A study of Mixed-Membership Models for Complex Networks Analysis

Adrien Dulac

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : Mathématiques et Informatique

Arrêté ministériel : 25 mai 2016

Présentée par

Adrien DULAC

Thèse dirigée par **Eric GAUSSIER**, UGA
et codirigée par **Christine LARGERON**

préparée au sein du **Laboratoire d'Informatique de Grenoble**

dans l'**École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Etude des modèles à composition mixée pour l'analyse de réseaux complexes

A study of Mixed-Membership Models for Complex Network Analysis

Thèse soutenue publiquement le **17 décembre 2018**,
devant le jury composé de :

Monsieur ERIC GAUSSIER
PROFESSEUR, UNIVERSITE GRENOBLE ALPES, Directeur de thèse
Monsieur MOHAMED NADIF
PROFESSEUR, UNIVERSITE PARIS 5, Rapporteur
Monsieur FABRICE ROSSI
PROFESSEUR, UNIVERSITE PARIS 1, Rapporteur
Madame CHRISTINE LARGERON
PROFESSEUR, UNIVERSITE JEAN MONNET - SAINT-ETIENNE, Examinateur
Monsieur PHILIPPE LERAY
PROFESSEUR, UNIVERSITE DE NANTES, Président du jury
Madame ADELINE LECLERCQ SAMSON
PROFESSEUR, UNIVERSITE GRENOBLE ALPES, Examinateur

A Study of Mixed-Membersip Models for Complex Network Analysis

Adrien Dulac

A thesis presented for the degree of
Doctor of Science

Supervised by:
Professor Eric Gaussier
Professor Christine Largeron

UNIVERSITÉ
Grenoble
Alpes

University Grenoble Alpes, France
september 2018

Se demander si un ordinateur sait penser est aussi intéressant que se demander si un sous-marin sait nager.

— E. W. Dijkstra

Remerciements

Cette thèse est le fruit d'un travail de recherche de quatre années mené au sein du Laboratoire d'Informatique de Grenoble. Celui-ci constitue un formidable écosystème dynamique où se mélangent diverses disciplines scientifiques conduites par des personnalité plurielles et pleines d'humanité, auxquelles je tiens à exprimer toute ma gratitude.

En premier lieu, je remercie très sincèrement mes deux directeurs de thèse. Christine Largeron, pour son soutien, son enthousiasme et ses précieux conseils tout au long de mon cheminement, et aussi pour son accueil dans le Laboratoire Hubert Curien où nous avons partagé de nombreuses conversations passionnées. Il me tient à cœur de formuler mon plus grand respect à Eric Gaussier pour m'avoir guidé, conseillé et épaulé tout au long de cette thèse, avec une patience et une sagesse qui m'ont permis de m'épanouir et d'acquérir en rigueur. Encore un immense merci à tout les deux pour votre énergie et pour m'avoir accompagné sans relâche durant cette longue période.

Je remercie l'ensemble du jury de m'avoir fait l'honneur d'assister à ma soutenance et tout particulièrement Fabrice Rossi et Mohamed Nadiff pour avoir relu mon manuscrit et apporté vos précieux commentaires et remarques. Merci également à vous Léa, Suzie et Dominique pour votre relecture.

Je remercie et salue l'ensemble des membres de l'équipe AMA au cœur de laquelle j'ai évolué durant ma thèse et qui ont animé la vie scientifique de ce séjour. Je salue également mes collègues de bureau, Camille, Yagmur, Jacques, Clément, Georgios et Thibaut avec lesquels nous avons traversé de rudes épreuves, entre ascenseurs dotés de parole et monologuant, régulations thermiques extrêmes et contreforts cartonnés pour lutter contre les rayonnements solaires, dans une bonne humeur imperturbable.

Je remercie le Club Alpin Français de Grenoble pour m'avoir initié à l'alpinisme entre deux équations ainsi que la communauté des grimpeurs, les collocs' de Gières, de St Martin le Vinoux, et du 106, et tout ceux, nombreux, avec lesquels j'ai partagé d'essentiels épisodes de retour à la Nature et d'exploration du massif alpin, étapes indispensables à un équilibre mental et physique, favorisant la bonne conduite de mes travaux de recherche.

Enfin, je remercie ma Famille pour son soutien inconditionnel, et sans laquelle rien de tout cela n'aurait été possible.

le jeudi 13 décembre 2018

Abstract

Relational data are ubiquitous in the nature and their accessibility has not ceased to increase in recent years. Those data, see as a whole, form a network, which can be represented by a data structure called a graph, where each vertex of the graph is an entity and each edge a connection between pair of vertices. Complex networks in general, such as the Web, communication networks or social network, are known to exhibit common structural properties that emerge through their graphs. In this work we emphasize two important properties called *homophilly* and *preferential attachment* that arise on most of the real-world networks. We firstly study a class of powerful *random graph models* in a Bayesian nonparametric setting, called *mixed-membership model* and we focus on showing whether the models in this class comply with the mentioned properties, after giving formal definitions in a probabilistic context of the latter. Furthermore, we empirically evaluate our findings on synthetic and real-world network datasets. Secondly, we propose a new model, which extends the former Stochastic Mixed-Membership Model, for weighted networks and we develop an efficient inference algorithm able to scale to large-scale networks.

Résumé

Les données relationnelles sont omniprésentes dans la nature et leur accessibilité ne cesse d'augmenter depuis ces dernières années. Ces données, vues comme un tout, forment un réseau qui peut être représenté par une structure de données appelée graphe où chaque nœud du graphe est une entité et chaque arête représente une relation ou connexion entre ces entités. Les réseaux complexes en général, tels que le Web, les réseaux de communications ou les réseaux sociaux sont connus pour exhiber des propriétés structurelles communes qui émergent aux travers de leurs graphes. Dans cette thèse, nous mettons l'accent sur deux importantes propriétés appelées *homophilie* et *attachement préférentiel* qui se produisent dans un grand nombre de réseaux réels. Dans une première phase, nous étudions une classe de modèles de graphes aléatoires dans un contexte Bayésien non-paramétrique, appelé *modèle de composition mixée*, et nous nous concentrons à montrer si ces modèles satisfont ou non les propriétés mentionnées, après avoir proposé des définitions formelles pour ces dernières. Nous conduisons ensuite une évaluation empirique pour mettre à l'épreuve nos résultats sur des jeux de données de réseaux synthétiques et réels. Dans une seconde phase, nous proposons un nouveau modèle, qui généralise un précédent modèle à composition mixée stochastique, adapté pour les réseaux pondérés et nous développons un algorithme d'inférence efficace capable de s'adapter à des réseaux de grande échelle.

Table of Contents

List of Figures

List of Tables

List of Symbols

G	a graph representing a network
$\mathcal{V}$	set of nodes for graph G
$\mathcal{E}$	set of edges for graph G
N	number of nodes in G
E	number of edges in G
Y	adjacency matrix of G
y_{ij}	edge(s) between node i and node j
d_i	degree of node i
K	number of latent classes/blocks/features in the graph G
$\mathcal{M}_g$ or Ω	set of hyper-parameters of a (Bayesian) model
$\mathcal{M}_e$ or Π	set of random parameters (equiv. latent variable) of a (Bayesian) model
$\mathbf{\Theta}$	matrix of nodes' latent classes/features
$\mathbf{\Phi}$	matrix of latent class/feature interaction strength
θ_i	vector of latent class/feature of node i
ϕ_{mn}	strengh of interaction between class/feature m and n
$\boldsymbol{Z}$	tensor of block membership assignments
$z_{i\rightarrow j}$	block membership of node i interacting with j
$z_{i\leftarrow j}$	block membership of node j interacting with i
γ_{ij}	matrix of variational parameters for the interaction between nodes i and j

Abbreviations

Topic

AI	**A**rtificial **I**ntelligence
ML	**M**achine **L**earning

Model

ER	**E**rdős-**R**ényi (model)
BA	**B**arabàsi-**A**lbert (model)
CRP	**C**hinese **R**estaurant **P**rocess
CRF	**C**hinese **R**estaurant **F**ranchise
DP	**D**irichlet **P**rocess
HDP	**H**ierarchical **D**irichlet **P**rocess
IBP	**I**ndian **B**uffet **P**rocess
ILFM	**I**nfinite **L**atent **F**eature **M**odel
MMSB	**M**ixed **M**embership **S**tochastic **B**lockmodel
IMMSB	**I**nfinite **M**ixed **M**embership **S**tochastic **B**lockmodel
SBM	**S**tochastic **B**lock **M**odel

Inference Scheme

EM	**E**xpectation **M**aximization
CGS	**C**ollapse **G**ibbs **S**ampling
ELBO	**E**vidence **L**ower **Bo**und
MCMC	**M**arkov **C**hain **M**onte **C**arlo
SVB/SVI	**S**tochastic **V**ariationnal **B**ayes/**I**nference
SCVB	**S**tochastic **C**ollapsed **V**ariationnal **B**ayes

Chapter 1

Introduction

1.1 Thesis Overview

Networks are ubiquitous data structures that allow the representation of interactions between objects. Complex networks encompass real-world networks with non-trivial structural properties such as social networks of individuals in the society or information networks constituted of interconnected documents. For instance, one could think about the Web as made up of web pages interconnected by hyperlinks, or academic papers connected by citations. Their study comes with many exciting questions concerning the laws governing them, their dynamics and their invariant properties. Such questions are motivated by a better understanding of the *emergent* phenomena raised by the formation of complex interacting systems as well as the development of tools to enhance their analysis. The analysis of complex networks is a modern field of science that started around the fifty with the introduction of graph theory, motivated by questions from sociologists and psychologist seeking for a mathematical representation in order to study the patterns and regularities resulting from relations between various entities (Harary & Norman 1953). It quickly gained interest as the data and the computing resources became largely accessible and "cheap"[1]. The science of networks represents an opportunity for many scientific communities to interact and share knowledge, due to the diversity of the sources from which networks arise on one side, and the common phenomena that occur in their midst in the other (Wasserman & Faust 1994).

On the other hand, Machine Learning is a discipline at the crossroads of applied

[1] See for example https://www.blogdumoderateur.com/chiffres-reseaux-sociaux/

mathematics and computer science that aims at building algorithms able to *explore* various kinds of observable information such as texts, images, time-series or networks. The purpose of this exploration is to shape an abstract view of the observable information, also called the training data (Michalski et al. 2013). This abstract view is referred to as a model and is used to accomplish some prediction tasks on unobserved or future outcomes. The process of (machine) learning consists of fitting a model with the data which, in general, is reduced to a mathematical optimization problem[2] of an objective function. This measures, in some sense, how well the model explains the data. A Machine Learning model can be seen as a data "observer" that tries to explain the data it observes. However, as one can imagine, the observation of "complex" data may result in many different interpretations depending on the beliefs of the observer, and thus choosing a "good" explanation, or model, is uncertain. Probabilistic Machine Learning is a general framework used to develop models that incorporate beliefs and deals with the uncertainty. A decisive question though, is the choice of the appropriate beliefs and degree of uncertainty according to the data observed, and ultimately the decisions to make regarding these data (Ghahramani 2015).

The context of this thesis is to study and develop probabilistic machine learning based approach towards complex networks analysis. This approach provides a rich framework to capture network properties and develop statistical procedures (i.e. inference) that efficiently extracts information from network data sets, predict their evolution, while controlling the uncertainty levels that are inherent to complex networks. Our objective is twofold. Firstly, we want to understand what particular models are adapted for complex networks and why. Our study led us to random graph models as they represent a general well-founded framework for modeling a very large set of networks and their ability to control the amount of knowledge "a priori", or "assumptions", about the networks by design. In particular, we study a family of probabilistic models, called Mixed-Membership models, characterized by hierarchical relations of latent variables that encodes the modular structure of networks trough the concept of "community" (Airoldi et al. 2014). This family of models extends the former Stochastic Blockmodel (Goldenberg et al. 2010) by allowing the entities of a network to belong to several communities[3]. An open question, though is the characterization of those models regarding the

[2] Most of the time, the learning process can be cast into an optimization problem, but not always, as for MCMC techniques (3.1) for example.

[3] Which is akin to soft clustering.

emergent properties found in networks (Newman 2010). We especially focused on the homophily, often refers to with the statement "*Birds of a feather flock together*" (McPherson et al. 2001), and the preferential attachment effect, frequently refers to with the statement "*the rich get richer and the poor get poorer*" (Barabási & Albert 1999). We proposed in this direction, formal definitions of those properties adapted to the probabilistic framework and we study to what extent the models satisfy those properties with respect to the proposed definitions. Secondly, we aim to generalize those models to overcome some of their limitations, such as modeling networks based on binary relations only and the scalability issues regarding large networks. To do this, we proposed an adaptation of the former Mixed-Membership Stochastic Blockmodel (Airoldi et al. 2009) in order to model weighted edge covariates based on the Poisson Distribution. Further, we proposed a flexible hierarchical prior over the rate parameters of the Poisson distributions to give flexibility to the connectivity strength within the latent communities.

