the information about the magnitude of the OD flows given by the such estimates in the refining stage — see for example Zhang et al. (2003) who use shrinkage to improve the

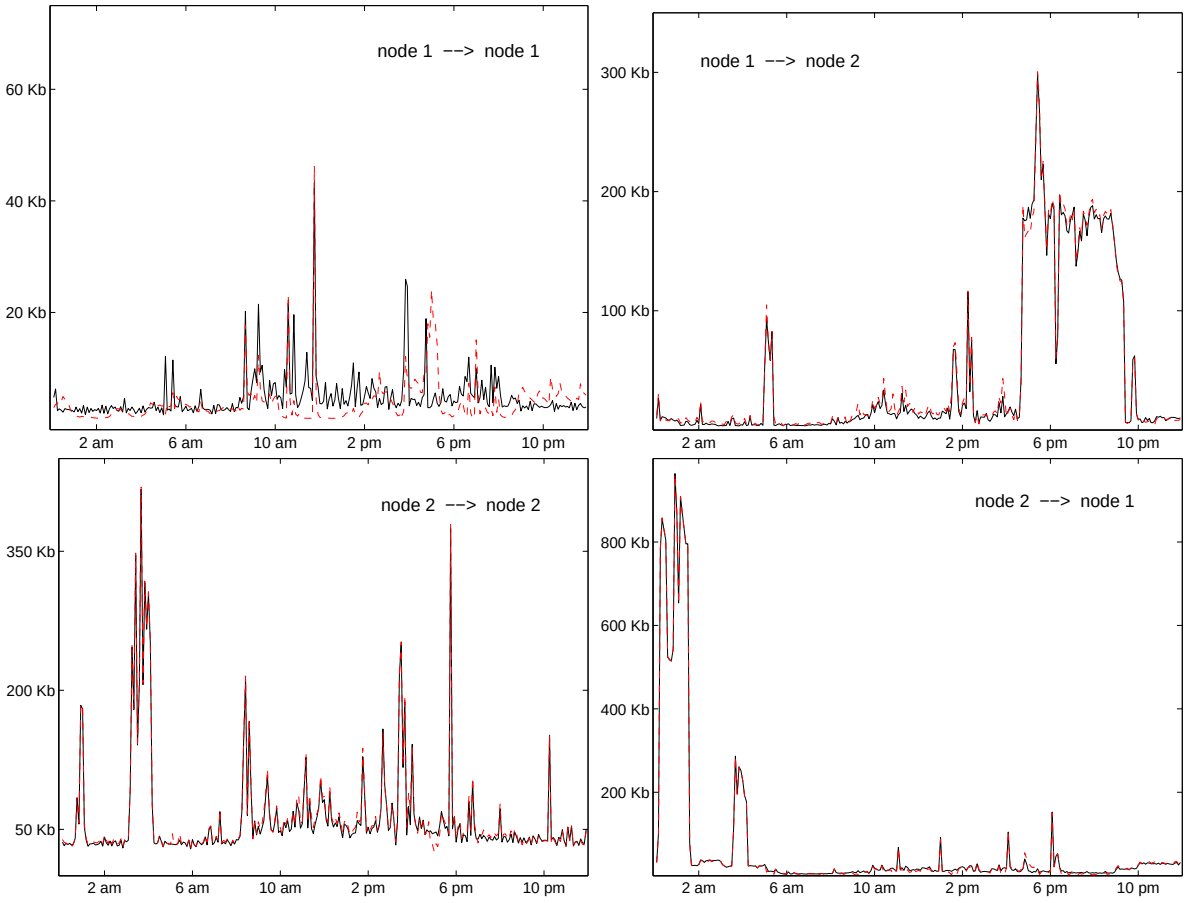

Figure 5.11: Example fits: actual latent flows (solid black lines) versus reconstructed flows (dashed red lines). IA manages to reconstructs several spikes.

solutions given by a gravity model. In our method, the fact that we retain also the information about the local dynamical behavior yields a significant jump in the final accuracy. Another channel through which informative priors help achieve a better accuracy is by reducing the bias entailed by multiple modes in the posterior distributions. Making the posteriors more *uni-modal* improves the precision of the point estimates of the OD flows (the posterior means) as we show in figure 10 below. Informative priors do drive the inferences about the OD flows towards the preliminary guesses, however the two layers of our model and the use of soft probabilistic constraints entail enough flexibility to capture several of the spikes in many cases, for an example see figure 12 below. Further our first-stage estimates are safely based on a model which entails a one-to-one relationship between OD flows and measurements, as it includes the model by Cao et al. (2000) as a special case. In the second stage the primary object of interest become the sequence of posterior distributions $P(\boldsymbol{x}(t) \mid \boldsymbol{y}(1), ..., \boldsymbol{y}(t))$. We use their means $\hat{\boldsymbol{x}(t)} = E(\boldsymbol{x}(t) \mid \boldsymbol{y}(1), ..., \boldsymbol{y}(t))$ as point estimates for the OD flows at time t. The Bayesian dynamical system brings further improvements, as we show in figure 7 above, due to the fact that we make use of all the observations up to time t in computing the posterior distributions $P(\boldsymbol{x}(t) \mid \boldsymbol{y}(1), ..., \boldsymbol{y}(t))$; conditioning on more observations yields a narrower variability. Local methods use fewer observations in a short window around t, instead.

Concluding, experimental evidence shows that the improvement IA models achieve goes beyond the contribution of state-of-the-art methods even when combined with recent resampling schemes which improve any given set of estimates. The modeling choices behind IA models are intuitive; first-stage estimates capture smooth average processes, second-stage estimates capture the spikes. Last, the estimation strategy of the dynamic IA model provides some insight in how to calibrate informative priors in Bayesian systems, where no clear guidance about the dynamic of the latent variables is available.

5.2 Co-Evolving Systems

The idea here is to revisit a classical model of social interactions and their evolution based on constructional theory (Carley, 1990, 1991), and to explore whether, and to what extent, its specifications fit within the statistical framework presented in the previous sections. In doing so, few points of discussion emerge that suggest a wide applicability of this approach.

The basic constructional model explains the dynamics of social interactions using three basic forces: (i) social interactions lead to shared knowledge; (ii) similar individuals tend to interact, and the more individuals interact the more similar they become; (iii) global social consensus emerges from diverse local conditions. Elements of the model portray a simplified society with N individuals. Culture is described in terms of individuals' knowledge about K facts, at any given period t, and encoded by Bernoulli variables $f^t(n,k)$ specific to individual-fact pairs. Social structure is defined in terms of individuals' probabilities of interaction with one another at any given period t, and encoded by scalars $p^t(n,m)$ specific to pairs of individuals. Social structure is assumed to be a deterministic function of the culture,

$$p^t(n,m) = \frac{\sum_k f^t(n,k) \cdot f^t(m,k)}{\sum_{o,k} f^t(n,k) \cdot f^t(o,k)}. \tag{5.7}$$

Actual interactions occur at any given period t, and are denoted by $i^t(n,m)$. Whenever two individuals interact, each shares knowledge about a single fact k, chosen uniformly among those that are known; this information is encoded by a pair of Bernoulli variables, $u^t(n,k)$, $u^t(m,l)$. And so culture evolves, an social structure changes.

The algorithm that specifies the evolution of social structure and culture in this model is as follows.

1. At epoch t

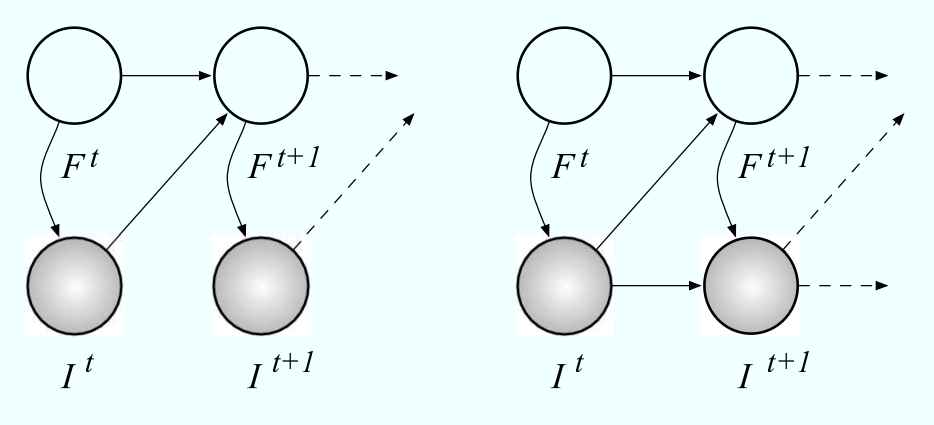

Figure 5.12: Graphical representation of the basic constructional model (Carley, 1990, 1991) and one of its extensions (Shreiber, 2006).

1.1. Compute social structure P^t given culture F^t

1.2. Sample interactions I^t given social structure P^t

1.3. Sample knowledge exchanged U^t given interactions I^t and culture F^t

1.4. Update culture F^{t+1} given previous culture F^t and knowledge exchanged U^t

The left panel of Figure 5.12 shows how the basic constructional model can be represented in the formalism of the statistical framework presented here.

Remark 2 (On Inference)**.** *Recall that it is possible to characterize the models of graphs based on the exponential family (Frank and Strauss, 1986; Wasserman and Pattison, 1996) with the formalism of undirected graphical models (Airoldi, 2006; Hanneke and Xing, 2007). The inference for the models of dynamics and evolution suggested by the constructional model of social interactions is tractable, although possibly computationally expensive: as long as we make use of probability distributions within the exponential family we can compute derivatives and likelihood and devise the corresponding EM algorithm—using approximation strategies such as variational methods and MCMC where necessary*

The constructional model of social interactions is essentially a data generating process that

involves probabilistic events and regularities; as such, its specifications can be subsumed within the statistical framework presented here. However, the goals of analyses that such a model (and its extensions) allows are not restricted to inference and parameter estimation. Through simulative experiments, for example, the constructional model allows to explore the space of possibilities that is consistent with a given set of structural hypotheses (Shreiber, 2006).

* * *

In this chapter, I described a strategy to introduce simple dynamics and evolution in the models of complex graphs developed so far. Edges within a network are no longer exchangeable in this temporal setting; exchangeability is substituted by other dependence structures.

I demonstrated how to fully specify model of network evolution—the Inverse Allocation model of Section 5.1—to solve an open problem in the context of dynamic network tomography. Network tomography constitutes an interesting application where a *locally smooth dynamic behavior* serves as the crucial constraint that allows the accurate estimation of origin-destination traffic from few aggregate traffic measurements. A conditional marked point process accommodates the extra variability due to bursts in the traffic.

I presented an overview of alternative (more complex) temporal modeling strategies and discussed the extent to which they provide a conceptual bridge between statistical models and agent-based models. I believe that this conceptual linkage suggests a new approach to calibration and validation issues that arise in agent-based models and simulations in general that is rooted in Bayesian statistics.

Chapter 6

Concluding Remarks

This thesis provides a methodological framework for the statistical analysis of complex graphs and dynamic networks. In it, I developed probabilistic algorithms that generate, evolve and integrate a heterogeneous collection of graphs, I studied the statistical models these algorithms implicitly specify, and I developed strategies for estimating the set of quantities on which they depend.

6.1 Conclusions

I have described a statistical approach to the analysis of complex systems. As it has emerged from the examples and case studies (either presented in details or referred to in published work) most of the models introduced here are tailored to the analysis of complex systems and their evolution, with special emphasis on applications to social and biological networks. The goals of the analysis in the various cases is different, but there is a binding theme: that of revealing non-observable mechanisms underlying social and biological processes by integrating a heterogeneous collection of measurements about diverse signals, i.e., networks, sequences, and attributes. Applications of the models presented here in the context of biological systems will be the main focus of future

research.

From a methodological perspective I introduced: (i) models for the analysis of complex networks; (ii) models for the analysis of multivariate attributes; (iii) strategies for integrating heterogeneous measurements; and (iv) models for the evolution of the system, within a coherent statistical framework. There are few basic ideas that get combined in various guises to derive full model specifications in this framework: (i) mixed membership; (ii) latent patterns; (iii) hierarchical structure in the likelihood; (iv) dynamics; and (v) sparsity. I found these ideas to be useful in applications to social and biological systems.

In future research, I plan to explore fundamental technical issues that are shared by Bayesian mixed membership models, and to some degree by hierarchical Bayesian mixture models.