1.2 Thesis Outline

This manuscript is organized into 5 chapters. The Chapters 2 and 3 review the state of the art of the topic. They are devoted to setting out the background of the main results of this thesis and to put them in perspective by exposing the associated relevant literature, and giving some fundamental theoretical building blocks. In particular, Chapter 2 focuses on complex networks analysis and mining aspects by reviewing some of the key properties observed in real networks and applications. In Chapter 3, we concentrate on the probabilistic framework for random graph models and theoretic foundations. In those preliminaries chapters our motivation is to provide elements of response to the following questions:

- What kind of network data are we interested in and what are their properties? (2.2, 2.3)
- What kind of prediction tasks do we want to solve with those models? (2.4)
- What models exist to extract knowledge from the data and infer new outcomes, and what are the theoretic foundations of the model of interest? (3, 3.4)
- What learning processes and inference methods are used to fit the models to the data? (3.1, 3.3)

The Chapters 4 and 5 constitute the heart of this thesis. The contributions can be resumed as follows:

- In Chapter 4, we ask whether a certain class of powerful probabilistic models, namely the class of stochastic block models, comply with two important properties found in real-world networks, namely preferential attachment and homophily. More specifically we study two probabilistic models, the Infinite Latent Feature Model (ILFM) (Miller et al. 2009) and the Infinite Mixed-Membership Stochastic Blockmodel (IMMSB) (Airoldi et al. 2009; Kim et al. 2013) and show that in their standard formulation they do not comply with the homophily and the preferential attachment. However, we show that IMMSB comply with the local preferential attachment where only edges within communities are considered.
- In Chapter 5, we proposed a new stochastic block model that extends the IMMSB model in order to model weighted networks. Particularly, a hierarchical Beta-Gamma prior is proposed to have a flexible block-block parameter distributions. We develop an efficient inference algorithm able to scale on networks with millions of edges, and we evaluate and compare the model with various types of large real-world networks. We empirically show that the performance on the link prediction task can be improved when the networks are partially observed.

Chapter 2

Complex Network Analysis

2.1 Context of the study

The study of complex networks is grounded by the graph theory and in particular, for the statistical analysis of networks, by the random graph theory (Albert & Barabási 2002; Mark E.J. Newman 2003). The latter approach is particularly well adapted for complex networks because, by definition, their associated graphs don't have a *rigid* topological structure such as being regular, acyclic, complete, or other specific symmetries in their connectivity patterns, or at least, such structures are not assumed *a priori*. This can be resumed by the statement: *"a complex network is governed by simple assumptions"*. Meanwhile, using random structures provides a rich formalism to encode data priors and model the uncertainty through representation (3) that are sound with the assumptions made on the network (Orbanz & Roy 2015). A major difficulty in modeling complex networks is that they have several degrees of uncertainty regarding their topological structure and in the law that controls their dynamic. This uncertainty makes the problem of finding a good model to explain the construction of a given network (and make prediction on it) ill-defined. Besides, there is another source of uncertainty that comes from the fact that there is no strong consensus about the semantics behind the construction of the networks (Krackhardt 1999); what does it mean that there is a connection between two nodes in a network? Why and when is a connection established between two nodes? There is no obvious answer to those questions, and worse, the answer may not be the same depending on the type of the network considered (2.2). Even for two networks of similar type, the answer may differ for two different couples of nodes. For example, the notion of friendship in a social network can vary according to the country or the culture considered. Or, for a hyperlink network, such as the Web[1], a web page could point to another one because their content is related in some way, or maybe because it refers to a sponsor, which are two completely different reasons. Hence, to face this uncertainty we focus our study on probabilistic models as they provide a natural framework to build powerful and flexible models in this context (Ghahramani 2015).

Therefore, to build pertinent probabilistic models, one needs to seek and identify the characteristic properties of the data being modeled, in order to propose suitable assumptions. In the rest of the section, we give a quick review of the type of networks we are interested in. Then, we recall the basic properties of graphs before exposing the applications of interest in the context of our study.

2.2 Network type

A large variety of domains exist in our environment from which networks can arise. Most of us are familiar with certain type of network that surrounds us, especially since Information Technology (IT) and online social networking platforms have been widely adopted by the population. Indeed, those platforms that connect people who are "friends" have somehow democratized the access to a number of datasets which are used by, among others, the machine learning community[2]. They are also of great interest for the social sciences for empirical evaluation. Nevertheless, as exposed in this section, there are many different types of networks that emerge in other disciplines as well, such as Linguistics and Natural Language Processing (NLP), but also the Economy, Ecology and Biology.

More generally, the category of **social interaction networks** represents type of networks used to study any kind of relation between individuals, humans or not. Here is a non-exhaustive list of such networks:

- *Social networks*: They represent sets of entities with some relationships pattern between them. The pattern can be the friendships between individuals or business relations between companies for example. They are the most representative networks in term of available datasets and academic research (Kunegis 2013; Mark E.J. Newman 2003).

[1] The World Wide Web

[2] See the statistics on https://icon.colorado.edu and http://konect.uni-koblenz.de/

- *Communication networks*: They represent communication patterns between entities. The pattern usually takes the form of a message or information delivered from a sender to a receiver. Such networks can be built from email exchange in a company (Klimt & Yang 2004) or from phone call patterns (Aiello et al. 2001) for example.
- *Economic networks*: They represent transaction patterns between entities and are mostly based on human activities. For instance, one can think about user ratings of movies, user clicks over web content, or again, financial transactions (Bell & Koren 2007). Those networks have gained a particular focus in the recommender system community (Burke 2002).
- *Sexual contacts*: Networks of sexual relations between individuals have also attracted some attention (Liljeros et al. 2001).

Another important source of networks resides in the relations that can be extracted from textual content or collection of documents. This category is formed by **information networks** (sometimes called knowledge network) and the semantic behind their construction is strongly dependent on the format used to represent the documents:

- *Citation networks*: The entities are documents and relations are the citations between them. The academic paper citations form the most studied representative of these networks (Leskovec et al. 2007).
- *Collaboration networks*: The entities are authors and there is a relation between two authors if they have collaborated on a paper. These networks are also often studied through academic paper collaborations (Yang & Leskovec 2015; Ley 2002).
- *Hyperlink networks*: The entities are web-pages and the relations between them are the hyperlinks. Those networks arise from the web, and often represent a small region of it such as the hyperlinks in a set of related blogs (Adamic & Glance 2005), or again the hyperlinks of the Wikipedia website (Preusse et al. 2013).
- *Lexical networks*: The entities are words and a relation can be built in several ways between them. For example, a relation can be present if two words are consecutive (Newman 2006) or alternatively if they co-occur in a document (Leacock & Chodorow 1998).
- *Ontologies*: Those networks, also called *knowledge graph*, are used to represent the relations between concepts, data, and entities that substantiate one or

many domains. They are also often used to formalize the structure of databases as a means to describe its semantics (Bollacker et al. 2008).

Aside from social individuals or textual documents, a source of interconnected entities arises from technological devices. They form the **technological networks** that are constituted of artificial networks (i.e. made by humans) of resources such as the electricity power grid (Watts & Strogatz 1998), the Internet as the network of interconnected physical machines, or the network of roads that connects cities (Kalapala et al. 2003).

The last category of networks that also contains an abundant number of examples in our environment corresponds to **biological networks**, which are used to represent the relations between biological structures. Those networks can again be divided into two distinct sub-types. The first concerns interactions between organisms (exogenous relations) in their environment such as *predator-prey networks*, which are used to represent food-web, where nodes are the species and the edge indicates feeding relations (Lafferty et al. 2006; Thompson & Townsend 2000). These networks are sometimes referred to as ecological network. The second sub-type concerns interactions inside an organism (endogenous relation). Major examples of those are *metabolic networks* that represent the functioning of cells at a chemical level (Jeong et al. 2000; Stelling et al. 2002), *protein interaction* networks that map physical interactions that proteins can have together (Maslov & Sneppen 2002), or the *gene regulatory* networks that express the relations between genes and proteins production (Guelzim et al. 2002; Kauffman 1992). Lastly, *neural networks* constitute a very important class of biological networks, and although they are very difficult to measure due to brain access, they have been investigated as in (Sporns 2002).

Note that some of the network types presented here can overlap in some case as there is no strict definition to state to which category a particular network belongs. For example, one could consider that the notion of friendship in a friendship network relates to a communication flow shared between individual and thus, call it a communication network. Or from another point of view, that an email exchange network or even a collaboration network underlay some social relations. Furthermore, the classification of networks presented here is based on the similarities of the *nodes* of the graph. But, a classification based on the similarities of the *edges* (i.e. the similarities between the patterns of interactions) could suggest different classification choices. Recently, interesting metrics have also been proposed in order to compare and classify networks. In (Onnela et al. 2012), the authors defined a measure of similarity for graphs to automatically construct taxonomies of networks. Another interesting methodology to compare pair of networks was proposed in (Asta & Shalizi 2015), where a network is approximated by continuous geometric object.

In this thesis, we aim to study general properties that occur in most of the complex networks. It turns out that several properties have been found to emerge in most of the real-world networks, that we briefly reviewed in section 2.3. In particular, it is known that social networks exhibit community structure, power-law degree distributions and are sparse (Barabási & others 2016). However, power-law degree distributions and sparsity are not verified by all the categories of real-world networks. For instance, it has been shown that most food webs appear to be dense (Dunne et al. 2002) and that may also be true for other biological networks such as metabolic network (Mark E.J. Newman 2003). In general, biological and technological networks may exhibit specific properties, which require different expert knowledge than would be required for social and information networks. Therefore, our work focuses primarily on properties that characterize social interaction networks and information networks and, our study and development of models in chapter 4 and 5 move in this direction as well.

2.2.1 Datasets

An important collection of networks have been collected in recent years by some universities and researchers to provide clean repositories of datasets and helping the research in networks science and the reproducibility of experiments.

The Stanford Network Analysis Project (SNAP) provides various useful resources to the community, such as events, tutorials, publications and datasets[3] (Leskovec & Krevl 2014). Some datasets are provided by[4] (Batagelj & Mrvar 2006). The University of California maintains a repository with links that point to datasets curated by individuals[5]. The page of Tore Opsahl contains a list of social networks datasets[6].

More recently, an effort has been made to build and maintain an index of the

[3] https://snap.stanford.edu/data/index.html
[4] http://vlado.fmf.uni-lj.si/pub/networks/data/
[5] https://networkdata.ics.uci.edu/resources.php
[6] https://toreopsahl.com/datasets/

existing network datasets found in the literature. They also provide descriptions, statistics and source references. There are two projects that provide such indexes:

- The KONECT (the Koblenz Network Collection) by the University of Koblenz–Landau[7] (Kunegis 2013).
- The ICON project based at the University of Colorado Boulder[8] (Aaron Clauset & Sainz 2016).

2.3 Network properties

We will first provide some basic definitions used throughout the manuscript, and then recall some of the properties found in real-world networks. The Table 2.1 recalls the basic terminology of networks analysis.

2.3.1 Definitions

Graph. We can represent a network by a graph. A graph G is defined by a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$ such that $G = (\mathcal{V}, \mathcal{E})$. We denote the number of nodes as $N = |\mathcal{V}|$ and the number of edges as $E = |\mathcal{E}|$. A graph may be *directed* or *undirected* and may contains *self-loop* or not. For undirected graph without self-loop, the number of possible edges is $\begin{pmatrix} N \\ 2 \end{pmatrix}$. In general, if not specified, we will consider undirected graph for convenience. A graph may be binary or weighted. For binary graph, let y_{ij} indicates the presence or absence of an edge with $y_{ij} = 1$ if there is an edge between i and j and $y_{ij} = 0$ otherwise. For weighted graph, let y_{ij} indicates the weight between i and j — for instance the number of calls in a communication network, or the number of links in a hyperlink network. The *adjacency matrix* of a graph G is a matrix of edge indicators where nodes are in rows and columns denoted $Y = (y_{ij})_{N \times N}$. For each node i, let d_i be its degree defined as the number of adjacent edges of i, such that $d_i = \sum_{j \in \mathcal{V}} y_{ij}$. A graph can be *unipartite* or *multipartite*. For multipartite graph, there are multiple classes of nodes and edges, which are drawn only between nodes of different classes (note that in this case the adjacency matrix is not square) — for instance, a bipartite graph can be a movie rating network where y_{ij} represents the rating of a movie j by an individual i. However, we will only consider unipartite graphs in the following,

[7] http://konect.uni-koblenz.de
[8] https://icon.colorado.edu

Table 2.1: Short glossary of terms (Newman 2003).

Graph: The mathematical representation of a networks defined in terms of vertices and edges.

Vertex (pl vertices): The fundamental unit of a network, also called a site (physics), a node (computer science), or an actor (sociology).

Edge: The line connecting two vertices. Also called a bond (physics), a link (computer science), or a tie (sociology).

Directed/Undirected: An edge is directed if it runs in only one direction (such as an email sent to a person), and undirected if it runs in both directions. A graph is directed if all of its edges are directed. An undirected graph can be represented by a directed one having two edges between each pair of connected vertices, one in each direction.

Self-loop: A self-loop indicates an edge from a vertex to itself.

Degree: The number of edges connected to a vertex. Note that the degree is not necessarily equal to the number of vertices adjacent to a vertex, since there may be more than one edge between any two vertices. A directed graph has both an in-degree and an out-degree for each vertex, which are the numbers of in-coming and out-going edges respectively.

Density: The density of a graph is its number of edges divided by the maximal number of edges.

Geodesic path: A geodesic path is the shortest path through the network from one vertex to another. Note that there may be and often is more than one geodesic path between two vertices.

Diameter: The diameter of a network is the length (in number of edges) of the longest geodesic path between any two vertices.

wich translates to a square adjacency matrix of size $N \times N$. For undirected graph, the adjacency matrix is symmetric such that $y_{ij} = y_{ji}$.

Additionally, a graph can be *multi-relational* where the edges can belong to several classes. In this case, the graph is represented by an adjacency tensor. However, in this work, we will only focus on *uni-relational* graph.

Finally, a graph can be *dynamic* (sometimes call temporal or time-varying network) if one or several of its characteristics can evolve over time. This can translate into a birth and/or death process for both nodes and edges. A dynamic graph $\mathcal{G}$ can be represented by a sequence of graph snaphsots $\mathcal{G} = (G_1, G_2, G_3, \dots)$, where each graph G_t represents the network at a time step t.

In this thesis, we will focus though on unipartite and static graphs only.

In the next section, we will introduce the emerging properties that are often observed in real-world networks.

2.3.2 Community structure

Real-world networks are known to exhibit a modular structure, where nodes are grouped together if they share some common topological patterns (Flake et al. 2002; Girvan & Newman 2002; Schwartz & Wood 1992). Those groups are often referred to as communities where a *community* is generally defined as a set of nodes that are more tightly connected to each other than those outside the community (Fortunato & Hric 2016; Fortunato 2010). The precise definition of what constitutes a community in an network though can vary depending on the authors and is still an active area of research (Rosvall et al. 2017).

A concurrent approach to identify groups or communities in networks is based on the notion of roles that are formally defined trough the formal concept of equivalence relation (Everett & Borgatti 1994; Holme & Huss 2005; Rossi & Ahmed 2015). In particular, two nodes are said to be *regular equivalent* (i.e. same role) if they have similar connections to nodes having same roles without being necessarily the same. A particular case of the regular equivalence is the *structural equivalence* where the nodes should share exactly the same neighbors.

A similar equivalence of the latter is the so called *stochastic equivalence* where two nodes are "stochastic equivalent" if their probability distributions of edges with other nodes are the same. Note that in this case, the neighborhood of two equivalent nodes has not to be exactly the same which provides more flexibility compared to the structural equivalence (Goldenberg et al. 2010; Wasserman & Faust 1994). The stochastic equivalence is at the basis of the so-called stochastic blockmodel that constitutes the building block of the models studied in this thesis 4.

For formal definitions of equivalence relation in graphs, the reader can refer to (White & P. Reitz 1983).

2.3.3 Mixing patterns and Homophily

Mixing patterns refers to the tendency of certain type of nodes to connect to another type. For example, studies on networks of married and unmarried couples have shown strong correlation between the ages of the partners (Garfinkel et al. 2002). In general, this kind of selective linking is based on the nodes attributes or characteristics, which are dependent of the type of network analyzed, to measure similarities between them. For social networks, it has been admitted that individuals tend to associate between those similar in some way. This is known as *assortative mixing* or *homophily* and it has been widely covered in the literature (McPherson et al. 2001; La Fond & Neville 2010; Kim & Altmann 2017). In contrast, when nodes tend to connect to those dissimilar, the networks is said to be *disassortative* or *heterophilic*. Several metrics have been proposed to measure to what extent a network exhibits assortative (homophilic) mixing patterns that usually rely on some side information about nodes (type, temporal or geographic features...) that are assumed to be known (Mark EJ Newman 2003). A sub-case of assortative mixing consists of measuring nodes pair similarity with the network topology information only, such as node's centrality. In particular, the measure of the degree has received some attention (Mark E.J. Newman 2003), where the idea is to ask if a node with a high degree prefers to connect to nodes with either high degree or low-degree; it appears that both situations can occur in some networks.

2.3.4 Preferential attachment

A key element to characterize the structural properties of networks is the study of the repartition of the node degrees. Notably, real-world networks have been found to exhibit degree distributions that lie in the family of *long-tailed* distributions, which are a sub-class of heavy-tailed distributions (Clauset et al. 2009). They represent functions with a heavier tail than the exponential distribution, that is, with a slower decay. The slow decay of heavy-tailed distributions is hence in contrast to typical Gaussian distributions that have an exponential decay[9]. The Long-tailed distributions family includes the power-law and the Pareto distribution among others. A long-tailed distribution is characterized by events (the nodes' degree in a graph for instance) with a high-frequency that concentrates a large part of the population located in the head of the distribution, followed by events with low frequency that gradually decrease asymptotically, called the tail[10]. This property has been coined in many domains in science, for example in social science, where it has been called the *Matthew effect* in reference to biblical texts of the *Gospel of Matthew* (Merton 1968). In statistics, the phenomenon is often referred to as the *Pareto principle* after the Pareto distribution, where the least frequently occurring items (e.g. the degrees) represent 80% of the population and the most frequent ones represent only 20%. Lastly, it has also been discovered in the word distribution of the natural language, and it is known as the *Zipf law* (Zipf 2016).

A proposed explanation of the emergence of fat-tailed distribution of degrees in networks is based on the *preferential attachment effect*, which encodes the idea that *the more you have, the more you will get*. It states that a node in the network will attach with higher probability to nodes that have a high degree (Barabási & Albert 1999), and leads to the famous Barabàsi-Albert (BA) model that generates networks with power law degree distribution such that

$$P_k \sim k^{-\alpha},$$

where P_k is the probability that a node chosen uniformly at random has degree k and α a constant exponent greater than zero. Formal proof of the raise of a degree distribution with power law from the preferential attachment models has been done in (Bollobás et al. 2001). Networks with power-law distributions are also referred to as *scale-free* networks as the power-law *"is the only distribution that is the same whatever scale we look at it on."* (Newman 2005).

2.3.4.1 Sparsity and scale-free networks

Sparsity is a property observed on most of the real-world network datasets and means that the number of edges E is very low compared to the network capacity, which naturally increases quadratically with the number of nodes N. The question

[9] This is also the case of the Poisson distribution often used to represent count processes.

[10] In this case we say that the distribution right-tailed, but is can also be left-tailed or both.

we ask here is how the sparsity is related to scale-free networks? We show that scale-free networks leads to sparse networks if $\alpha > 2$ (Note that an extended proof with similar arguments as below has been proposed in (Del Genio et al. 2011)).

Formally we say that a network G is sparse if the following limit is true:

$$\frac{E}{N^2} \to 0 \qquad \text{as} \quad N \to \infty \ .$$

Let G be a scale-free undirected network of size N with degree distribution $f(k) = Ck^{-\alpha}$ with C and α two positive constants. Let f_k be the number of nodes having a degree equal to k. The number of edges in the network is then

$$E = \frac{1}{2}\sum_{k=0}^{\infty} k f_k \ .$$

Further, when $N \gg 0$, we assume that the empirical degree distribution converges towards the true degree distribution such that $f(k) = \frac{f_k}{N}$. Thus, one obtains:

$$E = \frac{NC}{2}\zeta(\alpha - 1) \ ,$$

where ζ is the Riemann zeta function. Hence, E is not divergent if $\alpha > 2$ and in this case, E has a linear growth with N as $E = \mathcal{O}(N)$, which results to a sparse network.

2.3.5 Small-world effect

The small-world effect is an important emerging property that has been found in many real-world networks (Watts 2004). It has been popularized by the well-known experiment of Milgram (Travers & Milgram 1967) in which letters, passed from person to person, were able to reach any individual in a small number of steps. Though, it has been speculated in earlier work (Karinthy 1929; Sola Pool & Kochen 1978). A major practical implication of the small-world effect in real-world networks concerns the speed of the spread of information.

The phenomenon is related to the slow growth of the geodesic path as the network size increase. Let us define L as the mean length of geodesic (i.e. shortest) path between nodes pair in an undirected network:

$$L = \frac{2}{N(N+1)} \sum_{i<j} d_{ij}$$

where d_{ij} is the length of the geodesic path from node i to node j. Many real-world networks exhibit the small-world effect in the sense that it has been observed that the value of L scale logarithmically or slower with the number of nodes N such as $L \propto \log N$ (M. Newman 2001; Mark EJ Newman 2001). Interestingly, Many random graph models are also known to exhibit the small-world effect (Bollobás 1981). Another interesting question concerns the relation between scale-free networks and small-world phenomenon. A highlighting result provided by (Bollobás & Riordan 2004), showed that the mean geodesic path L in a network with power law degree distributions increases no faster than $\log N / \log \log N$.

2.4 Applications

Applications of network analysis aim at developing models and algorithms to perform various kinds of tasks. They can be divided into two opposite categories:

- *Generative based*: Here the objective is to generate networks according to a model. Applications include the development of simulators that can be used for visualization purpose or to provide synthetic datasets for the scientific communities. The challenge of this problem is to generate graphs that have relevant properties according to the type of networks we want to simulate.
- *Learning based*: Where the goal is to build models and algorithms that can extract knowledge from a given network (or a set of networks) and eventually predict future outcomes. The challenge here is twofold; the first is to make accurate predictions and the second is to scale the algorithms to large networks..

In this thesis, we focus on learning based applications. Specifically we are interested in two sub-tasks that have been widely studied in the literature, namely the *community detection* and the *link prediction*.

2.4.1 Community detection

Community detection is a task that consists of solving a clustering problem where one tries to find the best partition of the nodes according to a given criterion generally based on the network structure (Khan & Niazi 2017). In general, this criterion is a means of identifying *communities*. A standard criterion in this direction is the **modularity** used to find communities under the regular equivalence. The optimization method was originally proposed through a greedy algorithm (E.J. Newman & Girvan 2004). However, methods based on the modularity are known to have resolutions issues, where small communities are undetected (Fortunato & Barthelemy 2007). Modularity is also affected by the field of view limit. In contrast to the resolution limit this last one results in overpartitioning the communities with large diameter (Schaub et al. 2012). Recently, several scalable methods have been introduced to optimize the modularity such as simulated anealing approach and the Louvain algorithm (Chen et al. 2014). A concurrent approach is the clustering of network nodes in the context so-called *block models* (Breiger et al. 1975; White et al. 1976), which are essentially just partitions of the nodes into blocks (or classes) according to a criterion. The advantage of block modeling is that it offers more flexibility in the definition of the criterion than the modularity based approaches as they allow to capture communities either under the regular equivalence or the structural equivalence (Gopalan & Blei 2013; Karrer & Newman 2011). Furthermore, they can be cast into probabilistic model to allow more flexibility as exposed in section 4. The relation between (stochastic) block modeling and modularity has been studied in (Bickel & Chen 2009) and notably they established under what conditions their respective objectives are consistent. Probabilistic versions of the block model such as the Stochastic Block Model and the Mixed Membership Stochastic Blockmodel have also been used for community detection (Holland et al. 1983). Many others algorithms have been proposed to find community structures; we do not detail them here and refer the interested reader to (Coscia et al. 2011; Fortunato & Hric 2016; Newman 2004).

2.4.2 Link prediction

In the *link prediction* task, one assumes a partially observed network with missing links. The goal is then to predict the missing links, that is, to predict if either an edge exists or not between any unobserved relations between nodes (Al Hasan & Zaki 2011; Lü & Zhou 2011; Getoor & Diehl 2005). In the case of weighted graphs, the prediction concerns the number of edges or the weights between two nodes. In the learning based context, there are two major approaches to solve this problem[11]. The first is based on matrix factorization techniques (Menon & Elkan 2011) and the second on latent variable models (Wang et al. 2015), also referred to as probabilistic models that encompass the familly of Stochastic Block Models. The latter is the approach pursued in the thesis[12] because of the flexibility and the unifying aspect of the probabilistic framework, as exposed in section 4. Further, Stochastic Block Model and its extensions are generative models which are "cluster based" in the sense that the edges are generated on the basis of the nodes membership to some latent communities. Therefore, they provide the advantage of being able to be used, in addition than the link prediction task, for community detection as well as for graph simulation/generation. This latter task can be used to generate networks that mimic the properties of the training data sets, in the limit of the intrinsic properties of the models, which are explored in section 4.

2.4.3 Other applications

Though not in the scope of this thesis, it is worth mentioning another important application, which consists of studying the *diffusion processes* in networks. From the diffusion of innovations to the spread of disease or more generally, the propagation of information between network members, the question is to identify the impact of the structural properties of network on the diffusion process and localize the regions that either maximize or minimize the latter (Z.-K. Zhang et al. 2016; Pei & Makse 2013). A related notion is the network resilience that evaluates the impact of random deletions of nodes and/or edges, and which can be studied through the percolation theory (Callaway et al. 2000).

Yet another application is the exploitation of the graph topology for *Information Retrieval*. Canonical examples rely on random walk, which is at the basis of the Page-Rank algorithm used by search engine to extract relevant web pages on the web (Kleinberg 1999; Page et al. 1999).