6.2 Technical Issues

In working with latent aspects models of the sort described in this thesis, I have encountered four themes of a technical nature: (i) the mixed membership of objects to patterns, and the related allocation task; (ii) model selection and model choice; (iii) the presence of many local peaks in the likelihood, and strategies for finding one with a good substantive interpretation; and (iv) scalability of the approach to very large data sets. I briefly touch upon each of these in the following subsections. The context in each case is given by a specific model, but the discussion and results generalize to other models.

6.2.1 The Geometry of Allocation

The allocation task has a central role in the latent aspects models described in this thesis; resolving this task is equivalent to estimating the mixed membership map between objects and latent patterns.

Intuitively, we allocate objects to categories and we introduce a new category when the fit is bad on some scale using the current number of categories. It is possible to characterize the notion of allocation in terms of variance components, both analytically and with simulations.

Example 29. *In the classical Factor Analysis and linear Gaussian state-space models it is possible to derive in closed form the projections of data onto the lower dimensional spaces of factors and states, respectively. The projection allocates data to latent components according to the entries of the various variance-covariance matrices involved, assuming equal component weights. Consider, for example, a factor analysis model: we measure D-dimensional quantities $Y = AX$, components of which are assumed to be sums, through the matrix A, of K-dimensional latent factors, X. In a simple formulation of the problem we can assume unit elements $A_{(ij)} = 1$, and $X \sim Normal\,(0, \Sigma)$. It turns out that the allocation of observations to factors, in this model, is resolved by estimating latent factors with weighted averages of the observations, $\mathbb{E}\,[\, X_{(k)} \mid \{Y\}_{n=1}^{N}\,] = \sum_d\, \omega_{kd}\, \bar{Y}_{(d)}$. The allocation is specified through the optimal (mixed-membership) weights, which are functions of the elements of the variance-covariance matrix, $\omega_{kd}^{*} = \omega_{k}^{*}(\hat{\Sigma})$. Note that the variance-covariance matrix Σ is estimated as well (e.g., see Appendix A.3 in Airoldi, 2003).*

The seeming analytical intractability of these models presents us with some obstacles, and opens analytical opportunities at the same time. Below I provide some experimental evidence that is suggestive of the how the quality of the allocation of objects to patterns responds to the quality of the assumptions encoded by a model. In the future I plan to explore the extent to which a tractable lower bound for the log-likelihood and asymptotic derivations help characterize these ideas.

Experimental Evidence: Simulations The simulation takes place in the context of models of multivariate attributes I developed in Section 4.1, where we allocate genes to temporal expression profiles using models that encode independence among occurrences of the same gene versus models of contagion. Simulative experiments suggest that Dirichlet-Poisson model of Section 4.1 is

better at recovering membership than the independence model when realistic SAGE mean/variance ratio holds.

We first validate our models by examining to what extend they can recover the mixed-membership probabilities $\{\theta_n\}$, i.e., the soft cluster assignments of each gene, under various simulated conditions. We generated the ground truth using our generative processes, and we focused on scenarios where the "mean" expression level at the various epochs was lower than its corresponding "variance"— a realistic biological experimental scenario. We compare our models, normalized DiP and conditional DiP, with two other methods, the independence model (Pritchard et al., 2000; Minka and Lafferty, 2002; Blei et al., 2003), and the PoissonL model (Cai et al., 2004). Our models yield higher likelihoods of expression profiles in the test set (not shown), and more accurate predictions of the latent theme id of each gene based on their observed expression levels. Out of 1000 genes we simulated, for example, nDiP and cDiP achieved 75.95% and 70.32% accuracy, respectively, whereas the independence model reached only 63.25%. Strikingly, the independence model clustered all genes in one profile in several runs.

Experimental Evidence: A 20-gene Synthetic Data Set In small samples bearing realistic SAGE characteristics, although the recovered clusters differ only slightly, the estimated mixed-membership are sharper using DiP than with the independence model.

Here I report our analysis of a small dataset used in Cai et al. (2004), which contains the expression profiles of 20 genes over 5 temporal epochs. Eighteen of the 20 genes belong to one of 4 clusters (temporal themes), and the 2 remaining two are identified as outliers. The expression profiles are generated from 6 different latent themes, or clusters, which the authors reduce to 4 by ignoring the abundance of the gene tags observed on the transcripts sampled at each epoch. In particular, there are 3 profiles from theme 1, 4 from theme 2, 6 from theme 3, and 6 from theme 4. The raw data is plotted in Figure 6.1 on various scales. Among the profiles from theme 2, there is

1 with 10 times as many gene tags as the others, and similarly for theme 3—number 7 and number 13 in Figure 6.2. Note that these 2 profiles are "more expressed" but they follow an expression theme similar to the other expression profiles in the respective clusters.

Figure 6.2, displays the 4 themes learned by the normalized and conditional DiP models (bottom-left panel), versus those learned by PoissonL (Cai et al., 2004) and the independence model (top-left panel). A rough eyeballing shows that the gene expression themes learned by DiPs and the two competing methods are similar. However, a close examination reveals the following. Arguably, we obtain a more compact themes 3, as revealed by the lower degree of dispersion among genes

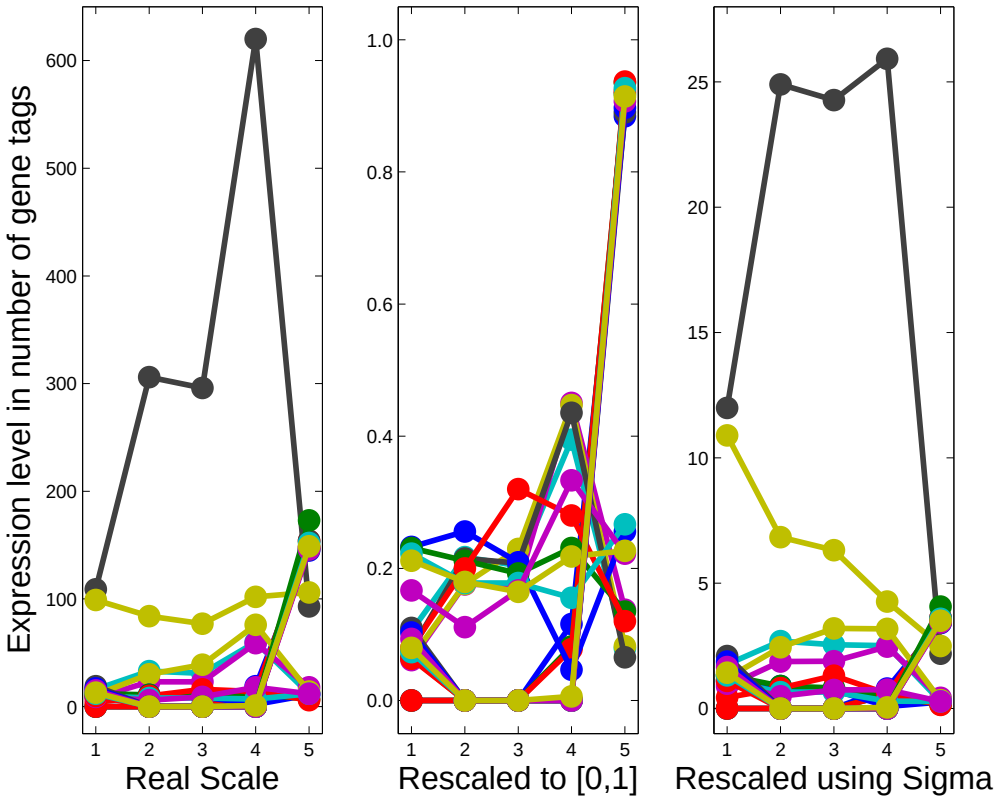

Figure 6.1: The raw example data in Cai et al. (2004), on the original expression scale (left); on a normalized expression scale, by gene, into $[0, 1]$ (center); and on a normalized expression scale, by epoch, using $\hat{\sigma}_{1:T}$ (right).

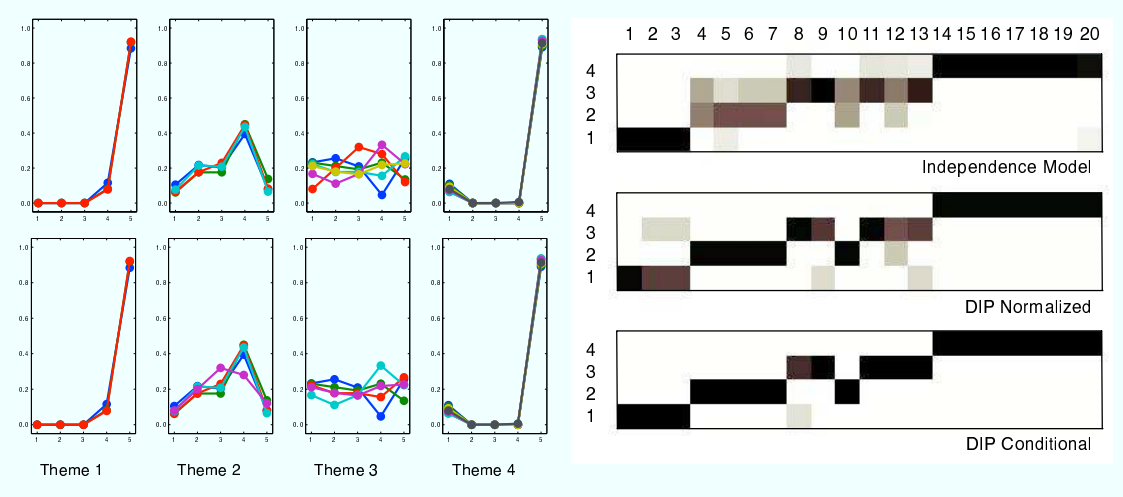

Figure 6.2: Left: Latent gene expression themes learned by different algorithms. Top: 4 themes (numbered 1 to 4 from left to right) learned by PoissonL and the independence model. Each theme is represented by the expression profiles of all the genes assigned to that theme base on MAP prediction using the estimated mix-membership vector θ_n. In this case, PoissonL and the independence model give the same membership prediction. Bottom: The 4 themes discovered by normalized DiP and conditional DiP. Note that due to overlap of the profile curves, the "occupancy" number of each theme is not apparent here. But in Fig. 6.2, one can see it more clearly. Right: The estimated membership probabilities, $\{\hat{\theta}_{nk}\}$, for the independence model (top), nDiP (middle), and cDiP (bottom). Each row correspond to a theme, and each column corresponds to a gene. The color shades of the cells correspond to values ranging from 1 (black) to 0 (white). The panel shows that cDiP yields the sharpest estimates.

assigned to this theme; but for theme 2, the genes assigned to it by the independence model and PoissonL are slightly more consistent. Overall, the software clustering assignment of each gene are compatible across all 4 algorithms, and as shown in Figure 6.2), but the mixed-membership probabilities inferred by the DiPs for each gene are sharper. If we compare the MAP assignment of each gene to a single most probable themes, the 19 of the 20 genes are consistent across all 4 algorithms, and their assignments agree with the true themes label given by the original dataset. The remain one, gene no. 10, is intriguing. It has an expression profile, $\{Y_{10}^{1:5}\} = (4, 10, 16, 14, 6)$, and is originally labeled as from theme 2, $\{\lambda_2^{1:5}\} = (10, 30, 30, 60, 10)$. Apparently profile $\{Y_{10}^{1:5}\}$ exhibits great variability with respect to its supposedly underlying theme. Using DiP, we infer the label of gene no. 10 to be theme 3, which has a prototype profile $\{\lambda_3^{1:5}\} = (10, 10, 10, 10, 10)$, and

indeed we found much of the variability in gene 10 is related to the overall abundance of all genes in different epochs, rather then its intrinsic trend. So we feel this assignment is arguable more plausible the the purported theme 2. As shown in Figure 6.2), the independence model inferred a split assigned, about equally probable to pattern 2 and 3.

The example suggests the role of model properties in latent allocation tasks. The intuition is that if the model cannot express, on average, the salient properties of the data, then it may lead to artifactual effects. Specifically, the unexplained variability will need to find a "place-holder", and it seems to increase the variability of parameter estimates.

6.2.2 Model Selection Strategies and Issues

Although there are pathological examples, where slightly different model specifications lead to quite different analyses and choices of key parameters, in real situations we expect models with similar probabilistic specifications to suggest roughly similar choices for the number of patterns (Airoldi et al., 2006e). In the applications presented or referred to throughout this thesis I explored the issue of model choice by means of different criteria.