[11] Other Standard non learning based approaches relies on ad-hoc similarity measures that use topological properties of nodes such as common neighborhood, short paths or other (Liben-Nowell & Kleinberg 2007)

[12] Note that the two approaches are closely related as in some case matrix factorization models can be equivalently interpreted in terms of probabilistic models. This question is also related to frequentist vs Bayesian reasoning debate.

2.5 Summary

We presented in this chapter several classes of complex networks found in the real world, and in particular social and information networks on which we focus in this thesis. We review some of the emergent properties that arise in those networks as they constitute observation evidence and will be used to guide our assumptions and evaluations of the network models throughout the next chapters. We also presented the typical applications that concern network analysis and especially the one we have been considering in our empirical evaluations in chapters 4 and 5.

Chapter 3

Bayesian Models

Machine Learning can be though of as inferring plausible models to explain observed data. A major difficulty of this task is that the data can be consistent with many models and choosing an appropriate model is uncertain. Therefore, being able to represent the uncertainty plays a key role in order to build flexible yet powerful model (Ghahramani 2015). For network analysis, one can think about a clustering task such as the *community detection* for instance; should the clusters to detect should satisfy the *regular equivalence* or the *structural equivalence*? Should the node memberships to clusters should be soft or strict? How many clusters should be detected? The answers to those questions can either be known or uncertain and therefore, one should be able to incorporate various kinds of prior knowledge and with different levels of confidence. In this direction, a well-grounded framework to control the uncertainty is the Bayesian Inference, grounded by the probability theory, and which allows the actual learning procedure behind probabilistic models.

In this chapter, we introduce the probabilistic framework upon which the models studied in this thesis are based. In particular, we expose the key distributions and the processes harnessed within the models considered throughout the manuscript, before reviewing the class of latent variable models (i.e. Bayesian models) used for network analysis and some fundamental result justifying their constructions.

3.1 Graphical models

Bayesian modeling is a probabilistic framework used to formalize causal theory (Fong 2013). Given a set of observable data X, let Π be a set of random variables and Ω a set of hyper-parameters. A Bayesian model defines the conditional relations of dependancy (and independancy) between those variables whom describe how the data are generated. In a nutshell, the generative process consists firstly to generate random parameters from a *prior distribution* $\Pi \sim P(\Pi|\Omega)$, then the data are generated from a *data likelihood distribution* given the parameters such that $X \sim P(X|\Pi)$. The model can be represented as a Directed Acyclic Graph (DAG), which is a Graphical model[1], where the conditional relations between the different variables are emphasized as illustrated in Fig. 3.1. The graphical model is a means of expressing the causal relations underlying a probabilistic model in a visual and synthetic way.

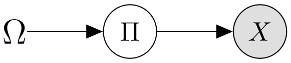

Figure 3.1: A simple Bayesian model. The circled nodes represent random variables, and the non-circled one represent constant generally called hyper-parameters. The grey circles represent the observed data and the white circles correspond to the model parameters. The directed edges represent causal relations between the variables.

Bayesian inference is an inversion procedure that consists of estimating the model parameters Π given the observable data. This inversion is realized through the Bayes' law who expresses the distribution of the model parameters given the observable data, called the *posterior distribution*, as

$$P(\Pi|X,\Omega) = \frac{P(X|\Pi)P(\Pi|\Omega)}{P(X|\Omega)}$$

where $P(X|\Omega) = \sum_{\Pi} P(X,\Pi|\Omega)$ is referred to as the *marginal likelihood* or model *evidence*. For simple model, the inference method of the posterior is generally based on either the Maximum Likelihood Estimation (MLE) or the Maximum a Posteriori (MAP) algorithm. These methods are called *point estimation* as they give a single value that tries to maximize the posterior distribution. However, for more complex models, the posterior cannot be computed directly, due to non closed-form expression of the evidence, and one must resort to approximate inference methods. Two main concurrent approaches have been explored in the literature to approximate the posterior distribution. The first relies on sampling techniques grounded by Markov Chain Monte Carlo (MCMC) theory (Neal 1993; Geyer 2011); MCMC based methods are stochastic procedures where successive

[1] Graphical models are used to represent various type of probabilist models. Directed Acyclic Graph (DAG) are used for Bayesian models. Another important type of probabilistic models are the Markov Random Field (MRF) which are represented by undirected Graphical models (Sutton et al. 2012).

sampling steps are performed to approximate the true posterior distribution. The second relies on Variational Inference (VI) (equiv. Variational Bayes) methods. In this inference scheme, one tries to minimize the divergence between the true posterior and a given proxy distribution. A major advantage of this approach is that it allows developing deterministic inference procedure and thus, open the door to the framework of gradient descent based algorithms. Nevertheless, the price to pay is that the proxy distribution incorporates a bias often hard to evaluate (Blei et al. 2017). As a final note on approximate inference, it is worth mentioning the Expectation-Maximization (EM) algorithm, which is a baseline method for MAP approximation when the posterior is tractable or for direct optimization of the prior parameters (i.e. frequentist approaches.).

The estimation of the posterior distribution is achieved through a fitting procedure that, for approximate inference, consists of iterative updates of an objective towards a maximizer. The (approximate) posterior can then be used to "answer questions" through the prediction of future outcomes as

$$P(x_{new}|X,\Omega) = \int P(x_{new}|\Pi, X)P(\Pi|X)d\Pi$$

And $P(x_{new}|X,\Omega)$ is referred to as the predictive distribution of an unobserved outcome x_{new}.

3.2 Exponential family

The Exponential family represents a class of parametric distributions that subsumes many common distributions. It includes the Normal, Poisson, Bernoulli, Multinomial, Beta, Dirichlet, Gamma, Exponential, Pareto etc. Briefly, a distribution in the exponential family can be expressed, in its canonical form, by:

$$P(X|\eta) = \exp(\eta^T S(X) - A(\eta) - H(X))$$

where $S(X)$ is a measurable map, A, called the log-partition function, and H are two known measurable real-valued functions.

Distributions in the Exponential family have convenient properties that make them theoretically appealing and hence, constitute a core topic in the Bayesian framework (Orbanz 2009) and noticeably in the field of Information Geometry (Amari & Nagaoka 2007). For instance, they convey the key notion of *sufficient* *statistic* regarding the measure of the information in a data sample. There are also at the basis of the *generalized linear models*, which allow straightforward generalization of simpler models to work in different contexts. Another very useful property of the exponential family, which we shall highlight in this section, is the existence of *conjugate priors*. More Precisely, let $P(X|\eta)$ be a model with parameter space Ω_η, and let $\mathcal{H}$ be a set of prior distribution on Ω_η. Then, the model $P(X|\eta)$ and the set $\mathcal{H}$ are said to be conjugate if for every prior $P_\eta \in \mathcal{H}$ and observation set $X = x$, the corresponding posterior $P(\eta|X)$ is an element of $\mathcal{H}$.

Conjugate priors, as well as the notion of sufficient statistics, are inextricably linked to the exponential family (Halmos et al. 1949). Notably, it has been showed that an exponential family representation always implies the existence of a sufficient statistic and a conjugate prior. Furthermore, under mild regularity conditions, the converse is also true.

A conjugate prior of $P(X|\eta)$ has the following representation

$$P(\eta|\tau, \eta_0) = \exp(\tau^T \eta - \eta_0 A(\eta) - H(\tau, \eta_0)) \tag{3.1}$$

The practical advantage of using conjugate prior is that it leads to closed-form updates for inference applications. This is in particular often the case for predictive distribution in conjugate model, which is of great interest to develop efficient inference scheme. Note, that the use of conjugate prior in machine learning model is further justified by its mathematical convenience than for its epistemological meaning (Blei et al. 2003). To conclude on exponential family, we recall a theorem proved in (Diaconis et al. 1979), that characterizes the predictive distribution in a conjugate model, and that highlight its linear form that the reader will be able to rediscover in different parts of this manuscript.

Theorem 3.2.1 (Diaconis-Ylvisaker characterization of conjugate priors) *Let $P_X(.|\Theta)$ be a natural exponential family model dominated by Lebesgue measure, with open parameter space $\Omega_\theta \subset \mathbb{R}^d$. Let P_Θ be a prior on Θ which does not concentrate on a singleton. Then P_Θ has a density of the form (3.1) w.r.t Lebesgue measure on $\mathbb{R}^d$ if and only if [81]*

$$\mathbb{E}_{P_\Theta(\Theta|X_1=x_1,\ldots,\mathcal{X}_n=x_n)}[\mathbb{E}_{P_X(X|\Theta=\theta)}[X]] = \frac{a + n\hat{x}}{b + n} \ .$$

That is, given observation $x_1, \ldots, x_n$ the expected value of new draw x under unknown value of the parameter is linear in the sample average $\hat{x} = \frac{1}{n}\sum x_i$.

3.3 Nonparametric processes

When modeling natural phenomena, we mentioned the importance of having flexible model able to adapt to the complexity of the data. Flexibility can be obtained through Bayesian nonparametric that refers to Bayesian models that use nonparametric processes as prior knowledge. The class of nonparametric process can be though as the generalization of parametric distributions[2] to an infinite dimensional parameter space. Hence, the dimension of the model becomes itself a random parameter that can be learned from the data.

Nonparametric processes include major models that are worth mentioning. The Gaussian Process, that generalizes the multivariate Normal distributions (Rasmussen 2004), is particularly adapted for modeling continuous data such as time series and image processing (Lawrence & Moore 2007). The Poisson Process can be defined as a counting process on a measurable space, where each region is associated to a finite-dimensional Poisson distribution. It is ubiquitous in queuing theory (Nelson 2013), and has been used in a wide range of applications, from earthquake occurrence modeling (Ogata 1988), to the analysis of photon emission (Jäger et al. 2009). In this work, we are particularly interested in the two others following nonparametric processes:

- The *Dirichlet Process* is the generalization of the Dirichlet distribution in the infinite case. It is adapted for categorical data and often used as a prior for infinite mixture models and clustering applications.
- The *Indian Buffet Process* is a prior over categorical matrices with infinite columns. It has been used for discrete matrix factorization, overlapping community detection and model selection.

In the rest of this section, we introduce the Dirichlet Process, its extension namely the Hierarchical Dirichlet Process and the Indian Buffet Process.

[2] Such as distributions in the exponential family.

3.3.1 Dirichlet Process

A Dirichlet process (DP) is a random probability measure G over a measurable space $(\mathcal{X}, P(\mathcal{X}))$, with base measure H, and concentration parameter $\alpha_0 \in \mathbb{R}_+^*$, if for any partition $(A_1, \ldots, A_k)$ of $\mathcal{X}$

$$(G(A_1), \ldots, G(A_k)) \sim \mathrm{Dir}(\alpha_0 H(A_1), \ldots, \alpha_0 H(A_k))$$

and we write $G \sim \mathrm{DP}(\alpha_0, H)$ and where $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$ is the Dirichlet distribution which is defined by a degenerate density on simplex $\Delta_k = \{p_1, \ldots, p_k\} \in \mathbb{R}_+^k$ such that $P(p_1, \ldots, p_k) \propto p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1}$ and $\sum_i p_i = 1$.

The existence of the DP was established by Ferguson (Ferguson 1973) and has the following properties (Teh 2010):

- The expectation of a DP for any sets $A, B \subset \mathcal{X}$ is:

$$\mathbb{E}[G(A)] = H(A),$$

  sometimes simply noted $\mathbb{E}G = H$. This result is analogous to the expectation of a Dirichlet distribution where $\mathbb{E}[p_i] = \alpha_i / \alpha_.$. Further, the variance of G is $\mathbb{V}[G(A)] = \frac{H(A)(1-H(A))}{1+\alpha_0}$ and its covariance is $\mathrm{Cov}(G(A), G(B)) = \frac{H(A \cap B) - H(A)H(B)}{1+\alpha_0}$.
- The marginal distribution of G(A) is Beta such that $G(A) \sim \mathrm{Beta}(\alpha_0 H(A), \alpha_0(1 - H(A)))$. One can again recognize the analogy with the Dirichlet distribution.
- The posterior distribution of i.i.d. draws $(X_1, \ldots, X_n)$ from a DP G is:

$$G \mid X_1, \ldots, X_n \sim \mathrm{DP}\left(\alpha_0 + n, \frac{1}{\alpha_0 + n}\left(\alpha_0 H + \sum_{i=1}^{n} \delta_{X_i}\right)\right) \tag{3.2}$$

  where δ_{X_i} represents the point mass located at X_i[3]. This result directly follows from the conjugacy between Dirichlet and Multinomial distributions.
- The predictive distribution of a new draw X_{n+1} is:

$$\begin{aligned} P(X_{n+1} \in A \mid G, X_1, \ldots, X_n) &= \mathbb{E}[G(A) \mid X_1, \ldots, X_n] \\ &= \frac{1}{\alpha_0 + n}\left(\alpha_0 H(A) + \sum_{i=1}^{n} \delta_{X_i}(A)\right) \end{aligned}$$

[3] δ is the Dirac operator.