Paremetric: Choice Informed by the Ability to Predict Cross-validation is a popular method to estimate the generalization error of a prediction rule (Hastie et al., 2001), and its advantages and flaws have been addressed by many in that context (e.g., Ng, 1997). More recently, cross-validation has been adopted to inform the choice about the number groups and associated patterns in hierarchical Bayesian models (Barnard et al., 2003; Wang et al., 2005). Guidelines for the proper use of cross-validation in choosing the optimal number of groups K, however, has not been systematically explored. One of the goals of our case studies is that of assessing to what extent cross-validation can be "trusted" to estimate the underlying number of topics or disability profiles. In particular, given the non-negligible influence of hyper-parameter estimates in the evaluation of the held-out

likelihood, i.e., the likelihood on the testing set, we discover that it is important not to bias the analysis with "bad estimates" of such parameters, or with arbitrary choices that are not justifiable using preliminary evidence, i.e., either in the form of prior knowledge, or outcome of the analysis of training documents. To this extent, estimates with "good statistical properties," e.g., empirical Bayes or maximum likelihood estimates, should be preferred to others (Carlin and Louis, 2005). Alternative approaches based on the predictive ability of a set of latent patterns have been recently proposed, e.g. in the context of clustering (Tibshirani and Walther, 2005).

Semiparametric: Stochastic Process Priors Positing a Dirichlet process prior on the number of latent topics is equivalent to assuming that the number of latent topics grows with the log of the number of, say, documents or individuals (Ferguson, 1973; Antoniak, 1974). This is an elegant model selection strategy in that the selection problem become part of the model itself, although in practical situations it is not always possible to justify. A nonparametric alternative to this strategy, recently proposed (McAuliffe et al., 2006), uses the Dirichlet Process prior is an infinite dimensional prior with a specific parametric form as a way to mix over choices of K. This prior appears reasonable, however, for static analyses of scientific publications that appear in a specific journal. Kumar et al. (2000) specify toy models of evolution which justify the scale-free nature of the relation between documents and topics using the Dirichlet process prior for exploratory data analysis purposes (Kleinberg et al., 1999; Kumar et al., 2000). However, has to be noted that the prior on the membership of the patterns induced by many such processes is not always desirable, and in certain applications is wrong. For example, in biological applications to protein interaction networks, the latent patterns correspond to stable protein complexes (i.e., groups of proteins) that are composed of 4 to 7 proteins on average (Krogan et al., 2006).

Other Criteria for Model Choice The statistical and data mining literatures contain many other criteria and approaches to deal with the issue of model choice, e.g., reversible jump MCMC

techniques, Bayes factors and other marginal likelihood methods, and penalized likelihood criteria such as the Bayesian Information Criterion (BIC) (Schwartz, 1978; Pelleg and Moore, 2000), the Akaike information criterion (AIC) (Akaike, 1973), the deviance information criterion (DIC) (Spiegelhalter et al., 2002), minimum description length (MDL) (Chakrabarti et al., 2004). See (Han and Kamber, 2000) for a review of solutions in the data mining community. AIC has a frequentist motivation and tends to pick models that are too large when then number of parameters its large—it does not pay a high enough penalty. BIC and DIC have Bayesian motivations and thus fit more naturally with the specifications in this paper. Neither is truly Bayesian; however DIC involves elements that can be computed directly from MCMC calculations, and the variational approximation to the posterior (described in detail below), allows us to integrate out the nuisance parameters in order to compute an approximation to BIC for different values of K.

A Simulation Study I conclude by presenting some anecdotal evidence we gathered from synthetic data with the aim of highlighting the dangers of fixing the hyper-parameters according to some ad-hoc strategy that is *not* supported by the data, e.g., fixing $\alpha = 50/K$ in the models of the Chapters 3 and 4. I simulated a set of 3,000 documents according to the latent dirichlet allocation model for generating textual documents described in Example 22, with $K^* = 15$ topics (the patterns) and a vocabulary of size 50. I then fitted the correct Bayesian mixed-membership model on a grid for $K = 5, 10, 45$ that included the true underlying number of groups and associated patterns, using a five-fold cross-validation scheme. In a first batch of experiments I fitted alpha using empirical Bayes Carlin and Louis (2005), whereas in a second batch of experiments I set $\alpha = 50/K$, following the analysis in Griffiths and Steyvers (2004). The held-out log-likelihood profiles are reported in Figure 6.3.

In this controlled experiment, the optimal number of non-observable topics is $K^* = 15$. This implies a value of $\alpha = \frac{50}{15} = 3.33 > 1$ for the ad-hoc strategy, whereas $\hat{\alpha} = 0.052 < 1$ according to the empirical Bayes strategy. Intuitively, the fact that $\alpha > 1$ has a disrupting effect on the model

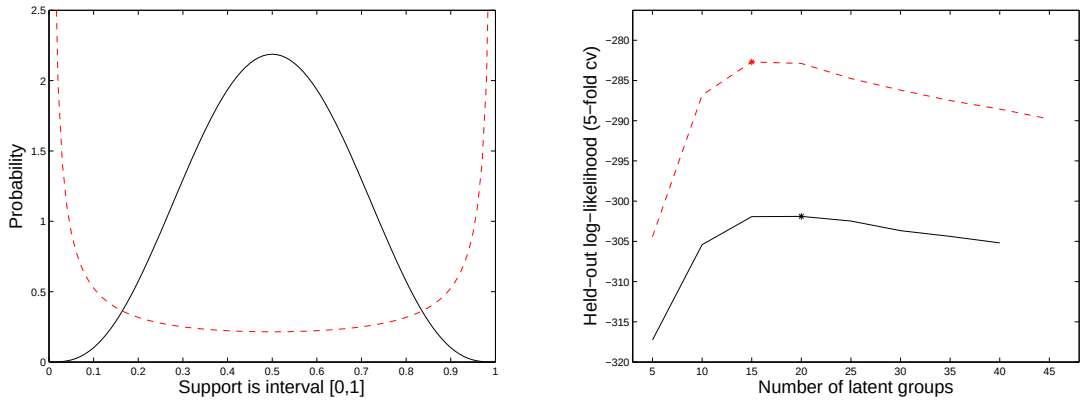

Figure 6.3: Left: 2D symmetric Dirichlet densities underlying mixed-membership vectors $\theta = (\theta_1, \theta_2)$, with parameter $\alpha = 4 > 1$ (solid, black line) and with parameter $\alpha = 0.25 < 1$ (dashed, red line). Right: held-out log-likelihood for the simulation experiments described in the text. The solid, black line corresponds to the strategy of fixing $\alpha = 50/K$, whereas the dashed, red line corresponds to the strategy of fitting α via empirical Bayes. K^* is denoted with an asterisk.

fit: each topic is expected to be present in each document, or in other words each document is expected to belong equally to each group/topic, rather than only to only a few of them, as it is the case when $\alpha < 1$. As an immediate consequence, the estimates of the components of mixed-membership vectors, $\{\theta_{nk}\}$, tend to be diffuse, rather than sharply peaked, as we would expect in text mining applications. Furthermore, in this simple simulation, setting the hyper-paramter α to a value greater than one when the data supports values in a dramatically different range, e.g., $0.01 < \alpha < 0.1$, ultimately bias the estimation of the number of latent patterns. Figure 6.3 shows that, ultimately, the empirical Bayes strategy correctly recovers $K^* = 15$, whereas the ad-hoc strategy finds recovers an erroneous number of latent patterns $K^* = 20$.

Concluding, experiments in a controlled setting suggest that it is desirable not to fix the hyper-parameters, e.g., the non-observable category abundances α, according to ad-hoc strategies, unless such strategies are supported by previous analyses. Ad-hoc strategies will affect inference about the number of non-observable patterns in non-controllable ways, and ultimately bias the analysis of data and the substantive conclusions. This effect can be observed in a real problem setting in

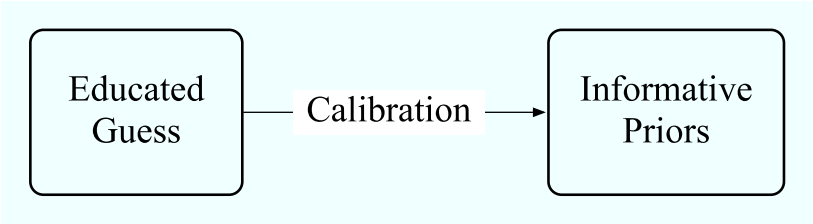

Figure 6.4: The non-parametric empirical Bayes approach at a glance.

Example 25 of Chapter 4 by looking at the entries in Figure 4.10. The plots in the right column display latent topics that were estimated using the strategy of fixing α; they are visibly more *diffuse* than the topics estimated by fitting the hyper-parameter α using the empirical Bayes strategy likelihood—plots in the left column.

6.2.3 Nonparametric Empirical Bayes

An issue with mixture models is that of multiple local peaks (e.g., Buot and Richards, 2006a,b). Depending on the signal-to-noise ratio in the data, this can lead to problematic inferences. However, even in those cases where the signal is buried in the noise it is possible to adopt estimation and inference strategies that minimize the issue.

Example 30. *In the application of the admixture of latent blocks model to protein interaction networks (Airoldi et al., 2006c) a two-stage approach is used; a model with no interactions among protein in different complexes is fit first, i.e., B is constrained to be the identity, and then the full model is fit, i.e., B is unconstrained. In the second stage, the mixed membership map is initialized to that recovered in with the simpler model. In this model, the strategy aims at resolving the inference among the two competing explanations for the interactions; namely, the mixed membership map between protein and stable protein complexes, and the block model that encodes interactions among proteins in different complexes.*

In order to perform inference in the models presented a multiple-stage approach to estimation

and inference is adopted—see Figure 6.4. In general, the non-parametric empirical Bayes approach is an engineering solution. The approach suggests fitting a sequence of models, from most simple to most complex, which are not necessarily nested. The results of estimation and inference in simpler models is used to inform (or calibrate) priors for teh parameters in the more complex models. In future work I plan to quantify both (i) the effect of signal-to-noise ratio that is needed to cause problems, and its interaction with (ii) the effect of the distance between the true model the starting model on the probability of successful estimation and inference.

6.2.4 Scalability

The scalability of posterior inference algorithms in models of relational data, e.g., the stochastic block models of mixed membership of Section 4.2.2, is a crucial practical issue given the size of social and biological networks of interest that arise in modern applications. Below, I illustrate a possible solution to perform fast posterior inference in the context of a specific model. Notably, the proposed *nested variational* inference strategy is applicable to other models of relational data and makes posterior inference feasible in applications that involve large graphs and networks.

Consider the admixture of latent blocks model of Section 3.1. To achieve fast convergence of the proposed posterior inference algorithm in that case, I employed a highly effective *nested* variational inference scheme based on a non-trivial scheduling of variational parameters updating. The resulting algorithm is also parallelizable on a computer cluster.

In a naïve iteration scheme for variational inference, one would initialize the variational Dirichlet parameters $\vec{\gamma}_{1:N}$ and the variational multinomial parameters $(\vec{\phi}_{p\to q}, \vec{\phi}_{p\leftarrow q})$ to non-informative values, and then iterate until convergence the following two steps: (i) update $\vec{\phi}_{p\to q}$ and $\phi_{p\leftarrow q}$ for all edges (p, q), and (ii) update $\vec{\gamma}_p$ for all nodes $p \in \mathcal{N}$. In such algorithm, at each variational inference cycle we need to allocate $NK + 2N^2K$ scalars. Experimental evidence (Airoldi et al.,

Outer loop

1. initialize $\vec{\gamma}^0_{pk} = \frac{2N}{K}$ for all p, k
2. **repeat**
3. **for** $p = 1$ to N
4. **for** $q = 1$ to N
5. get **variational** $\vec{\phi}^{t+1}_{p \to q}$ and $\vec{\phi}^{t+1}_{p \leftarrow q} = f\ \left(R(p,q), \vec{\gamma}^t_p, \vec{\gamma}^t_q, B^t \right)$
6. partially update $\vec{\gamma}^{t+1}_p$, $\vec{\gamma}^{t+1}_q$ and B^{t+1}
7. **until** convergence

Figure 6.5: The nested (two-layered) variational inference algorithm for γ and $(\phi^{\rightarrow}, \phi^{\leftarrow})$. The inner layer consists of Step 5. The function g is described in details in Figure 6.6.

2006d) suggests that the naïve variational algorithm often fails to converge, or converges after a large number of iterations. I attribute this behavior to a dependence that the two main assumptions (block model and mixed membership) induce between $\vec{\gamma}_{1:N}$ and B, which is not satisfied by the naïve algorithm. Some intuition about why this may happen follows. From a purely algorithmic perspective, the naïve variational EM algorithm instantiates a large coordinate ascent algorithm, where the parameters can be semantically divided into coherent blocks. Blocks are processed in a specific order, and the parameters within each block get all updated each time.[1] At every new iteration the naïve algorithm sets all the elements of $\vec{\gamma}^{t+1}_{1:N}$ equal to the same constant. This dampens the likelihood by suddenly breaking the dependence between the estimates of parameters in $\hat{\vec{\gamma}}^t_{1:N}$ and in $\hat{B}^t$ that was being inferred from the data during the previous iteration.

Instead, the nested variational inference algorithm maintains some of this dependence that is being inferred from the data across the various iterations. This is achieved mainly through a different scheduling of the parameter updates in the various blocks. To a minor extent, the dependence is maintained by always keeping the block of free parameters, $(\vec{\phi}_{p \to q}, \vec{\phi}_{p \leftarrow q})$, optimized given the other variational parameters. Note that these parameters are involved in the updates of parameters in $\vec{\gamma}_{1:N}$ and in B, thus providing us with a channel to maintain some of the dependence among

[1] Within a block, the order according to which (scalar) parameters get updated is not expected to affect convergence.

Inner loop

1. initialize $\phi^0_{p\to q,g} = \phi^0_{p\leftarrow q,h} = \frac{1}{K}$ for all g, h
2. **repeat**
3. **for** $g = 1$ to K
4. update $\phi^{s+1}_{p\to q} \propto f_1\left(\vec{\phi}^s_{p\leftarrow q}, \vec{\gamma}_p, B\right)$
5. normalize $\vec{\phi}^{s+1}_{p\to q}$ to sum to 1
6. **for** $h = 1$ to K
7. update $\phi^{s+1}_{p\leftarrow q} \propto f_2\left(\vec{\phi}^s_{p\to q}, \vec{\gamma}_q, B\right)$
8. normalize $\vec{\phi}^{s+1}_{p\leftarrow q}$ to sum to 1
9. **until** convergence

Figure 6.6: Details Step 5. in Figure 6.5; the inference algorithm for the variational parameters $(\phi^{\rightarrow}_{nm}, \phi^{\leftarrow}_{nm})$ corresponding to the basic observation y_{nm}. The functions g_1 and g_2 are updates for $\phi^{\rightarrow}_{nmg}$ and $\phi^{\leftarrow}_{nmh}$ described in the text of Section 4.1.3.

them, i.e., by keeping them at their optimal value given the data. Further, the nested algorithm has the advantage that it trades time for space thus allowing us to deal with large graphs; at each variational cycle we need to allocate $NK + 2K$ scalars only. The increased running time is partially offset by the fact that the algorithm can be parallelized and leads to empirically observed faster convergence rates. This algorithm is also better than MCMC variations (i.e., blocked and collapsed Gibbs samplers) in terms of memory requirements and convergence rates.

Complexity Recall that attribute measurements taken on individual objects in a population of interest can be represented as a bipartite graph, and that relational measurements taken on pairs of objects in a population of interest can be represented as a unipartite graph. In both cases, denote the number of edges in the graph by I, the number of objects by N, the number of attributes by M, the number of latent patterns by K, and the number of iterations till convergence of the posterior inference algorithm employed by T.

In summary, the complexity of fitting a model of multivariate attributes that follows the general

specifications of Section 4.2.1 is

$$O\left(I + NMKT + K^2T\right),$$

whereas the complexity of fitting a model of multivariate relations that follows the general specifications of Section 4.2.2 is

$$O\left(I + N^2KT + K^2T\right).$$

Appendix A

Proof of Lemma 2

The proof is based on the following result.

Fact 1. *There exists a permutation ρ of the columns of $A_{(\ell\times\kappa)}$ such that $[A]_{(i,\rho(j))} = [A_1 \,|\, A_2]$, where A_1 is $(\ell \times \ell)$ and has full rank, and A_2 is $(\ell \times (\kappa - \ell))$.*

As a consequence we can permute the components of X to get $[X]'_{\rho(i)} = [X_1 \,|\, X_2]'$, and $Y = A\,X = A_1\,X_1 + A_2\,X_2$, and finally express X_1 in terms of X_2 and Y, like so:

$$X_1 = A_1^{-1} \cdot (Y - A_2\,X_2)$$

Proof. The Gibbs sampler scheme involves iterative sampling from the full conditional distributions $P(Z_i|Z_{(-i)} = z_{(-i)})$, for $i = 1, ..., N$ and Z vector. A sufficient condition to ensure the irreducibility of the chain, Besag (1974), requires that the support of the full conditional distributions is positive where that of the joint distribution of Z is positive, that is:

$$\text{if} \quad P(Z_i = z_i, Z_{(-i)} = z_{(-i)}) > 0 \quad \Rightarrow \quad P(Z_i|Z_{(-i)} = z_{(-i)}) > 0. \tag{A.1}$$

2D case: we show that condition A.1 holds. Specifically consider the situation displayed in figure 6 above, where there are $\kappa - \ell = 2$ components of X_2 that we need to sample from. The chain is at a point $X_2 > 0$ where the joint support is positive and $A_1^{-1}(Y - A_2X_2) > 0$, and it moves by $(+\epsilon, +\epsilon)'$ to the point $X_2 + (\epsilon, \epsilon)'$ where the joint support is also positive and $A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, \epsilon)')\,) > 0$. We want to show that whenever both X_2 and $X_2 + (\epsilon, \epsilon)'$ are feasible, it is possible to pass from the former to the latter by means of component-wise moves, as we would with Gibbs moves; that is, the support of the full conditionals must be positive either at $A_1^{-1}(Y - A_2\,(X_2 + (0, \epsilon)')\,)$ or at $A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, 0)')\,)$. In other words we want to show that

$$\{\, A_1^{-1}(Y - A_2X_2) \geq 0 \quad \wedge \quad A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, \epsilon)')\,) \geq 0 \,\} \tag{A.2}$$

implies

$$\{\, A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, 0)')\,) \geq 0 \quad \vee \quad A_1^{-1}(Y - A_2\,(X_2 + (0, \epsilon)')\,) \geq 0 \,\}. \tag{A.3}$$

Assume that A.2 holds. Notice that $A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, \epsilon)')\,) = A_1^{-1}(Y - A_2X_2 - \epsilon(A_2^{11}, A_2^{21})' - \epsilon(A_2^{12}, A_2^{22})') \geq 0$. Add $A_1^{-1}(Y - A_2X_2) \geq 0$, non negative by assumption, and rearrange terms to get $A_1^{-1}(Y - A_2X_2 - \epsilon(A_2^{11}, A_2^{21})') + A_1^{-1}(Y - A_2X_2 - \epsilon(A_2^{12}, A_2^{22})') \geq 0$ which cannot be the sum of two negative quantities. QED.

Similar derivations show that whenever the joint support has positive probability at $A_1^{-1}(Y - A_2\,(X_2 - (\epsilon, \epsilon)')\,)$ then it also possible for the chain to get there either through $A_1^{-1}(Y - A_2\,(X_2 - (0, \epsilon)')\,)$ or through $A_1^{-1}(Y - A_2\,(X_2 - (\epsilon, 0)')\,)$; and that the same condition holds as we consider the moves to the points $A_1^{-1}(Y - A_2\,(X_2 + (\epsilon, -\epsilon)')\,)$ and $A_1^{-1}(Y - A_2\,(X_2 + (-\epsilon, \epsilon)')\,)$.

General case: the proof is exactly the same as in the 2D case, but more tedious. Now X_2 and $(\epsilon, ..., \epsilon)'$ are $\kappa - \ell = n$-dimensional. Assume a $A_1^{-1}(Y - A_2X_2) \geq 0$ and $A_1^{-1}(Y -$

$A_2\,(X_2+(\epsilon,...,\epsilon)')\,)\geq 0$ hold true. Rewrite $A_1^{-1}(Y-A_2\,(X_2+(\epsilon,...,\epsilon)')\,)$ as $A_1^{-1}(Y-A_2\,X_2-\epsilon(A_2^{11},A_2^{21},...,A_2^{n1})'-...-\epsilon(A_2^{1n},A_2^{2n},...,A_2^{nn})')\,)\geq 0$. Add $(n-1)\times A_1^{-1}(Y-A_2X_2)\geq 0$, non negative by assumption, and rearrange terms to get $A_1^{-1}(Y-A_2X_2-\epsilon(A_2^{11},A_2^{21},...,A_2^{n1})')+...+A_1^{-1}(Y-A_2X_2-\epsilon(A_2^{1n},A_2^{2n},...,A_2^{nn})')\geq 0$, which cannot be the sum of n negative terms. QED.

Again similar derivations show that condition A.1 holds as we consider moves to other points $X_2+(\pm\epsilon,...,\pm\epsilon)'$. □

Appendix B

Full Conditionals for the Gibbs Sampler

Say $\Theta = (\lambda_1, ..., \lambda_\kappa, \phi)'$ then $P(X, \Theta) = \prod_{i=1}^{\kappa} P(X_i|\Theta)\, P(\Theta) = \prod_{i=1}^{\kappa} P(X_i|\lambda_i, \phi)\, P(\lambda_i)\, P(\phi)$. We want $\lambda_i \in (0, \infty)$ and $\phi \in (0, \infty)$. As an example, assume priors for λ_i and $1/\phi$ proportional to a constant, and $\tau = 1$. Then, noticing that $P(\Theta|X, Y) = P(\Theta|X)\, I_{\{A^{-1}Y\}}(X)$, the following full conditional distributions can be derived.

$$\begin{aligned} P(\lambda_i|X, Y) &\propto \prod P(X_i|\lambda_i, \phi) \cdot P(\lambda_i) \\ &\propto \frac{1}{\lambda_i^{\frac{Ik}{2}}} e^{-\frac{1}{2\phi}\left(\frac{\log(X_i)-\lambda_i}{\lambda_i^k}\right)^2} \end{aligned}$$

$$\begin{aligned} P(\phi|X, Y) &\propto P(X_i|\lambda_i, \phi) \cdot P(\phi) \\ &\propto \frac{1}{\phi^{\frac{I}{2}+2}}\, e^{-\frac{1}{2\phi}\sum_i\left(\frac{\log(X_i)-\lambda_i}{\lambda_i^k}\right)^2}. \end{aligned}$$

In order to compute $P(X|Y,\Theta)$ we use the fact in Appendix A to conclude that $P(X|Y,\Theta) = P(X_2|Y,\Theta) \times \times P(X_1(X_2)|Y,\Theta)$; hence for $X_i \in X_2$ and $X_j \in X_1$ it follows:

$$\begin{aligned} P(X_i|X_{(-i)},Y,\Theta) &\propto P(X_i|\Theta)\cdot P(X_1|Y,\Theta) \\ &= log\text{-}Normal_{X_i}(\lambda_i,\phi\lambda_i)\cdot\textstyle\prod_j log\text{-}Normal_{X_j}(\lambda_j,\phi\lambda_j)\, I_{\{A^{-1}Y\}}(X_j) \end{aligned} \tag{B.1}$$

In the analysis, we explored the various posterior distributions using the Gibbs sampler with Metropolis steps. In order to sample from $P(X_i|Y,\Theta)$ and $P(\lambda_i|X,Y)$, we used χ^2 and Uniform proposals, improper priors on the lambdas (all proportional to a constant), and several flavors for the improper prior on ϕ (proportional to a constant, to $\frac{1}{\phi}$, and oto $\frac{1}{\phi^2}$).

Appendix C

Compendium of Network Models

Models for graphs of various types are scattered across research in the social, physical, mathematical, statistical, and computing sciences. In this review of the literature, I emphasize those statistical models that attempt to express the dependencies between objects in the system, in some sense.

C.1 Static Graphs

The works in this section take as input measurements about objects that start from a network as given.