Finally by marginalizing out G, we obtain:

$$X_{n+1} \mid X_1, \ldots, X_n \sim \left(\frac{\alpha_0}{\alpha_0 + n} H + \frac{n}{\alpha_0 + n} \frac{\sum_{i=1}^{n} \delta_{X_i}}{n} \right) \tag{3.3}$$

Figure 3.2 represents the graphical model of the DP for the simple sampling scheme just described.

The predictive distribution for X_{n+1} is therefore equal to the base measure of the posterior distribution of G. This sequence of predictive distributions refers to the Blackwell-MacQueen urn scheme which has been used to show the existence of the DP (Blackwell et al. 1973). Furthermore, it emphasizes several important properties of the DP:

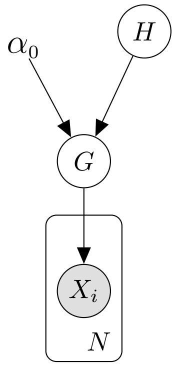

Figure 3.2: Graphical model of a sequence generated through the Dirichlet Process.

Exchangeability

Using the predictive distribution, Eq. 3.3, one can construct the distribution over the sequence $X_1, X_2, \ldots$ by iteratively drawing each X_i given $X_1, \ldots, X_{i-1}$ such that

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_{i-1}, \ldots, X_1).$$

It is straightforward to show that this joint distribution is infinitely exchangeable, meaning that the probability of the sequence $X_1, \ldots, X_n$ is the same for any other order of that sequence. Or said differently, the probability of that sequence doe not depend on the order in which we draw samples. Formally, given any permutation σ over natural number, one has

$$P(X_1, \ldots, X_n) = P(X_{\sigma(1)}, \ldots, X_{\sigma(n)}).$$

On the other hand, de Finetti's theorem states that (Kingman & others 1978) for any infinitely exchangeable sequence $X_1, X_2, \ldots$ there is a random measure G such that the sequence is composed of i.i.d draws from it[4]:

$$P(X_1, \ldots, X_n) = \int \prod_{i=1}^{n} G(X_i) dP(G).$$

In the Blackwell-MacQueen urn scheme, the prior over the random measure $P(G)$ is precisely the Dirichlet Process $DP(\alpha_0, H)$, and therefore justify its existence.

Discrete Distribution and clustering property

A characteristic of the predictive distribution of the DP is that its draws belong to some points mass (or atoms) and that there is a positive probability that new draws will take the value of a preceding atom. Therefore, the sequence $(X_1, \ldots, X_n)$ will take values in a set $(X_1^*, \ldots, X_T^*)$ with $T \leq n$. Hence, the posterior distribution is a weighted average sum of new draw from the base measure H and the empirical distribution. Let $(n_1, \ldots, n_T)$ be the counts associated to uniques atom values, one can rewrite the empirical distribution of draws as

$$\sum_{i=1}^{n} \frac{\delta_{X_i}}{n} = \sum_{k=1}^{T} n_k \frac{\delta_{X_k^*}}{n}.$$

It follows that, by rewriting Eq. 3.3, it makes appear a **rich-get-richer** phenomenon over the atoms, as the probability to draw an element equal to a given atom value X_k^* increases (is proportional) with the number of atoms having this value n_k. This clustering effect of the DP leads to Chinese Restaurant Process metaphor (3.3.2) while the discrete aspect of the DP leads to another construction of it called the Stick Breaking Process (3.3.3). Note that this discrete aspect of the DP is true whether the base measure is discrete or continuous.

It is worth mentioning that the DP has other interesting properties such as being self-similar (fractal property) and tail-free. We refer the interested reader to (Ferguson et al. 1992) for further details.

3.3.2 Chinese Restaurant Process

The discreteness and clustering property of the DP, as mentioned previously, make repeated draws $(X_1, \ldots, X_n)$ a particular partition taking values into $(X_1^*, \ldots, X_T^*)$.

[4] Or says differently, that draws are conditionally independent given G.

The predictive distribution can then be rewritten as:

$$X_{n+1} \mid X_1, \dots, X_n \sim \frac{1}{\alpha_0 + n} \left(\alpha_0 H + \sum_{k=1}^{T} n_k \delta_{X_k^*} \right) \tag{3.4}$$

With n_k the number of atoms for cluster k such that $n_k = \sum_{i=1}^{n} \delta_{X_i}(A_k)$.

The Chinese Restaurant Process (CRP) is the process associated to the predictive distribution of the DP where G has been marginalized out (Eq. 3.4). It illustrates the infinite mixture model induced by the DP. It is defined as follows:

- Assume a Chinese Restaurant with an infinite number of tables, each table can welcome an unlimited number of customers and table k serves dish X_k^*.
- First customer sits at first table.
- Suppose there are T tables occupied when the i-th customer comes. He can either:
  - sit a table $1 \leq k \leq T$ with probability $\frac{n_k}{\alpha_0+i-1}$, and one set $X_i = X_k^*$.
  - sit at a new table with probability $\frac{\alpha_0}{\alpha_0+i-1}$, and one increases T to $T+1$, draw $X_T^* \sim H$ and set $X_i = X_T^*$.

Figure 3.3 gives an illustration of tables an customers in the CRP.

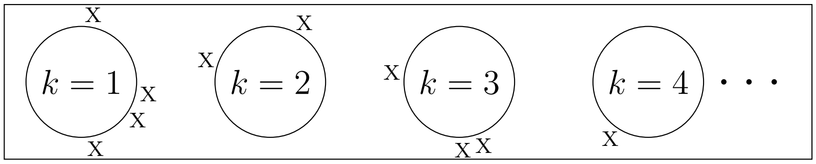

Figure 3.3: Illustration of a Chinese Restaurant Process. Each customer (X) can seat at a table k or start on new one according to a rich-get-richer phenomenon.

In addition, the CRP gives hint about the distribution of the number of clusters (i.e. tables) depending on the number of data n (i.e. customers). Let m be the number of tables generated by the DP. From the CRP, one knows that the probability of generating a new table for each draw is $\frac{\alpha_0}{\alpha_0+i-1}$, and so is independent of the previous number of tables. Thus, the mean and variance of the number of tables for n draws are

$$\begin{aligned}\mathbb{E}[m|n] &= \sum_{i=1}^{n} \frac{\alpha_0}{\alpha_0 + i - 1} = \alpha_0(\psi(\alpha_0 + n) - \psi(\alpha_0)) \\ &\approx \alpha_0 \log(1 + \frac{n}{\alpha_0}) \qquad \text{for } N, \alpha_0 \gg 0\end{aligned}$$

$$\begin{aligned}\mathbb{V}[m|n] &= \sum_{i=1}^{n} \frac{\alpha_0}{\alpha_0 + i - 1} = \alpha_0(\psi(\alpha_0 + n) - \psi(\alpha_0)) + \alpha_0^2(\psi'(\alpha_0 + n) - \psi'(\alpha_0)) \\ &\approx \alpha_0 \log(1 + \frac{n}{\alpha_0}) \qquad \text{for } N > \alpha_0 \gg 0\end{aligned}$$

where ψ is the Digamma function.

A final note about the CRP is that this process is useful to construct Gibbs Sampler of models that use DP prior, in particular, the CRP equations relate to the Collapse Gibbs Sampling updates for DPs based model as the base measure is marginalized out (Antoniak 1974).

3.3.3 Stick Breaking Process

The Stick-Breaking construction goes beyond the previous definition of predictive distribution of the DP. It provides a more general and constructive process to make explicit the random measure G. As mentioned, the DP is a discrete distribution made of weighted sum of point mass such that

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{X_k^*} .$$

The stick proportion π_k for each cluster is then build in a way which can be seen as if one recursively broke a proportion of a stick, according to the following process:

$$\begin{aligned}\beta_k &\sim \text{Beta}(1, \alpha_0) \qquad X_k^* \sim H \\ \pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)\end{aligned}$$

Then $G \sim \text{DP}(\alpha_0, H)$ and we write the Stick-Breaking construction of proportions as $\boldsymbol{\pi} \sim \text{GEM}(\alpha_0)$. This name stands for Griffiths, Engen and McCloskey who discovered it. The Stick-Breaking process has been used to develop and improve inference techniques via Variational Inference and MCMC sampling methods.

3.3.4 Two-parameter extension

The DP has a two-parameter extension called the Pitman-Yor Process (PYP) with discount parameter $0 \leq d \leq 1$ and concentration parameter α_0. If G is a PYP with the given parameters and base measure H, we write $G \sim \text{PYP}(d, \alpha, H)$. When $d = 0$, the PYP reduces to the DP. The stick breaking process for the PYP is generalized as follows:

$$\begin{aligned} \beta_k &\sim \text{Beta}(1-d, \alpha_0 + kd) & X_k^* &\sim H \\ \pi_k &= \beta_k \prod_{l=1}^{k-1}(1-\beta_l) & G &= \sum_{k=1}^{\infty} \pi_k \delta_{X_k^*} \end{aligned}$$

and we refer to the distribution of $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots) = GEM(d, \alpha_0)$.

The CRP for the PYP generalizes to the following urn scheme:

$$X_i | X_1, \dots, X_{i-1}, \alpha, d, H = \frac{\alpha_0 + Kd}{\alpha_0 + i - 1} H + \sum_{k=1}^{K} \frac{n_k - d}{\alpha_0 + i - 1} \delta_{X_k^*}$$

The PYP preserves the clustering and the rich-get-richer properties of the DP. Furthermore, it exhibits a power-law nature as under its Stick-breaking formulation one has $\mathbb{E}[\beta_k] = \mathcal{O}(k^{-1/d})$ if $0 < d < 1$, which means that the cluster sizes decay according to a power-law. This fact makes the PYP often a better choice than the DP to model natural phenomena.

3.3.5 Hierarchical Dirichlet Process

The Hierarchical Dirichlet Process (HDP) is a two-stage DP process (Teh et al. 2006). It allows sharing the same mixture components[5] across the data, which is not possible with a unique DP (i.e. all tables have a different dish in the CRP metaphor.). By overcoming this limitation, the HDP constitutes a flexible prior to build hierarchical models able to capture complex semantic patterns in the data. In his canonical form, the HDP assumes that the observed data are composed of J instances, and each instance is composed of N data point occurrences. For example, in the context of network analysis, the instances would be the nodes and data occurrences the edges. In the context of text analysis (topic modeling), the instances typically represent documents and the data point occurrences the words.

[5] The term component is used in a general case here but depending on the context, the component may be called differently. For instance the term of latent topic is common for text analysis and block, class or communities for network modeling.

In the following, we go through the original HDP model proposed which was used for topic modeling. Notice that the adaption of the HDP for relational models need a slight modification that will be further examined in chapter 4. The generative process of the HDP is as follows:

$$\begin{aligned} G_0 \mid \alpha_0, H &\sim \mathrm{DP}(\gamma, H) \\ G_j \mid \gamma, G_0 &\sim \mathrm{DP}(\alpha_0, G_0) \\ \theta_{ji} \mid G_j &\sim G_j \\ y_{ji} \mid \theta_{ji} &\sim F(\theta_{ji}) \end{aligned}$$

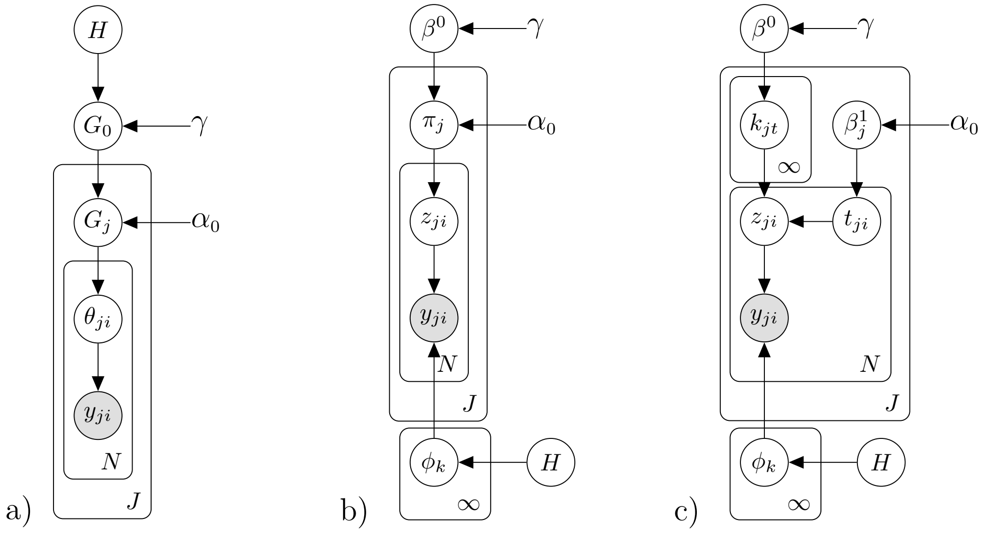

Figure 3.4: Graphical models for the HDP: a) The most compact way of representing the generative model. b) the first DP level is unwrap using the Stick-Breaking construction. c) both DPs level are unwrap.