Exponential Random Graph Models. Under the assumption that two possible social ties are independent only if a common actor is involved in both[1] Frank and Strauss (1986) devised the following characterization for the probability distribution of undirected Markov graphs.

$$P_\theta \{Y = y\} = \exp\left(\sum_{k=1}^{n-1} \theta_k S_k(y) + \tau T(y) + \psi(\theta, \tau) \right) \quad y \in \mathcal{Y}, \tag{B.1}$$

[1] This is the intuitive definition of Markov property for spatial processes on a lattice in Besag (1974).

where the statistics S_k and T count specific structures, such as edges, triangles, and k-stars, $\{\theta_k\} = \theta$ and τ are the parameters, and $\psi(\theta, \tau)$ is the normalizing constant. Frank and Strauss (1986) worked mainly with the three parameter models, where $\theta_3, \ldots, \theta_{n-1} = 0$. They proposed the pseudo-likelihood estimation method to estimate the complete vector of parameters by maximizing the following pseudo-likelihood function.

$$\ell(\theta) = \sum_{i<j} \log \left(P_\theta \left\{ Y_{ij} = y_{ij} \,|\, Y_{uv} = y_{uv} \text{ for all } u < v, (u,v) \neq (i,j) \right\} \right) . \tag{B.2}$$

Wasserman and Pattison (1996) proposed the current formulation of Exponential Random Graph Models (ERGM), also referred to as p^* models, as a generalization of the Markov graphs of Frank and Strauss. For both directed and undirected graphs, they maintain a similar characterization of the probabilities where the statistics S_k and T are substituted for arbitrary statistics U. This leads to the probability functions of the form

$$P_\theta \left\{ Y = y \right\} = \exp \left(\theta^\top u(y) - \psi(\theta) \right) . \tag{B.3}$$

More recently, Snijders et al. (2004) have proposed a variant of these models where the major problem of double-counting[2] is mitigated, but not overcome. Hunter and Handcock (2004) propose an alternative estimation scheme that corrects parameter estimates for double-counting. This estimation procedure can be used for models based on distributions in the curved exponential family. Park and Newman (2004) formally characterize sensitivity issues.

Remark A. It is possible to express the current formulation of exponential random graphs using the formalism of undirected graphical models, let us write the likelihood of an arbitrary undirected

[2]The statistics $S_i(y)$ count graph structures. Although they are not independent, i.e., they count overlapping sets of edges, they are assumed independent in the pseudo-likelihood. Ignoring the correlations is a bad idea, and causes extreme sensitivity of the predicted number of edges to small changes in the value of certain parameters.

graph.

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{\prod_{c\in\mathcal{C}} \psi(\boldsymbol{x}_c|\boldsymbol{\theta}_c)}{z}, \tag{B.4}$$

where $\boldsymbol{x}_c$ denotes the nodes in clique c, $\boldsymbol{\theta}_c$ denotes the corresponding set of parameters, ψ are non-normalized potentials over the cliques, and $z = \sum_{c\in\mathcal{C}} \prod_{c\in\mathcal{C}} \psi(\boldsymbol{x}_c|\boldsymbol{\theta}_c)$ is the normalization constant. If the likelihood is in the exponential family, we can write:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \exp\left\{ \boldsymbol{\theta}^\top u(\boldsymbol{x}) - \log z \right\}.$$

Remark B. Models in this family are not "generative models" in that no assumptions are present to explain how the sufficient statistics are generated. However, it is possible to posit a generative model that includes exponential random graph models, or any other conditional model, as part of the emission model (Airoldi et al., 2006b).

Latent Variable Models. The notions of equivalence, structural equivalence, and blocks are introduced by Lorrain and White (1971) and further explored by many, notably by Faust (1988). A comprehensive treatment of models that use blocks to express the complexity of the data is given in Doreian et al. (2004). A summary of models and notions relevant for social networks developed in the social sciences can be found in Wasserman and Faust (1994).

Stochastic block models, the probabilistic treatment of blocks, have appeared early in the statistical sciences (Holland and Leinhardt, 1975) and widely studied (Fienberg and Wasserman, 1981; Fienberg et al., 1981; Holland et al., 1983; Fienberg et al., 1985; Wang and Wong, 1987; Wasserman and Anderson, 1987; Anderson et al., 1992) and recently rediscovered (Snijders and Nowicki, 1997; Nowicki and Snijders, 2001), including non-parametric treatment of the number of blocks (Kemp et al., 2004), and integration with non-relational information to infer the blocks (Wang et al., 2005). A general stochastic block model of mixed-membership has been recently proposed (Airoldi et al.,

2005b, 2006c), along with a framework to integrate external information of different types (Airoldi et al., 2006b), that relaxes the historical assumption of single-membership of objects to blocks, and estimates block-to-block connectivity patterns in a Bayesian fashion.

Remark C. A general framework for integration of a different nature is described by Carley (2002).

An alternative approach latent space models, where observed interactions are projected on a latent space through a generalized linear model (Hoff et al., 2002; Hoff, 2003b,a). Hoff et al. (2002) use MCMC to infer latent space positions, treated as hyper-parameters. Hoff (2003b) specifies a Gaussian prior over the latent space, thus giving to the model fully generative flavor, with the goal of modeling reciprocity.

Remark D. It is possible to posit a generative model that includes generalized linear models as part of the emission model (Airoldi et al., 2006b). The connection between stochastic block models (SBM) and latent space models (LSM) is more subtle, though.

Both SBM and LSM seek to define a conditional probability distribution for relations $\{y_{nm}\}$ among actors in a way that reflects some latent semantics (i.e., roles, topics, functions, etc.) of the actors. Let Z_n denote a latent variable capturing the latent semantic representation of the actor n, the SBM usually defines a generative model $y_{nm} \sim f(\cdot|Z_n, Z_m)$, of which the Z's typically act as indicators of context-dependent edge generating processes. On the other hand, an LSM maps the observed relation y_{nm} to some latent semantic differences between the two actors via a regression function, of which the Z's typically represent the projections of the actors onto some latent metric space where their differences can be measured via a Euclidian metric. Specifically, in LSM the Zs are multivariate/continuous, e.g., could be drawn from a mv-normal, and their realizations indicate the position of actors in the latent space. In SBM the Zs are multivariate/discrete, e.g., could be drawn from a $multinomial(\theta)$, and their realizations indicate which group an actor belongs to, for each observed interaction. In other words, the dimensionality of the Zs in SBM reflects how

many latent groups to be captured in a domain, whereas in LSM the dimensionality of the Zs does not have an explicit interpretation in terms of groups. In fact, in LSM we need to run a clustering procedure (e.g., k-means) in the latent space where the actors are projected to, in order to decide how many groups there are. Thus, the two types of network models are different: SBM focuses on latent membership of each actor and underlines the importance of modeling the "grouping" of actors, whereas LSM focuses on latent distances and therefore stress more on modeling proper projections of actors into a latent manifold. Hoff's formulation of LSM is not a soft version of SBM. As a results, SBM and LSM have some orthogonal advantages in modeling network data.

Remark E. Connections have been highlighted to MDS and other linear methods (Breiger et al., 1975), to unsupervised learning, e.g., PCA, FA, (Ghahramani, 2004), and to matrix factorization (Ding, 2005; Xing and Jordan, 2003).

Spectral Methods. Research on by Gaussian unit ensembles provides a probabilistic connection to spectral decompositions (Metha, 2004). In the computer science literature, there is a stream of works in this area well summarized by Saul (2005), who discusses comparison to PCA, MDS and other linear methods. Briefly, isomap (Tenenbaum et al., 2000), local linear embedding (Roweis and Saul, 2000), laplacian eigenmaps (Belkin and Niyogi, 2002), Hessian eigenmaps (Donoho and Grimes, 2003), maximum variance unfolding (Weinberger and Saul, 2004; Weinberger et al., 2004; Sun et al., 2006), conformal eigenmaps (Coifman et al., 2005a,b; Lafon and Lee, 2006) and its asymptotics (Nadler et al., 2005), and the recent reformulation of problems and solutions in terms of tensors (He et al., 2005).

Simple Models of Real-World Phenomena and their Mathematical Properties. Much of the research across communities concerns the study of real-world graphs and their properties with the aim of building toy models that capture such properties. For example, Newman and Park (2003) study transitivity and assortative mixing (i.e., positive correlation of degrees of adjacent vertices)

via group structure; Hoff (2003a) studies transitivity, reciprocity and balance; Barabasi (2005a) studies burst and heavy tails in human dynamics; Zheng et al. (2005) study the size of individuals' social networks and means of estimating them from a certain type of survey questions; and Ganesh et al. (2005) study the effects on epidemics of the topological properties of graphs.

Research originating in mathematics and physics posit simple algorithms for generating graphs that replicate observed properties, which are amenable to probabilistic analysis. Bollobás and Riordan (2003) review few of such algorithms for popular graph types (Barabasi and Albert, 1999; Kumar et al., 2000; Cooper and Frieze, 2003), and present an extended analysis of the "LCD" model of Buckley and Osthus (2004). Other notable analytical investigations concern sampling, and asymptotic results. Park and Newman (2004, 2005) give analytic solutions for the 2-star network and for clustered networks; Milo et al. (2004c) analyze sampling algorithms; Kleinberg and Kleinberg (2005) describe asymptotics of isomorphism and embedding; Stumpf et al. (2005) find that sub-samples of scale-free graphs are not scale-free, and present a way to study properties of a sub-sample based on moment generating functions; Flaxman et al. (2005) describe the behavior of high degree vertices and eigenvalues in scale-free graphs; Chung and Lu (2003) characterize average distances given expected degrees; and Caldarelli et al. (2004) study the formation of cycles. A series of works is concerned with models and methods to find "statistically significant" motifs, i.e., recurring edge patterns over sets of difference nodes (Berg and Lassig, 2004; Shen-Orr et al., 2002; Milo et al., 2002; Artzy-Randrup et al., 2004; Milo et al., 2004a; Kashtan et al., 2004; Milo et al., 2004b). Newman (2003b) portrays networks as mixtures of patterns; and Vázquez et al. (2004) present the only investigation to date of how global patterns may arise from the composition of local ones. Few comprehensive reviews are available, which summarize many of these findings (Barabasi et al., 1999; Albert and Barabasi, 2002; Dorogovtsev and Mendes, 2002; Newman, 2003a; Amaral and Ottino, 2004).

A notion that recently captured the attention of funding agencies and high profile journals is

that of "topology types". Airoldi and Carley (2005) present a review and a critique of such notion. They survey generative algorithms for random graphs (Erdös and Rényi, 1960), Poisson graphs and others that lead to heavy tails for the corresponding degree distributions (Simon, 1955; Bollobás, 1985, 2001; Barabasi, 2005a), scale-free graphs (Faloutsos et al., 1999; Barabasi and Albert, 1999; Huberman and Adamic, 1999; Adamic and Huberman, 2000; Barabasi et al., 2000; Barabasi and Bonabeau, 2003), small-world graphs (Milgram, 1967; Watts and Strogatz, 1998; Kleinberg, 1999a; Amaral et al., 2000; Liben-Nowell et al., 2005), core-periphery graphs (Borgatti and Everett, 1999), and cellular graphs and networks (Frantz and Carley, 2005a; Airoldi and Carley, 2006). Several of these topology types are presented in heuristic terms, vaguely consistent across communities[3]. Airoldi and Carley (2005) show that the slight differences in the sampling algorithms, which generate topologies that adhere to the heuristic requirements of a specific type, are not stable in terms of the topological properties of the graphs they lead to. That is, slight differences in the operational definitions for the same topology type lead to separable graphs in terms of the set of common metrics used by practitioners in the various communities[4]. A different set of concerns is explored in "robustness" studies, which measure the stability of topological properties of graphs and networks of specific types to disruption and other stress situations (Borgatti et al., 2005; Frantz and Carley, 2005b). These works are simulation studies that approach the sub-sampling issues discussed above from another perspective.

Remark F. Alternatively, being able to embed the various topology type in a smooth parametric continuum (e.g., Erdös random, to small-world, to ring lattice; see Watts and Strogatz, 1998) would help understanding the boundaries. Unfortunately, also this strategy is not practical. There is a potpourri of necessary conditions that have to be satisfied by such a smooth parametric

[3]A notable survey is that of Mitzenmacher (2004), who presents a brief history of power-laws and lognormal distributions, and discusses some of their connections from a generative perspective. Newman (2005) discusses the connections among of power-laws, the Pareto distribution, and Zipf's law. Airoldi and Shalizi (2006) present a clear analytical overview of these connections.

[4]Few works survey network metrics and visualization tools; notables are Carley and Reminga (2004) and (Frank, 2000).

continuum, which appear in the heuristic definitions of the various topology types, e.g., the same degree distribution for all nodes or not, or shortest path as the only notion of distance, or shortest path and metric embedding. Although it is possible to posit a generative process that satisfies all the necessary conditions, such a generative process relies on a non-smooth parametric continuum. Specifically, we would need to introduce in such a process a discrete parameter that controls the number of different "probabilistic treatments" for the nodes, e.g., the number of degree distributions. The problem is that, on one hand, the value of such a discrete parameter is difficult to estimate. On the other hand, its correct estimation is fundamental in correctly assigning the topology type. Ultimately, the diversity in the notions of topology types translates into the hardness of the estimation task, upon the success of which depends our ability to discriminate among types.

To summarize, we can organize the various works according to few aspects: (a) the notion(s) of distance between pairs of nodes that are needed; (b) the use, or not, of the descriptive statistics, as well as their nature, i.e., local versus global; (c) the existence of dependence constraints among neighborhoods; (d) the focus on node patterns (groups) versus edge patterns (motifs), where we do no distinguish similar edge configurations among different sets of nodes. These aspects have to be crossed with the nature of the models: (i) "generative" models and algorithms, both probabilistic and deterministic; (ii) models and algorithms that contain "generative" ideas, both probabilistic and deterministic; (iii) other models and algorithms.

Problems. More works have introduced methodological innovations in the context of specific problems. Notable research in this sense concerned how to find communities in networks (Girvan and Newman, 2002; Newman, 2004a,b), and in bipartite graphs (Mishra et al., 2004). To this extent, Doreian et al. (2004) summarize relevant works in the social sciences and develop a theory of generalized block models. A cluster of research is about link-mining (Domingos, 2003; Jensen, 1999; Getoor, 2003; Getoor and Diehl, 2005), graph mining (Chakrabarti, 2005), link prediction (Getoor et al., 2002; Liben-Nowell and Kleinberg, 2003; Goldenberg and Moore, 2004), and link ranking. (Brin and Page,

1998; Kleinberg, 1999b; Cohen et al., 1999; Ng et al., 2001). Other notable works are concerned with the information flow within a network; the emergence of deadlines (Papadimitriu and Servan-Schreiber, 1999), the dynamics of information (Kleinberg, 2001), the dynamics of information exchange (Dodds et al., 2003), how to maximize influence spread (Kempe et al., 2003), decentralized information processing (Van Zandt, 1997), and decentralized search (Kleinberg, 2000, 2004). A practical set of concerns inspired methods for entity disambiguation (Malin et al., 2005), and classification of relational data (Macskassy and Provost, 2005). Solan et al. (2005) propose a model to learn grammar-like rules in natural languages.

Empirical Studies. Another portion of research concerns findings that influenced the development of theoretical aspects. Notable empirical studies include the web (Faloutsos et al., 1999; Albert et al., 1999; Kleinberg and Lawrence, 2001), air traffic (Guimera et al., 2005), the creative enterprise (Barabasi, 2005b), scientific collaborations (Newman, 2001), metabolic networks (Guimera and Amaral 2005), decentralized search in email network (Adamic and Adar, 2005), transcriptional regulatory network, (Balazsi et al., 2005), words (Steyvers and Tenenbaum, 2005), the organization within the cell (Barabasi and Oltvai, 2004), politics (Porter et al., 2005), complex brain networks (Sporns et al., 2004), and more (Newman et al., 2002).

C.2 Dynamics and Evolution

Most existent works focus on static networks, however, there are few that consider methodology to deal with dynamics and evolution. Notables are the stream of works on cellular automata (Ilachinski, 2001), the early works on diffusion (Coleman et al., 1957), the treatment of dynamics with Markov-chains Monte-Carlo (Wasserman, 1980), dynamic random fields on undirected graph (Shalizi, 2003), link-copying processes (Kleinberg et al., 1999; Leskovec et al., 2005b,a), cascad-

ing behaviors (Watts, 2002), network tomography and latent allocation (Airoldi and Faloutsos, 2004), dynamics in the social space (Banks et al., 2005), and models that attempt at replicating real-world phenomena such as opinion formation (Wu and Huberman, 2005), and evolution (Doreian and Stokman, 1997).

Empirical Studies. Very few studies exist to guide theoretical developments in a dynamic setting. Few notables are communication networks (Airoldi, 2003; Airoldi and Faloutsos, 2004), email networks (Kossinets and Watts, 2006), nucleic acid chain dynamics (Sales-Pardo et al., 2005), and scientific collaborations (Barabasi et al., 2002).

C.3 Building Graphs from Data

The works in this section share the intuition that measurements about objects are inherently noisy; the various authors attempt to model the uncertainty associated with the measurements in order to make decisions whether two objects are related or not, and create a graph. A popular approach is that of associating a random variable with each "object", e.g., Bernoulli, define the process through which "observations" relate to binary outcomes, and estimate the parameter of a Bayesian network (Heckerman, 1999) that describes the observations best, through dependencies among objects. The estimated Bayesian network provides a probabilistic model for the observed co-occurrences that can be used to predict missing links, or to assess the likelihood of existing ones, (Getoor et al., 2002; Friedman and Koller, 2003; Heckerman et al., 2004; Goldenberg and Moore, 2004; Teyssier and Koller, 2005). Important applications based on variations this approach have been used for building recommender systems (Breese et al., 1998), social networks (Breiger, 2003), and complex cellular networks (Friedman, 2004; Segal et al., 2005).

C.4 Inadequacies of the Current Research

There are several dimensions that are relevant to statistical analyses of graphs and networks. Unfortunately no single approach develops, or at least allows for, all of them

Dimensions of interest are: (a) a "proper" likelihood function; (b) the fully generative nature of the model; (c) replicability of interesting properties at both global and local level; (d) the focus on edges or nodes; (e) notions of distance and embedding of graphs in a metric space; (f) identifiability issues that need be explicitly identified; (g) hierarchical relations between dyads, triangles, k-stars, k-triangles and other basic structural (connected) components that are used to summarize and characterize an observed graph; (h) dependencies among relevant quantities, i.e., the sufficient statistics, corresponding to a decomposition of the observed graph into cliques or other structures of interest need be identified; (i) goodness of fit must be assessed—current models tend to over-fit observed graphs and can not be easily extended as the observed network grows; (j) the possibility of integrating data on different object types; and other dimensions.

With respect to this last dimension, i.e., integration of multiple object types and data about them, most existent work tend to concern or assume specific types of data representations, e.g., temporal and sequential data in attribute space, or relational data represented by graphs or networks. We view learning problems along this line as "type-specific-learning" problems. Typically, one can develop solutions to type-specific-learning problems by devising novel domain- and data-specific models and algorithms that leverage domain knowledge and semantics of interest for particular applications. Integrating heterogeneous data types under a unified model remains a challenge, however, especially for complex graphs that are simultaneously described by intrinsically different types of characteristics, such as features in attribute space and links in relational space (Airoldi et al., 2006b).

As we discussed in the previous section, there is a wide range of research questions that an

elegant solution to the issues above may help us answer. It is useful to keep those questions in mind in order to guide our technical choices.

Bibliography

L. A. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.

L. A. Adamic and B. A. Huberman. Comment on “Power-law distribution of the world wide web”. *Science*, 287:2115a, 2000.

E. M. Airoldi. Advances in network tomography. Technical Report CMU-CALD-03-101, Carnegie Mellon University, 2003.

E. M. Airoldi. Sampling algorithms for pure network topologies. Technical Report CMU-ISRI-05-111, School of Computer Science, Carnegie Mellon University, 2005.

E. M. Airoldi. Hierarchical bayesian mixture models of graphs and networks: A probabilistic approach to complexity via motifs, dynamics & integration. PhD thesis proposal, February 2006.

E. M. Airoldi. Comment on “Model-based clustering for social networks”. *Journal of the Royal Statistical Society, Series A*, 170, 2007.

E. M. Airoldi and K. M. Carley. Sampling algorithms for pure network topologies: Stability and separability of metric embeddings. *ACM SIGKDD Explorations*, 7(2):13–22, 2005.

E. M. Airoldi and K. M. Carley. The emergence of cellular structures. Manuscript, January 2006.

E. M. Airoldi and C. Faloutsos. Recovering latent time-series from their observed sums: network tomography with particle filters. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 10, pages 30–39, 2004.

E. M. Airoldi and X. Lin. Sparse factor analysis with application to microarrray data. Manuscript, June 2006.

E. M. Airoldi and C. R. Shalizi. The analytical nexus from Zipf's law to lognormal distributions. Manuscript, July 2006.

E. M. Airoldi and E. P. Xing. Non-observable birth-death processes. Manuscript, 2006a.

E. M. Airoldi and E. P. Xing. Bayesian analysis of graphs via excahngeable-edge models. Manuscript., June 2006b.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic block models for relational data with application to protein-protein interactions. Manuscript, November 2005a.

E. M. Airoldi, D. M. Blei, E. P. Xing, and S. E. Fienberg. A latent mixed-membership model for relational data. In *Workshop on Link Discovery: Issues, Approaches and Applications*, 2005b. In conjunction with the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

E. M. Airoldi, W. W. Cohen, and S. E. Fienberg. Bayesian models for frequent terms in text. In *Proceedings of the Classification Society of North America and INTERFACE Annual Meetings*, 2005c.

E. M. Airoldi, E. P. Xing, and C. Faloutsos. Super-resolution models for dynamic network analysis with application to network tomography. Manuscript, October 2005d.

E. M. Airoldi, A. G. Anderson, S. E. Fienberg, and K. K. Skinner. Who wrote Ronald Reagan radio addresses? *Bayesian Analysis*, 1(2):289–320, 2006a.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Latent mixed-membership allocation models of relational and multivariate attribute data. Manuscript, 2006b.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Admixtures of latent blocks with application to protein interaction networks. Manuscript under review, January 2006c.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Stochastic block models of mixed membership. Manuscript under review, June 2006d.

E. M. Airoldi, S. E. Fienberg, C. Joutard, and T. M. Love. Discovering latent patterns with hierarchical Bayesian mixed-membership models and the issue of model choice. Technical Report CMU-ML-06-101, School of Computer Science, Carnegie Mellon University, April 2006e.

E. M. Airoldi, S. E. Fienberg, and E. P. Xing. Latent aspects analysis for gene expression data. Manuscript, January 2006f.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, A. Goldenberg, E. P. Xing, and A. X. Zheng, editors. *Statistical Network Analysis: Models, Issues and New Directions*. Lecture Notes in Computer Science. Springer-Verlag, 2007a.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Combining stochastic block models and mixed membership for statistical network analysis. In E. M. Airoldi, D. M. Blei, S. E. Fienberg, A. Goldenberg, E. P. Xing, and A. X. Zheng, editors, *Statistical Network Analysis: Models, Issues and New Directions*, Lecture Notes in Computer Science. Springer-Verlag, 2007b. Forthcoming.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N.

Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, 1973.

R. Albert and A. L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(47), 2002.

R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world wide web. *Nature*, 401:130, 1999.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland, 4th edition, 2002.

L. A. N. Amaral and J. M. Ottino. Complex networks. *European Physics Journal B*, 38:147–162, 2004.

L. A. N. Amaral, A. Scala, M. Barthélémi, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97:11149–11152, 2000.

C. J. Anderson, S. Wasserman, and K. Faust. Building stochastic blockmodels. *Social Networks*, 14:137–161, 1992.

T. W. Anderson. R. A. Fisher and multivariate analysis. *Statistical Science*, 11(1):20–34, 1996.

C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone. Comment on "Network motifs: Simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". *Science*, 305:1107c, 2004.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubinand, and G. Sherlock. Gene ontology:

Tool for the unification of biology. The gene ontology consortium. *Nature Genetics*, 25(1): 25–29, 2000.

G. Balazsi, A. L. Barabasi, and Z. N. Oltvai. Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli. *Proceedings of the National Academies of Science*, 102(22):7841–7846, 2005.

D. Banks and K. M. Carley. Models for network evolution. *Journal of Mathematical Sociology*, 21:173–196, 1996.

H. T. Banks, A. F. Karr, H. K. Nguyen, and J. R. Samuels, Jr. Sensitivity to noise variance in a social network dynamics model. Technical Report 2005-10, Statistical and Applied Mathematical Sciences Institute, 2005.

A. L. Barabasi. The origins and heavy tails in human dynamics. *Nature*, 435:207–211, 2005a.

A. L. Barabasi. Network theory—the emergence of the creative enterprise. *Science*, 308:639–641, 2005b.

A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

A. L. Barabasi and E. Bonabeau. Scale-free networks. *Scientific American*, pages 50–59, May 2003.

A. L. Barabasi and Z. N. Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

A. L. Barabasi, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physyca A*, 272:173–187, 1999.

A. L. Barabasi, R. Albert, H. Jeong, and G. Bianconi. Response to power-law distribution of the world wide web. *Science*, 287:2115a, 2000.