In order to capture in a more constructive way the distributions of the variables involved in the process, the levels of DPs can be decomposed or unwrap, by using their Stick-Breaking constructions. The first level of DP is rewritten as follows:

$$\boldsymbol{\beta}^0 \sim \mathrm{GEM}(\gamma) \qquad \phi_k \sim H$$
$$G_0 = \sum_{k=1}^{\infty} \beta_k^0 \delta_{\phi_k}$$

This first DP level represents the shared components ϕ_k across the observed data of the HDP with the proportion $\boldsymbol{\beta}^0$. Then, one can expess the data instance's distribution of components using π_j as a function of a DP parameterized by $\boldsymbol{\beta}^0$.

Thus, the generative process at the instance level becomes:

$$\begin{aligned}\pi_j &\sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta}^0)\\ z_{ji} &\sim \mathrm{Mult}(\pi_j)\\ y_{ji} &\sim F(\phi_{z_{ji}})\end{aligned}$$

The second level of DP corresponds to π_j. It defines the mixture of the shared components at the instance level of the data. The latent variables can again be expressed using the Stick-Breaking construction of the DP:

$$\begin{aligned}k_{jt} &\sim \mathrm{Mult}(\boldsymbol{\beta}^0) \qquad & \boldsymbol{\beta}^1_j &\sim \mathrm{GEM}(\alpha_0)\\ t_{ji} &\sim \mathrm{Mult}(\boldsymbol{\beta}^1_j) \qquad & z_{ji} &= k_{jt_{ji}}\end{aligned}$$

The graphical models for the three representations of HDP are shown in Figure 3.4. When F is a Multinomial distribution, the model is sometimes referred to as the HDP-LDA as a generalization of the Latent Dirichlet Allocation (LDA) with a potentially infinite number of topics.

As for the DP, one can obtain closed-form expressions for the predictive distributions of the HDP by marginalizing out the base measures. We discuss this approach in the next section.

3.3.5.1 Chinese Restaurant Franchise

The Chinese Restaurant Franchise (CRF) is a metaphor to describe the process behind draws from the HDP and the form of the associated predictive distributions, which generalize the CRP. The HDP is composed of two levels of DPs. The draws from the first level are associated to dishes shared across a restaurant franchise. At the second level, each restaurant is composed by a possibly infinite number of tables with infinite capacity, and each table is assigned to one dish. Finally, customers who sit at a table, share the same dish. The CRF aims at writing the predictive distribution that a customer sits at a particular table and that a particular table will serve a specific dish. The indexes k, t, i respectively represent the dishes, the tables and the customers of a restaurant. What we called the instance level in the previous section, corresponds to the restaurants of the franchise, and is indexed by j. Table 3.1 presents the different variables involved in the CRF.

Let's denote the marginal count of indexes by a dot. For example, the total

Random Variables	**Description**	**CRF metaphor**
θ_{ji}	Draws from G_j	Customer i in restaurant j.
ψ_{jt}	Draws from G_0, which represent a component for values in θ_j.	Table t in restaurant j.
ϕ_k	Draw from base measure H. Represent the distinct values for ψ_{jt}.	Dish k, shared in all restaurants.
t_{ji}	Index of ψ_{jt} associated to θ_{ji}.	The table taken by customer i in restaurant j.
k_{jt}	Index of ϕ_k associated to ψ_{jt}.	Dish ordered by table t in restaurant j.
m_{jk}	The number of times ψ_{jt} takes value in ϕ_k.	Number of tables that ordered dish k in restaurant j.
n_{jtk}	The number of times θ_{ji} takes values in ϕ_k for ψ_j index at t.	Number of customer in restaurant j, at table t, eating dish k.

Table 3.1: Random variables involved in the Chinese Restaurant Franchise.

number of tables is denoted by $m_{\cdot\cdot}$ and the total number of customers sitting at a table t in restaurant j by $n_{jt\cdot}$. In this setting, we can write the predictive distributions for θ_{ji} and ψ_{jt}, where respectively G_j and G_0 are integrated out, following Eq. 3.4, as:

$$\theta_{ji} \mid \theta_{j1}, \ldots, \theta_{j,i-1}, \alpha_0, G_0 \sim \frac{1}{i-1+\alpha_0} \left(\alpha_0 G_0 + \sum_{t=1}^{m_{j\cdot}} n_{jt\cdot} \delta_{\psi_{jt}} \right)$$

$$\psi_{jt} \mid \psi_{-jt}, \gamma, H \sim \frac{1}{m_{\cdot\cdot} + \gamma} \left(\gamma H + \sum_{k=1}^{K} m_{\cdot k} \delta_{\phi_k} \right)$$

The predictive distribution of θ_{ji} and ψ_{jt} have thus a closed-form expression and constitute the starting point to develop inference scheme for a practical usage of the HDP.

3.3.5.2 Inference

Several inference methods have been proposed for the HDP in the literature that mostly, either rely on Markov Chain Monte Carlo method (MCMC) or Variational Inference. In their seminal paper (Teh et al. 2006), the authors propose different sampling schemes. In this section, we give the main results needed to derive the Gibbs updates for the model parameters.

Sampling by Direct Assignment

In this sampling scheme, based on the CRF, and who is akin to a Collapse Gibbs Sampling (CGS), we aim to sample iteratively, for the observation (ij), the component assignment z_{ij} (the dish chosen by the customer) given all the others data assignments. Furthermore, one also needs to concurrently sample the potential new components, which is achieved through the sampling of the number of tables m_{jk} and auxiliary variable $\boldsymbol{\beta}$. The latter is used to make an explicit construction of G_0. More precisely, as each ψ_{jt} is a draw from G_0, by conditioning it by the ψ_{jt} and exploiting Eq. 3.2 one has

$$G_0 \mid \boldsymbol{\psi}, H, \gamma \sim \mathrm{DP}\left(\gamma + m_{..}, \frac{\gamma H + \sum_{k=1}^{K} m_{.k}\delta_{\psi_k}}{\gamma + m_{..}}\right)$$

To accomplish the construction of G_0, one can resort to an augmented representation in order to have an explicit construction of it (Teh et al. 2006), written as

$$\boldsymbol{\beta} = (\beta_1, \dots, \beta_K, \beta_u) \sim Dir(m_{.1}, \dots, m_{.K}, \gamma) \qquad G_u \sim \mathrm{DP}(\gamma, H)$$
$$G_0 = \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \beta_u G_u$$

Under this representation, by omitting the reference to α_0 and γ, one can write the conditional distribution for a component k:

$$P(z_{ji} = k \mid \boldsymbol{y}, \boldsymbol{z}^{-ji}, \boldsymbol{m}, \boldsymbol{\beta}) \propto \begin{cases} (n_{j.k}^{-ji} + \alpha_0\beta_k) f_k^{-y_{ji}}(y_{ji}) & \text{if } k \text{ previously used,} \\ \alpha_0 \beta_u f_{k^{\text{new}}}^{-y_{ji}}(y_{ji}) & \text{if } k = k^{\text{new.}} \end{cases} \tag{3.5}$$

where $f_k^{-y_{ji}}$ denotes the conditional likelihood of y_{ij} under the component k given all data except y_{ij} such that $f_k^{-y_{ji}} = p(y_{ij} | \boldsymbol{z}^{-ij}, z_{ji} = k)$ and $f_{k^{\text{new}}}^{-y_{ji}}(y_{ji}) = \int f(x_{ij}|\phi)h(\phi)d\phi$ is simply the prior density of y_{ij} and $f(.|\phi)$ and $h(.)$ are respectively the density of $F(\phi)$ and H. The sampling of the table configuration $\boldsymbol{m}$ can be done using the unsigned Stirling of the first kind $s(n, m)$ (Antoniak 1974):

$$P(m_{jk} = m \mid \boldsymbol{z}, \boldsymbol{m}^{-jk}, \boldsymbol{\beta}) = \frac{\Gamma(\alpha_0\beta_k)}{\Gamma(\alpha_0\beta_k + n_{j.k})} s(n_{j.k}, m)(\alpha_0\beta_k)^m$$

We notice that if $h(.)$ and $f(.)$ are usually chosen to be conjugate because it allows then to obtain a closed-form for the updates of the components assignment given in Eq. 3.5. The given updates complete the sampling procedure of the CRP since θ_{ji} and ψ_{jt} can be reconstructed from their index variables.

Optimization of concentration parameter

In the CRF, G_0 and G_j have been integrated out, thus the component assignment are only conditioned by the base measure H and the concentration parameters γ

and α_0. One way to optimize those concentration parameters, proposed by (Teh et al. 2006), is to use auxiliary variable sampling method (Escobar & West 1995). In this scheme, auxiliary variables u and v are introduced, and we assume that priors for concentration parameters are gamma distributed.

Keeping the CRF notations and given the tables configuration $m_{j\cdot}$ and the client configuration $n_{j\cdot\cdot}$ in the CRF, we have for the parameter α_0, governing the number of tables $m_{\cdot\cdot}$, the following posterior distribution:

$$\begin{gathered}
\alpha_0 \sim \mathcal{G}(a_\alpha, b_\alpha) \\
u_j \sim \text{Bernoulli}(\frac{n_{j\cdot\cdot}}{n_{j\cdot\cdot} + \alpha_0}), \quad v_j \sim \text{beta}(\alpha_0 + 1, n_{j\cdot\cdot}) \\
\alpha_0 \mid u_j, v_j \sim \mathcal{G}(a_\alpha + m_{\cdot\cdot} - \sum_j u_j, b_\alpha - \sum_j \log v_j)
\end{gathered}$$

And similarly, for the parameter γ governing the number of classes K, we obtain:

$$\begin{gathered}
\gamma \sim \mathcal{G}(a_\gamma, b_\gamma) \\
u \sim \text{Bernoulli}(\frac{m_{\cdot\cdot}}{m_{\cdot\cdot} + \gamma}), \quad v \sim \text{beta}(\gamma + 1, m_{\cdot\cdot}) \\
\gamma \mid u, v \sim \mathcal{G}(a_\gamma + K - 1 + u, b_\gamma - \log v)
\end{gathered}$$

3.3.6 Indian Buffet Process

The Indian Buffet Process (IBP) is a stochastic process analogous to the CRP but, instead of being a prior over exchangeable partition, the IBP is a prior over sparse binary matrices (Griffiths & Ghahramani 2011). Let F be a binary matrix of size $N \times K$ drawn from an IBP with a hyper-parmeter α. The probability distribution of F can be derived from the following process; imagine an Indian restaurant with an infinite number of dishes, and where N customers enter one after another, and let the entry $f_{ik} = 1$ if customer i selects dish k:

- The first customer starts selecting dishes, and stops after having selected $\text{Poi}(\alpha)$ dishes,
- The i-th customer comes, and starts selecting dishes with probability $\frac{m_k}{i}$, where m_k is the number of times dish k has been selected $m_k = \sum_{i=1}^{N} f_{ik}$. When all previously sampled dishes have been tried, he selects $\text{Poi}(\frac{\alpha}{i})$ new dishes.

Each row i of the matrix F obtained can be interpreted as the (latent) “features” of i and, in the sampling process described, one can see that the distribution over row depends on i. Indeed, the new features (dishes) are not ordered arbitrarily and the number of active features increases with i. However, the law of the matrix F generated by the process should be invariant under row-permutations. Therefore, an operation on the matrix F is required to make the matrix independent of the ordering of the rows as well as the columns. This operation consists of finding an equivalence class of all the matrices that are equivalent by a permutation of the columns (note that this also makes it row-exchangeable), and it is called the *left-ordering-form* (*lof*) of the matrix. Under this transformation, the probability of any matrix F of the exchangeable IBP is given by

$$P(F \mid \alpha) = \frac{\alpha^{K_+}}{\prod_{h=1}^{2^N-1} K_h!} \exp(-\alpha H_N) \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!} \tag{3.6}$$

where H_N is N-th harmonic number $H_N = \sum_{j=1}^{N} \frac{1}{j}$, K_h denotes the number of features (dishes) having history h[6] and $K_+ = \sum_{h=1}^{2^N-1} K_h$ is the number of features for which $m_k > 0$.

We mention here some additional important properties of the IBP (Tutorial 2012):

- The IBP gives birth to a rich-get-richer phenomenon at the feature level; the more a feature is active, the more it will be.
- The distribution of the number of active features in F is $K_+ \sim \text{Poi}(\alpha H_n)$.
- The distribution of the total of non-zeros entries in F is $\text{Poi}(\alpha N)$.

A constructive approach of the IBP is obtained through the infinite limit of a Beta-Bernoulli process. Let F defined by the following process:

$$\pi_k \sim \text{Beta}(\frac{\alpha\beta}{K}, \beta),$$
$$f_{ik} \mid \pi_k \sim \text{Bernoulli}(\pi_k)$$

When $K \to \infty$, one can show that $P(F|\alpha)$ is consistent with equation 3.6, modulo the size of the equivalence class (under the *lof* transformation).

[6] A feature k has history h if $\sum_{i=1}^{N} f_{ik} 2^{i-1} = h$.