A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientifc collaborations. *Physica A*, 311:590–614, 2002.

K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

M. J. Beal and Z. Ghahramani. The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics*, volume 7, pages 453–464. Oxford University Press, 2003.

M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

J. Berg and M. Lassig. Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences*, 101:14689–14694, 2004.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 1974.

N. Bhardwaj and H. Lu. Correlation between gene expression profiles and protein–protein interactions within and across genomes. *Bioinformatics*, 21(11):2730–2738, 2005.

S. Blackshaw, S. Harpavat, J. Trimarchi, L. Cai, H Huang, W. P. Kuo, R. E. Fraioli, S. H. Cho, R. Yung, and E. Asch. Genomic analysis of mouse retinal development. *PLoS Biology*, 2004.

D. M. Blei and S. E. Fienberg. Comment on "Model-based clustering for social networks". *Journal of the Royal Statistical Society, Series A*, 170, 2007.

D. M. Blei and J. D. Lafferty. Dynamic topic models. In *International Conference on Machine Learning*, volume 23, pages 113–120, 2006.

D. M. Blei, A. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

B. Bollobás. *Random Graphs*. Academic Press, London, 1985.

B. Bollobás. *Random Graphs*. Academic Press, New York, 2nd edition, 2001.

B. Bollobás and O. Riordan. Mathematical results on scale-free graphs. In S. Bornholdt and H. Schuster, editors, *Handbook of graphs and networks*, pages 1–34. Wiley-VCH, 2003.

S. P. Borgatti and M. G. Everett. Models of core / periphery structures. *Social Networks*, 21: 375–395, 1999.

S. P. Borgatti, K. M. Carley, and D. Krackhardt. Robustness of centrality measures under conditions of imperfect data. *Social Networks*, 2005. Forthcoming.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

J. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Uncertainty in Artificial Intelligence*, volume 14, 1998.

R. L. Breiger. Emergent themes in social network analysis: Results, challenges, opportunities. In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, 2003.

R. L. Breiger, S. A. Boorman, and P. Arabie. An algorithm for clustering relational data with applications to social network analysis and comparison to multidimensional scaling. *Journal of Mathematical Psychology*, 12:328–383, 1975.

S. Brin and L. Page. The anatomy of a large-scale hypertextual (Web) search engine. In *Proceedings of the International World Wide Web Conference*, volume 7, 1998.

P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, 1991.

P. G. Buckley and D. Osthus. Popularity based random graph models leading to a scale-free degree sequence. *Discrete Mathematics*, 282(1–3):53–68, 2004.

W. Buntine and A. Jakulin. Applying discrete PCA in data analysis. In *Uncertainty in Artificial Intelligence*, 2004.

W. L. Buntine and A. Jakulin. Discrete components analysis. In C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection Techniques*. Springer-Verlag, 2006. to appear.

M.-L. G. Buot and D. St. P. Richards. Counting and locating the solutions of polynomial systems of maximum likelihood equations, I. *Journal of Symbolic Computation*, 41:234–244, 2006a.

M.-L. G. Buot and D. St. P. Richards. Counting and locating the solutions of polynomial systems of maximum likelihood equations, II: The Behrens-Fisher problem. Manuscirpt, 2006b.

L. Cai, H. Huang, S. Blackshaw, J. S. Liu, C. L. Cepko, and W. H. Wong. Clustering analysis of SAGE data using a Poisson approach. *Genome Biology*, 5(7):R51, 2004.

G. Caldarelli, R. Pastor-Satorras, and A. Vespignani. Structure of cycles and local ordering in complex networks. *European Physics Journal B*, 38:183—186, 2004.

J. Canny. GaP: A factor model for discrete data. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.

J. Cao, D. Davis, S. Van Der Viel, and B. Yu. Time-varying network tomography: router link data. *Journal of the American Statistical Association*, 95:1063–75, 2000.

J. Cao, D. Davis, S. Van Der Viel, B. Yu, and Z. Zu. A scalable method for estimating network traffic matrices from link counts. Technical report, Bell Labs, 2001.

K. M. Carley. Group stability: A socio-cognitive approach. *Advances in Group Processes*, 7:1–44, 1990.

K. M. Carley. A theory of group stability. *American sociological Review*, 56:331–354, 1991.

K. M. Carley. Smart agents and organizations of the future. In L. Lievrouw and S. Livingstone, editors, *The Handbook of New Media*, pages 206–220, 2002.

K. M. Carley and J. Reminga. ORA: Organizational Risk Analyzer, 2004. Available for download at http://www.casos.cs.cmu.edu/projects/ora/.

K. M. Carley, D. B. Fridsma, E. Casman, A. Yahja, N. Altman, L.-C. Chen, B. Kaminsky, and D. Nave. BioWar: Scalable agent-based model of bioattacks. *IEEE Transactions on Systems, Man, and Cybernetics—Part A*, 36(2):252–265, 2006.

B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, 2005.

D. Chakrabarti, S. Papadimitriou, D. Modha, and C. Faloutsos. Fully automatic cross-associations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 10, pages 79–88, 2004.

Deepayan Chakrabarti. *Tools for Large Graph Mining*. PhD thesis, School of Computer Science, Carnegie Mellon University, 2005.

F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Internet Mathematics*, 1:91–114, 2003.

William C. Cohen, Robert E. Schapire, and Yoram Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270, 1999.

D. Cohn and T. Hofmann. The missing link—A probabilistic model of document content and hypertext connectivity. In *Advances in Neural Information Processing Systems 13*, 2001.

R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102:7426–7431, 2005a.

R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods. *Proceedings of the National Academy of Sciences*, 102:7432–7437, 2005b.

J. Coleman, E. Katz, and H. Menzel. The diffusion of an innovation among physicians. *Sociometry*, 20(4):253–270, 1957.

C. Cooper and A. M. Frieze. A general model of web graphs. *Random Structures and Algorithms*, 22(3):311–335, 2003.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

C. Ding. Principal component analysis and matrix factorizations for learning. *International Conference on Machine Learning*. Tutorial, 2005.

P. S. Dodds, D. J. Watts, and C. F. Sabel. Information exchange and the robustness of organizational networks. *Proceedings of the National Academy of Sciences*, 100(21):12516–12521, 2003.

P. Domingos. Prospects and challenges for multirelational data mining. *SIGKDD Explorations*, 5 (1):80–83, 2003.

D. L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

P. Doreian and F. N. Stokman, editors. *Evolution of Social Networks*. Gordon and Breach Publishers, 1997.

P. Doreian, V. Batagelj, and A. Ferligoj. *Generalized Blockmodeling*. Cambridge University Press, 2004.

S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51:1079–1187, 2002.

A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 5:290–297, 1959.

P. Erdös and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

E. A. Erosheva. *Grade of membership and latent structure models with application to disability survey data*. PhD thesis, Carnegie Mellon University, Department of Statistics, 2002.

E. A. Erosheva and S. E. Fienberg. Bayesian mixed membership models for soft clustering and classification. In C. Weihs and W. Gaul, editors, *Classification—The Ubiquitous Challenge*, pages 11–26. Springer-Verlag, 2005.

E. A. Erosheva, S. E. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 97(22):11885–11892, 2004.

M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the ACM SIGCOMM Conference*, pages 251–261, 1999.

K. Faust. Comparison of methods for positional analysis: Structural and general equivalences. *Social Networks*, 10:313–341, 1988.

T. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1: 209–230, 1973.

S. E. Fienberg and S. K. Lee. On small world statistics. *Psychometrika*, 40(2):219–228, 1975.

S. E. Fienberg and S. Wasserman. Categorical data analysis of single sociometric relations. In S. Leinhardt, editor, *Sociological Methodology*, pages 156–192. San Francisco: Jossey-Bass, 1981.

S. E. Fienberg, M. M. Meyer, and S. Wasserman. Analyzing data from multivariate directed graphs: An application to social networks. In *Interpreting Multivariate Data*, pages 289–306. New York: Wiley, 1981.

S. E. Fienberg, M. M. Meyer, and S. Wasserman. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80:51–67, 1985.

R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

R. A. Fisher. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

R. A. Fisher. The use of multiple treatments in taxonomic problems. *Annals of Eugenics*, 7: 179–188, 1936.

R. A. Fisher. The statistical utilization of multiple measurements. *Annals of Eugenics*, 8:376–386, 1938.

A. Flaxman, A. Frieze, and Trevor Fenner. High degree vertices and eigenvalues in the preferential attachment graph. *Internet Mathematics*, 2(1), 2005.

O. Frank. Structural plots of multivariate binary data. *Journal of Social Structure*, 1(4):1–19, 2000.

O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81: 832–842, 1986.

T. Frantz and K. M. Carley. A formal characterization of cellular networks. Technical Report CMU-ISRI-05-109, School of Computer Science, Canregie Mellon University, 2005a.

T. Frantz and K. M. Carley. Relating network topology to the robustness of centrality measures. Technical Report CMU-ISRI-05-117, School of Computer Science, Canregie Mellon University, 2005b.

N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303:799–805, 2004.

N. Friedman and D. Koller. Being bayesian about bayesian network structure: A bayesian approach to structure discovery in bayesian networks. *Machine Learning*, 50(1–2):95–125, 2003.

A. Ganesh, L Massoulie, and D. Towsley. The effect of network topology on the spread of epidemics. In *Proceedings of the 24th IEEE INFOCOM Annual Conference*, 2005.

A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, and et. al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.

A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman & Hall, London, 1995.

L. Getoor. Link mining: A new data mining challenge. *SIGKDD Explorations*, 5(1):84–89, 2003.

L. Getoor and C. Diehl. Link mining: A survey. *ACM SIGKDD Explorations*, 7(2):3–12, 2005.

L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models with link uncertainty. *Journal of Machine Learning Research*, 2002.

Z. Ghahramani. Unsupervised learning. In O. Bousquet, G. Raetsch, and U. von Luxburg, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 72–112. Springer-Verlag, 2004.

E. N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30:1141–1144, 1959.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99:7821—7826, 2002.

A. Goldenberg and A. W. Moore. Tractable learning of large Bayes net structures from sparse data. In *Proceedings of the Intenational Conference on Machine Learning*, volume 21, 2004.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, 2004.

T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems*, volume 17, pages 537–544, 2005.

R. Guimera and L. A. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005.

R. Guimera, S. Mossa, A. Turtschi, and L. A. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102:7794—7799, 2005.

J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society, Series A*, 170:1–22, 2007.

S. Hanneke and E. P Xing. Discrete temporal models of social networks. In E. M. Airoldi, D. M. Blei, and S. E. Fienberg, editors, *Statistical Network Analysis: Models, Issues and New Directions*, Lecture Notes in Computer Science. Springer-Verlag, 2007.

P. C. Hansen. *Rank-deficient and discrete ill-posed problems: Numerical aspects of linear inversion*. SIAM, 1998. ISBN 0-89871-403-6.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

X. He, D. Cai, and P. Niyogi. Tensor subspace analysis. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

D. Heckerman. A tutorial on learning with Bayesian networks. In M. Jordan, editor, *Learning in Graphical Models*. MIT Press, 1999.

D. Heckerman, C. Meek, and D. Koller. Probabilistic models for relational data. Technical Report MSR-TR-2004-30, Microsoft Research, 2004.

Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, and K. Boutilier et. al. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–183, 2002.

P. D. Hoff. Random effects models for network data. In R. Breiger, K. Carley, and P. Pattison, editors, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, pages 303–312. National Academies Press, 2003a.

P. D. Hoff. Bilinear mixed effects models for dyadic data. Technical Report 32, University of Washigton, Seattle, 2003b.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97:1090–1098, 2002.

P. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: Some first steps. *Social Networks*, 5:109–137, 1983.

P. W. Holland and S. Leinhardt. Local structure in social networks. In D. Heise, editor, *Sociological Methodology*, pages 1–45. Jossey-Bass, 1975.

B. A. Huberman and L. A. Adamic. Growth dynamics of the world-wide web. *Nature*, 401:131, 1999.

D. Hunter and M. Handcock. Inference in curved exponential family models for networks. Technical Report TR0402, Department of Statistics, Penn State University, 2004.

A. Ilachinski. *Cellular Automata: A Discrete Universe*. World Scientific, 2001.

W. James and C. M. Stein. Estimation with quadratic loss. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379, 1961.

D. Jensen. Statistical challenges to inductive inference in linked data. In *Proceedings of the 17th International Workshop on Artificial Intelligence and Statistics*, 1999.