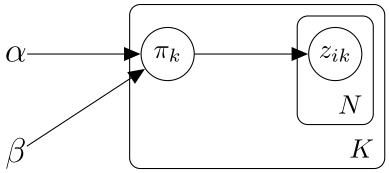

Figure 3.5: Graphical model for beta-bernoulli process

3.3.6.1 Inference

Gibbs Sampler

Developing MCMC sampler for the IBP is much simpler than for the DP. Indeed, one can show that the update rules for the entries of the matrix F are given by:

$$\begin{aligned} P(f_{ik} = 1 \mid F^{-(ik)}) &= \frac{m_{-i,k}}{N} \\ P(f_{ik} = 0 \mid F^{-(ik)}) &= 1 - \frac{m_{-i,k}}{N} \end{aligned}$$

After having sampled all the features for given row i of the matrix, one samples new features from $\mathrm{Poi}(\frac{\alpha}{N})$.

Optimizing α

In order to learn the hyper-parameter α of the IPB prior, controlling the speed of growth of the feature matrix, we can put a conjugate prior on it. A Gibbs sampling step can then be inserted in the sampling loop following the approach used in (Görür et al. 2006).

We know from the IPB that the probability of generating a feature matrix F is

$$P(F \mid \alpha) \propto P(\alpha \mid F) P(F)$$

Then, one can isolate the part of the equation depending only on α to be

$$P(F \mid \alpha) \propto \alpha^{K_+} \exp(-\alpha H_N) \propto \mathrm{Gamma}(1 + K_+, 1/H_N)$$

where $\mathrm{Gamma}(x, y)$ the Gamma distribution with shape x and scale y. Finally, if we suppose that $\alpha \sim \mathrm{Gamma}(a, b)$, for given hyper-parameter a and b, we obtain the following posterior distribution for α:

$$P(\alpha \mid F) = \mathrm{Gamma}(a + K_+, \frac{1}{b + H_N}) \tag{3.7}$$

Two-parameter extension

A notable limitation of the standard IPB is that the dimensionality of the matrix F and its sparsity are coupled through α. The two-parameter extension of the IBP adds a parameter β to be able to control the dimensionality K and the sparsity of F independently. In the constructive process of the IBP, the sampling step from the Beta distribution is changed to $\pi_k \sim \text{Beta}(\frac{\alpha\beta}{K}, \beta)$. Therefore, the two-parameter IBP sampler is impacted as follows:

1. A feature is activated with probability $\frac{m_k}{\beta+N-1}$,
2. New columns/features are draw from $\text{Poi}(\frac{\alpha\beta}{\beta+N-1})$.

It follows that the expected number of non-zeros entry per row is still $\text{Poi}(\alpha)$ but, the distribution of active features becomes $K_+ \sim \text{Poi}(\alpha \sum_{i=1}^{N} \frac{\beta}{\beta+i-1})$.

3.4 Random graph models

Graph theory is historically concerned by the study of graphs with well-established structure such as regular graphs or planar graphs (Albert & Barabási 2002). In random graph theory, the motivation is to discover properties satisfied by all graphs that can be generated from limited design assumptions i.e. a random graph model. The field has been popularized by a series of papers (ERDdS & R&WI 1959; Erdos & Rényi 1960; Erdős & Rényi 1961). They proposed a very simple model, namely the Erdős-Rényi (ER) model, and yet discovered meaningful properties that largely inspired the community.

3.4.1 ER Model

In the ER model, an undirected graph with N nodes is generated by connecting each node with a probability p, thus, one has $y_{ij} \sim \text{Bern}(p)$ and the class of possible graph generated is referred to as $G_{N,p}$. It can be shown that the distribution of a degree d of a randomly chosen node has a closed-form expression such that

$$P(d = n) = \binom{N}{n} p^n (1-p)^{N-n} \ .$$

A case of interest is for the so-called large graph, when $N \to \infty$. In this case, the degree distribution converges to a Poisson law as $P(d = n) \approx \text{Poi}(n; z)$ where

$z = p(N-1)$ is the mean degree[7]. The class of graphs $G_{N,p}$ generated by the ER model are consequently sometimes referred to as the *Poisson random graph*. This result leads to another interesting property of the ER model wich is worth mentioning; let's define a *component* as a maximal subset of nodes that can all be reached from another in the same subset (through edge traversal). There is a so called *phase transition*, from low value of p, where there are many small components with exponential size distribution, to high value of p with very few small components and one, so-called, *giant component*. A consequence of this is that the ER model satisfies the small world effect as its typical distance between two nodes is $l = \frac{\log N}{\log z}$ (Bollobás 1998). Nevertheless, the ER model do not satisfy all the other properties found in real-world networks. Its degree distribution is Poisson and thus has exponential decay, it has random mixing patterns and no clustering structure. Though, it is not adapted for modeling real systems, the ER model still gives insight on the way network can behave and constitutes a baseline regarding the emergence of phase transitions and giant components that are also studied in other random graph models.

Among the many extensions of the ER model proposed in the literature (Goldenberg et al. 2010), we focus here on the Stochastic Block Model that provides a very general framework to model graph with community structure.

3.4.2 Stochastic Block Model

The Stochastic Block Model (SBM), originally proposed in (Holland et al. 1983), has since been extensively covered in the literature[8]. It is a random graph model where nodes belong to some latent blocks (or communities) and the probability that two nodes bind depends only on the membership of the nodes to blocks. Let N be the total number of nodes and K the total number of blocks. Let the model parameters $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ and $(\phi_{kk'})_{K\times K}$ be respectively a probability vector and a probability matrix (or weight matrix). The generative process of the SBM is then as follows:

$$\begin{aligned} c_i &\sim \mathrm{Cat}(\boldsymbol{\pi}) \qquad && \text{for } i \in \{1, \dots, K\} \\ y_{ij} &\sim \mathrm{Bern}(\phi_{c_i c_j}) \qquad && \text{for } i, j \in \mathcal{V} \times \mathcal{V} \end{aligned}$$

[7] Where the self-loops are not considered.

[8] See http://bactra.org/notebooks/stochastic-block-models.html

As one can notice, if $K = 1$ the model reduces to the ER model, otherwise all the sub-networks, restricting nodes to a single block interaction (inner block interaction if $k = k'$, and outer block interaction if $k \neq k'$), locally behave like the ER model. The main challenge in the SBM is to infer a "good" partition of the nodes with the corresponding weight matrix. Note that the combination of possible assignments is K^N. The optimization of parameters is generally accomplished with a variant of the EM algorithm (Latouche et al. 2012). Recently, efficient inference based on MCMC has also been proposed, with revisited versions of the SBM (Peixoto 2017).

While the SBM is a strong baseline, it has several limitations due to the lack of prior over its parameters. In particular, the assumption of a fixed number of blocks is difficult to justify and a bad choice of K can lead to sub-optimal solutions. This problem has been addressed by using DP prior over the node assignment vector. The resulting model is referred to as the Infinite Relational model (IRM) (Kemp et al. 2006). Another limitation is the hard assignment of nodes to blocks, which may not be adapted to capture the diversity of interaction in real-world networks. Those limitations have been addressed in several directions that we will explore in the rest of the manuscript.

3.4.3 Mixed-Membership Models

As mentioned, a limitation of the SBM is that each node belongs to only one latent block. This assumption is often considered too restrictive for modeling complex network and complex system in general. More expressive models can be built by relaxing the hard assignment and instead allowing the nodes to belong to several latent blocks. The edge likelihood can hence be viewed as a mixture model over the multiple node memberships to the latent blocks (akin to overlapping communities and soft clustering topic). Mixture models have a long history that is related to matrix factorization used to decompose a data matrix into latent factors (Buntine & Jakulin 2005). A whole framework named *mixed-membership model*, has emerged to study and generalize mixture models in the case where the latent variables can themselves be shared among data instances, which have found many successful applications reviewed in (Airoldi et al. 2014). In this setting, we are particularly interested in two subclasses of mixed-membership model for networks analysis, referred to as *latent class model* and *latent feature model*. Many models have been proposed in this direction, though we do not intent to go into the details of each of them as it may not be relevant, we provide in Table 3.2 a comprehensive comparison

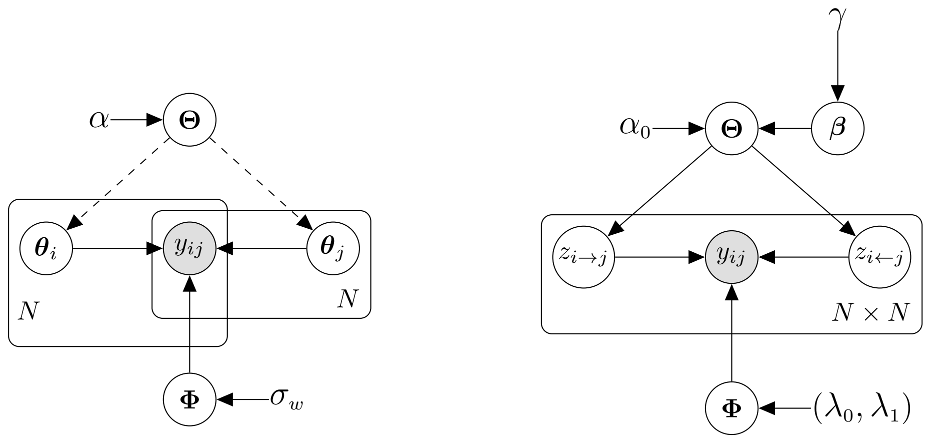

Figure 3.6: The two graphical representations of (left) the latent feature model (ILFM) and (right) the latent class model (IMMSB). The difference between the two graphical structures of models lies in the way representations are associated to nodes: a fixed representation is used in the case of the latent feature model, whereas the representation in the latent class model varies according to the link considered.

of these models regarding their modeling assumptions. Generally, in latent class models, one supposes that the nodes belong to some latent communities on which depend the mixing pattern of the graph. Whereas, for latent feature model, one supposes that the nodes own some latent feature, which control the mixing pattern. Interestingly, models that combine both aspects have been proposed, following (Mackey et al. 2010).

We are particularly interested in two two general representatives of the latent class model and latent feature model that we further describe and study in chapter 4. That is the IMMSB and ILFM model, that both allow overlapping communities with a possibly infinite number of communities, the former based on the HDP and the latter on the IBP. The graphical representations for both models are given in Figure 3.6.

The two type of models can be expressed in a common framework. Let $\Theta = (\theta_{ik})_{N\times K}$ and $\Phi = (\phi_{kk'})_{K\times K}$ be two random matrices with N be the number of nodes and K the dimension of the latent space. The edge likelihood is then parameterized by a bilinear product such that

$$y_{ij} \sim \text{Bern}(f(\theta_i \Phi \theta_j^T)) \tag{3.8}$$

where f is a bijective function to map the support of the bilinear product to a probability space if necessary, otherwise f is the identity.

One can easily see that this representation encompasses the ER and SBM models. Precisely, the only difference between those random graph models, including the difference between latent class and latent feature models, within this representation is the prior knowledge and parameters space of Θ and Φ:

- if $K = 1$ and $\phi_{11} = p$ one falls onto the ER model.
- if $\theta_i \sim \text{Mult}(1, \pi_k)$ and $\phi_{kk'} \sim \text{Uniform}[0, 1]$ with given hyper-parameter $\pi = (\pi_1, \dots, \pi_k)$ one falsl onto the SBM model. Additionally, if $\pi_k \sim GEM(\gamma)$, the model is equivalent to the Infinite Relational Model (IRM), where the number of class k can vary.
- if $\theta_i \sim \text{Dir}((\alpha_1, \dots, \alpha_K))$ and $\phi_{kk'} \sim \text{Beta}(a, b)$ the model is equivalent to the Mixed-Membership Stochastic Blockmodel (MMSB).[9]
- if $\Theta \sim \text{IBP}(\alpha)$ and $\phi_{kk'} \sim \text{Normal}(0, \sigma)$ with the function f being the sigmoid, then the model is the Infinite Latent Feature Model (ILFM).
- etc.

Type	Model	Observations	Prior	Mixed-membership	Generalize
Latent class	SBM (Holland 83)	Bernoulli	Multinomial	no	ER (Erdos 59)
	IRM (Kemp 06)	Bernoulli	DP	no	SBM
	IHRM (Xu 06)	Bernoulli	DP	no	IRM
	MMSB (Airoldi 09)	Bernoulli	Multinomial-Dirichlet	yes	SBM
	IMMSB (Kim 12)	Bernoulli	HDP	yes	MMSB, IRM
Latent feature	LFM (Ghahramani 95)	Gaussian	-	yes	CVQ (Ackley 95)
	LFL (Menon 10)	Bernoulli	-	no	softmax
	ILFM (Miller 09)	Bernoulli	IBP	yes	IRM
	IMRM (Morup 11)	Bernoulli	IBP	yes	ILFM
	BPM (Palla 12)	Bernoulli	IBP	yes	IRM

Table 3.2: Comparison of latent class and feature models found in the literature.