N. L. Johnson, S. Kotz, and A. W. Kemp. *Univariate Discrete Distributions*. John Wiley, 1992.

I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

J. B. Kadane, G. Shmueli, T. P. Minka, S. Borle, and P. Boatwright. Conjugate analysis of the Conway-Maxwell-Poisson distribution. *Bayesian Analysis*, 1(2):363–374, 2006.

A. F. Karr. *Point Processes and their Statistical Inference*. Marcel Dekker, 1991.

N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, 2004.

R. E. Kass and D. Steffey. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84(407):717–726, 1989.

C. Kemp, T. L. Griffiths, and J. B. Tenenbaum. Discovering latent classes in relational data. Technical Report AI Memo 2004-019, MIT, 2004.

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.

D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

J. Kleinberg. The small-world phenomenon: An algorithmic perspective. Technical Report 99-1776, Department of Computer Science, Cornell University, 1999a.

J. Kleinberg. Navigation in a small world. *Nature*, 845, 2000.

J. Kleinberg. Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems 14*, 2001.

J. Kleinberg. The small-world phenomenon and decentralized search. *SIAM News*, 37(3), 2004.

J. Kleinberg and S. Lawrence. The structure of the web. *Science*, 294:1849–1850, 2001.

J. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models and methods. In *Proceedings of the International Conference on Combinatorics and Computing*, pages 1–17, 1999.

J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5): 604–632, 1999b.

R. Kleinberg and J. Kleinberg. Isomorphism and embedding problems for infinite limits of scale-free graphs. In *Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms*, 2005.

L. Kontorovich, J. D. Lafferty, and D. M. Blei. Variational inference and learning for a unified model of syntax, semantics and morphology. Technical Report CMU-CALD-06-100, School of Computer Science, Canregie Mellon University, April 2006.

G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311: 88–90, 2006.

N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt. Global landscape of protein complexes in the yeast Saccharomyces Cerevisiae. *Nature*, 440 (7084):637–643, 2006.

R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Random graph models for the web graph. In *Annual Symposium on Foundations of Computer Science*, pages 57–65, 2000.

S. Lafon and A. B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006. Forthcoming.

S. Lafon, Y. Keller, and R. Coifman. Data fusion and multi-cue data matching by diffusion maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006. Forthcoming.

J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graphh generation and evolution, using Kronecker multiplication. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2005a.

J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005b.

D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the 12th Annuam ACM International Conference on Knowledge Management*, pages 556–559, 2003.

D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102:11623—11628, 2005.

F. Lorrain and H. C. White. Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, 1:49–80, 1971.

S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. Technical Report CeDER-04-08, Stern School of Business, New York University, 2005.

B. Malin, E. M. Airoldi, and K. M. Carley. A social network analysis model for name disambiguation in lists. *Journal of Computational and Mathematical Organization Theory*, 11(2):119–139, 2005.

K. G. Manton, M. A. Woodbury, and H. D. Tolley. *Statistical Applications Using Fuzzy Sets*. Wiley, 1994.

J. McAuliffe, D. Blei, and M. Jordan. Nonparametric empirical Bayes for the Dirichlet process mixture model nonparametric empirical bayes for the dirichlet process mixture model. *Statistics and Computing*, 2006. Forthcoming.

A. McCallum, X. Wang, and N. Mohanty. Joint group and topic discovery from relations and text. In *Statistical Network Analysis: Models, Issues and New Directions*, Lecture Notes in Computer Science. Springer-Verlag, 2007.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, second edition, 1989.

M. L. Metha. *Ranomd Matrices*. Elsevier, 2004.

H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, and et. al. Mips: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, 32:D41–44, 2004.

S. Milgram. The small world phenomenon. *Psychology Today*, 1(61), 1967.

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.

R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, and U. Alon. Response to comment on "Network motifs: Simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". *Science*, 305:1107d, 2004a.

R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303:1538–1542, 2004b.

R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequence. *ArXiv Condensed Matter e-prints*, 2004c.

T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Uncertainty in Artificial Intelligence*, 2002.

N. Mishra, D. Ron, and R. Swaminathan. A new conceptual clustering framework. *Machine Learning Journal*, 56(1–3):115–151, 2004.

M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

C. Morris. Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–65, 1983. With discussion.

F. Mosteller and D.L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964.

F. Mosteller and D.L. Wallace. *Applied Bayesian and Classical Inference: The Case of "The Federalist" Papers*. Springer-Verlag, 1984.

C. L. Myers, D. A. Barret, M. A. Hibbs, C. Huttenhower, and O. G. Troyanskaya. Finding function: An evaluation framework for functional genomics. *BMC Genomics*, 7(187), 2006.

B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98:404—409, 2001.

M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003a.

M. E. J. Newman. Mixing patterns in networks. *Physics Reviews E*, 67, 2003b.

M. E. J. Newman. Analysis of weighted networks. *Physics Reviews E*, 70, 2004a.

M. E. J. Newman. Detecting community structure in networks. *European Physics Journal B*, 38: 321–330, 2004b.

M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46: 323–351, 2005.

M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Physics Reviews E*, 68, 2003.

M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99:2566—2572, 2002.

A. Ng. Preventing "overfitting" of cross-validation data. In *Intenational Conference on Machine Learning*, volume 14, 1997.

A. Y. Ng, A. X. Zheng, and M. I. Jordan. Link analysis, eigenvectors, and stabiliy. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 17, 2001.

K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96:1077–1087, 2001.

C. Papadimitriu and E. Servan-Schreiber. The origins of the deadline: Optimizing communication in organizations. In *Complexity in Economics*, 1999.

J. Park and M. E. J. Newman. Solution of the 2-star model of a network. *Physics Reviews E*, 70, 2004.

J. Park and M. E. J. Newman. Solution for the properties of a clustered network. *Physics Reviews E*, 72, 2005.

D. Pelleg and A. W. Moore. X-means: Extending K-means with efficient estimation of the number of clusters. In *International Conference on Machine Learning*, volume 17, pages 727–734, 2000.

M. A. Porter, P. J. Mucha, M. E. J. Newman, and C. M. Warmbrand. A network analysis of committees in the united states house of representatives. *Proceedings of the National Academy of Sciences*, 102:7057–7062, 2005.

C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park. Scan statistics on enron graphs. *Computational and Mathematical Organization Theory*, 11(3):229–247, 2005.

J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. Genetic structure of human populations. *Science*, 298:2381–2385, 2002.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 22:2323–2326, 2000.

M. Sales-Pardo, R. Guimera, A. A. Moreira, J. Widom, and L. A. N. Amaral. Mesoscopic modeling for nucleic acid chain dynamics. *Physics Reviews E*, 71:1–13, 2005.

F. S. Sampson. *A Novitiate in a period of change: An experimental and case study of social relationships*. PhD thesis, Cornell University, 1968.

L. Saul. Spectral methods for dimensionality reduction. *Advances in Neural Information Processing Systems*. Tutorial, 2005.

Mark J. Schervish. *Theory of Statistics*. Springer, 1995.

G. Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

E. Segal, D. Pe'er, A. Regev, D. Koller, and N. Friedman. Learning module networks. *Journal of Machine Learning Research*, 6:503–556, 2005.

C. R. Shalizi. Optimal nonlinear prediction of random fields on networks. *Discrete Mathematics and Theoretical Computer Science*, AB(DMCS):11–30, 2003.

S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia Coli. *Nature Genetics*, 31:64–68, 2002.

Craig Shreiber. *Human and organizational risk modeling: Critical personnel and leadership in network organizations*. PhD thesis, School of Computer Science, Carnegie Mellon University, 2006.

H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

T. A. B. Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 2002.

T. A. B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14:75–100, 1997.

T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 2004.

Z. Solan, D. Horn, E. Ruppin, and S. Edelman. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, 102:11629–11634, 2005.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–639, 2002.

O. Sporns, D. R. Chialvo, M. Kaiser, and C. C. Higetag. Organization, development and function of complex brain networks. *Trends in Cognitive Science*, 8(9):418–425, 2004.

M. Steyvers and J. B. Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78, 2005.

M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences*, 102(12): 4221–4224, 2005.

J. Sun, S. Boyd, L. Xiao, and P. Diaconis. The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM Review*, 2006. Forthcoming.

B. Taskar, M. F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Neural Information Processing Systems 15*, 2003.

C. Tebaldi and M. West. Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, 93:557–576, 1998.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

M. Teyssier and D. Koller. Ordering-based search: A simple and effective algorithm for learning bayesian networks. In *Uncertainty in Artificial Intelligence*, volume 21, pages 584–590, 2005.

R. Tibshirani and G. Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.

O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in S. cerevisiae). *Proceedings of the National Academy of Sciences*, 100(19):10623–10628, 2003.

D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P. A. Pevzner. Identification of post-translational modifications via blind search of mass-spectra. *Nature Biotechnology*, 23:1562–1567, 2005.

T. Van Zandt. Decentralized information processing in the theory of organizations. In M. Sertel, editor, *Contemporary Economic Development Reviewed*, volume 4: The Enterprise and its Environment. MacMillan Press Ltd., 1997.

R. J. Vanderbei and J. Iannone. An em approach to od matrix estimation. Technical Report SOR 94-04, Princeton University, 1994.

Y. Vardi. Network tomography: estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, 91:365–377, 1996.

A. Vászquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. N. Oltvai, and A.-L. Barabási. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proceedings of the National Academy of Sciences*, 101:17940—17945, 2004.

V. E. Vesculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270:484–487, 1995.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley, 2003.

X. Wang, N. Mohanty, and A. K. McCallum. Group and topic discovery from relations and text. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

Y. J. Wang and G. Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82:8–19, 1987.

S. Wasserman. Analyzing social networks as stochastic processes. *Journal of the American Statistical Association*, 75:280—294, 1980.

S. Wasserman and C. J. Anderson. Stochastic a posteriori blockmodels: Construction and assessment. *Social Networks*, 9:1–36, 1987.

S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

S. Wasserman and P. Pattison. Logit models and logistic regression for social networks: I. an introduction to markov graphs and p^*. *Psychometrika*, 61:401–425, 1996.

S. Wasserman, G. Robins, and D. Steinley. Statistical models for networks: A brief review of some recent research. In E. M. Airoldi, D. M. Blei, S. E. Fienberg, A. Goldenberg, E. P. Xing, and A. X. Zheng, editors, *Statistical Network Analysis: Models, Issues and New Directions*, Lecture Notes in Computer Science. Springer-Verlag, 2007.

D. J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99:5766—5771, 2002.

D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393: 440–442, 1998.

K. Weinberger and L. Saul. Unsupervised learning of image manufolds by semidefinite programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

K. Weinberger, F. Sha, and L. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *International Conference on Machine Learning*, 2004.

M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, 1997.

C. Wiuf, M. Brameier, O. Hagberg, and M. P. H. Stumpf. A likelihood approach to analysis of network data. *Proceedings of the National Academy of Sciences*, 103:7566–7570, 2006.

M. A. Woodbury, J. Clive, and A. Garson. Mathematical typology: Grade of membership technique for obtaining disease definition. *Computational Biomedical Research*, 11(3):277–298, 1978.

F. Wu and B. A. Huberman. Social structure and opinion formation. Online Manuscript, 2005.

E. P. Xing. On topic evolution. Technical Report CMU-CALD-05-115, School of Computer Science, Canregie Mellon University, 2005a.

E. P. Xing. Dynamic non-parametric Bayesian models. Technical Report CMU-CALD-05-114, School of Computer Science, Canregie Mellon University, 2005b.

E. P. Xing and M. I Jordan. On semidefinite relaxation for normalized k-cut and connections to spectral clustering. Technical Report CSD-03-1265, Division of Computer Science, University of California at Berkeley, 2003.

E. P. Xing, M. I. Jordan, R. M. Karp, and S. Russell. A hierarchical Bayesian markovian model for motifs in biopolymer sequences. In *Advances in Neural Information Processing Systems*, volume 16, 2003a.

E. P. Xing, M. I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Uncertainty in Artificial Intelligence*, volume 19, 2003b.

E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russel. Distance metric learning with applications to clustering with side information. In *Advances in Neural Information Processing Systems*, volume 16, 2003c.

Y. Zhang, M. Roughan, C. Lund, and D. Donoho. An information-theoretic approach to traffic matrix estimation. In *Proceedings of SIGCOMM*, 2003.

T. Zheng, M. Salganik, and A. Gelman. How many people do you know in prison? *Journal of the American Statistical Association*, 2005. Forthcoming.