3.4.4 Representation theorem for exchangeable graphs

In section 3.3, we mentioned the concept of exchangeability of a random sequence and illustrated how it is related to the construction of the Dirichlet Process. What we have learned, is that the exchangeability assumption over a sequence of observable data is equivalent to the existence of an integral decomposition (a mixture) of the probability density of this sequence under which the observations are i.i.d given a random probability measure. This result is known as the de Finetti' theorem and constitutes a justification for the existence of latent variable models under the exchangeability assumption. An interesting question to ask, though is if an equivalent representation theorem exists for exchangeable random graphs. The

[9] Proof: $P(y_{ij}|\Theta, \Phi) = \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} P(y_{ij}|\Phi, z_{i\rightarrow j} = k_1, z_{i\leftarrow j} = k_2) P(z_{i\rightarrow j} = k_1|\theta_i) P(z_{i\leftarrow j} = k_2|\theta_j) = \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \phi_{k_1k_2}\theta_{ik_1}\theta_{jk_2} = \theta_i^T \Phi \theta_j$.

answer is yes, and it is known as the Aldous-Hoover theorem, that we shall recall here (Orbanz & Roy 2015). This theorem has a version adapted for bipartite graphs, but we will focus here on unipartite graph and thus square adjacency matrix.

Let's consider an undirected graph and its adjacency matrix Y (an array) of infinite size.

Definition 3.4.1 (Jointly exchangeable array) *A random array $(y_{ij})_{i,j\in\mathbb{N}}$ (denoted simply (y_{ij}) for short) is called jointly exchangeable if*

$$P((y_{ij})) = P((y_{\pi(i)\pi(j)}))$$

holds for every permutation π of $\mathbb{N}$.

As for exchangeable sequences, exchangeable graphs mean that the probability of a graph should not depend on the order in which we observe the data.

Theorem 3.4.1 (Aldous-Hoover for jointly exchangeable array) *Let $\boldsymbol{Y}$ be a sample space. A random array $(y_{ij})_{i,j\in\mathbb{N}}$ is jointly exchangeable if and only if it can be represented as follows: There is a random measurable function $F : [0,1]^3 \to \boldsymbol{Y}$ such that*

$$P((y_{ij})) = P((F(U_i, U_j, U_{\{i,j\}})))$$

where $(U_i)_{i\in\mathbb{N}}$ and $(U_{\{i,j\}})_{i,j\in\mathbb{N}}$ are, respectively, a sequence and an array of i.i.d $\mathrm{Uniform}[0,1]$ *random variables.*

In Bayesian language, the theorem states that there is a prior distribution μ over measurable functions such that an exchangeable graph is always generated by a model of the form

$$\begin{aligned} F &\sim \mu \\ U_i &\sim \mathrm{Uniform}[0,1] \qquad \forall i \in \mathbb{N} \\ U_{\{i,j\}} &\sim \mathrm{Uniform}[0,1] \qquad \forall i,j \in \mathbb{N} \\ y_{ij} &:= F(U_i, U_j, U_{\{i,j\}}) \end{aligned}$$

This powerful theorem again gives a justification for the use of latent variables and most important, the form of their priors (and a model is entirely determined by the choice of the prior on F) for exchangeable graphs. Although intuitive, it is not straightforward to prove that the representation for mixed-membership models given in Eq. 3.8 generates exchangeable graphs, it appears that it is a special case of theorem 3.4.1, which has been shown in (Aldous 1981) and (Kallenberg 2006). It derives from the fact that the edge probabilities are conditionally independent and that the nodes are associated to i.i.d random variables. In particular, the exchangeability of IRM, IMMSB and ILFM models are also illustrated in (Orbanz & Roy 2015). Note that the Aldous-Hoover theorem also generalizes for higher dimensional arrays (i.e. tensors), which is akin to the representation of exchangeable multi-relational graphs.

A notable corollary of the theorem 3.4.1 is that exchangeable graphs are either dense (i.e. the number of edges growth quadratically with N) or empty since their expected number of edges is independent of N. This may seem like a misspecification for the modeling of real-world networks. In response to that, the study of representation for graphs in the sparse regime has emerged as an active and growing field of research (Veitch & Roy 2015; Caron & Fox 2017; Le et al. 2015; Bollobás & Riordan 2011; Borgs et al. 2014).

3.5 Summary

The chapter presented the mathematical framework underlying the latent variable models for complex networks. We recalled some fundamental results and notions of probability theory and exposed a flexible class of Bayesian nonparametric priors. We also presented the class of latent variable models (Mixed-Membership model) that will be studied in the two next chapters, and we provided a modest literature survey. We finally presented the theoretical foundation (through representation theorem) that justify the construction and exposed the limitations of this class of models.

Chapter 4

Stochastic Mixed Membership Models and Their Properties

4.1 Introduction

Several powerful relational learning models have been proposed to solve the problem commonly referred to as *link prediction* that consists in predicting the likelihood of a future association between two nodes in a network (Liben-Nowell & Kleinberg 2007; Hasan & Zaki 2011). Among such models, the class of stochastic mixed membership models has received much attention as such models can be used to discover hidden properties and infer new links in social networks. Two main models in this class have been proposed and studied in the literature: the latent feature model (Meeds et al. 2006) and its non-parametric extension (Miller et al. 2009), and the mixed-membership stochastic block model (Airoldi et al. 2009), and its non parametric extension (Koutsourelakis & Eliassi-Rad 2008; Fan et al. 2013). These models fall in the category of mixed-membership models that encompass a wide range of models (such as admixture and topic model) able to learn complex patterns from structured data (Airoldi et al. 2014).

Nevertheless, although drawn from a wide range of domains, real-world social networks exhibit general properties, as exposed in section 2.3, and one can wonder if these models are able to capture these properties. In this work, we focus on the *homophily* and the *preferential attachment* effect (Newman 2010; Barabási 2003). Homophily is verified in a network when similar vertices tend to be more connected than dissimilar ones. On the other hand, preferential attachment states that a nodes is more likely to create connections with nodes having many connections. In graph theory, preferential attachment is used to explain the emergence of scale-free networks that are characterized by a power-law degree distribution. In social network analysis, the interest of these properties has been widely emphasized notably for modeling networks but also for improving the results obtained in classical tasks such as community detection or link prediction. The aim of our study is to assess to which extent stochastic mixed membership models comply with those properties

The remainder of the chapter is organized as follows: Section 4.2 discusses related work. Section 4.3 describes the two main stochastic mixed membership models used for link prediction in social networks and the settings in which they are used. Section 4.4 and 4.5 respectively introduces formal definitions of homophily and preferential attachment in a probabilistic settings and studies how stochastic mixed membership models relate to them. Section 4.6 illustrates the theoretical results on two synthetic and two real networks and Section 4.7 concludes the study.

4.2 Related work

Recently, the class of stochastic mixed membership models have been successfully used for link prediction and structure discovery in social networks. For example, in (Gopalan & Blei 2013), the authors propose an adaptation of mixed-membership stochastic block model (MMSB) called a-MMSB, where "a" stands for assortative, and they use it for discovering overlapping communities in large networks having millions of nodes. The weight matrix is constrained to have a fixed small value outside its diagonal. A non parametric dynamic version of MMSB model has also been introduced to handle temporal networks (Fan et al. 2013). The latent feature model (LFM) has also been extended in several ways, to handle non-negative weights in (Mørup et al. 2011) and with a more subtle latent feature structure in (Palla, Knowles, et al. 2012). Nevertheless, the characterization of these models with regards to the properties of the networks remains to be explored, with several challenges and oppurtunities as mentioned in (Jacobs & Clauset 2014).

In this chapter, we focus on two important properties of social networks, namely *homophily* and *preferential attachment* (Newman 2010; Barabási 2003). Those property has been emphasized in previous studies, for example for modeling and generating artificial networks reflecting properties of real networks, as in the model by Barab‘asi-Albert (Albert & Barabási 2002), the model by Buckley and Osthus (Buckley & Osthus 2001), which integrates a preferential attachment mechanism, or in the Dancer model for generating dynamic attributed networks with community structures and homophilic networks (Largeron et al. 2017). Preferential attachment has also been exploited for improving methods for solving classical tasks such as community detection (Ciglan et al. 2013) or link prediction (Zeng 2016). That said, few theoretical works have been conducted to study to what extent stochastic models comply with this property. Orbanz and Roy pointed out that models belonging to the family of infinitely exchangeable Bayesian graph models cannot generate sparse networks and are thus less compatible with power law degree distributions (Orbanz & Roy 2015). Consequently, Lee *et al.* proposed a random network model in order to capture the power law typical of the degree distribution in social networks (Lee et al. 2015). However the model remains challenging to use in practice, especially for link prediction, due to the relaxation of the exchangeability assumption.

Concerning the homophily effect, (Hoff 2008) pointed out that the latent eigen model (called MLFM, an extension of LFM) can comply with both homophily and stochastic equivalence in undirected graphs but without providing a formal definitions of these properties. Furthermore, Li *et al.*, suggest that the latent eigen model MLFM fails to model homophily for directed graphs and, for correcting that, designed the GLFM model (Li et al. 2011).

A preliminary version of this study was published in (Dulac et al. 2017). However, the definitions of preferential attachment and local degrees we proposed in this previous work are not entirely satisfying inasmuch as the dynamic aspect of preferential attachment was not taken into account. The definitions we propose here and the developments concerning stochastic block models are new and we believe better founded than in this previous work.

We study, in a theoretical way, how the non-parametric versions of the classical stochastic mixed membership models handle homophily and preferential attachment. For this purpose, we introduce formal definitions of this phenomenon and then study how the models behave with respect to these definitions but, first, we present these models and the settings in which we study their behavior.

4.3 Stochastic Mixed Membership Models

Stochastic mixed membership models are generative models that rely on latent factors (also called latent *classes* or *features*) for modeling relational data such as links in social networks represented by a graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a a set of nodes and $\mathcal{E}$ a set of edges between these nodes.

In the remainder, we denote by N the number of nodes in this graph ($N = |\mathcal{V}|$) and by Y the adjacency matrix of the graph G ($y_{ij} = 1$ if there is a link between nodes i and j, $1 \leq i, j \leq N$; $y_{ij} = 0$ otherwise). Without loss of generality, we assume that the graph is undirected, the directed case being a special case of the undirected one.

Stochastic mixed membership models are characterized by the fact that each node can "belong" to several latent factors, which reflects the fact that each individual usually has several properties, for example can belong to several communities[1]. The relation between a node i and the latent factors is encoded in a vector denoted $\boldsymbol{\theta}_i$, of finite dimension K in standard versions of the models, and of infinite dimension in non-parametric versions. The collection of all vectors $\boldsymbol{\theta}_i$ ($1 \leq i \leq N$) constitutes the factor matrix $\boldsymbol{\Theta}$. Furthermore, a weight matrix $\boldsymbol{\Phi}$ is used to encode the relations between the latent factors.

Stochastic mixed membership models differ on the way the vectors $\boldsymbol{\theta}_i$ ($1 \leq i \leq N$) and the matrix $\boldsymbol{\Phi}$ are generated. As mentioned before, and to be as general as possible, we consider here the non-parametric versions of the latent feature model (Miller et al. 2009), referred to as ILFM, and of the mixed-membership stochastic block model (Koutsourelakis & Eliassi-Rad 2008; Fan et al. 2013), referred to as IMMSB. This leads to a dynamic number of classes that allows the dimensions of the models to grow with the complexity of the data. This is done in practice by the use of non-parametric prior, the Indian Buffet Process (IBP) for ILFM and the Hierarchical Dirichlet Process (HDP) for IMMSB. All our results are nevertheless also valid for the finite versions of these models.

[1] As mentioned in (Goldenberg et al. 2010), the reader should however bear in mind that the notion of latent factors is of stochastic nature and is an approximation of the notions of communities and shared properties.

4.3.1 Infinite Latent Feature Model (ILFM)

In the latent feature model, each node is represented by a finite vector of binary features. The probability of linking two nodes is then based on a weighted similarity between their feature vectors, the weight matrix being generated according to a normal distribution. In its non-parametric version ILFM, the feature vectors are now generated according to an IBP, leading to feature vectors of infinite dimensions (even though only a finite number of dimensions is actually active). The following steps summarize this process:

1. Generate a feature matrix $\mathbf{\Theta}_{N\times\infty}$ representing the feature vector of each node:

$$\mathbf{\Theta} \sim \text{IBP}(\alpha)$$

2. Generate a weight matrix for each latent feature:

$$\boldsymbol{\phi}_{mn} \sim N(0, \sigma_w),\ m, n \in \mathbb{N}^{+*}$$

3. Generate or not a link between any node i and any node j according to:

$$y_{ij} \sim \text{Bern}(\sigma(\boldsymbol{\theta}_i \mathbf{\Phi} \boldsymbol{\theta}_j^\top))$$

where $^\top$ dentotes the transpose and $\sigma()$ is the sigmoid function, mapping $[-\infty, +\infty]$ values to [0,1], and where y_{ij} is a binary variable indicating that a link has been generated ($y_{ij} = 1$) or not ($y_{ij} = 0$). We will denote by $\boldsymbol{Y}$ the $N \times N$ matrix with elements y_{ij}. Finally, $\boldsymbol{\theta}_i$ denotes the row feature vector corresponding to the i^{th} row of $\mathbf{\Theta}$.

This model makes use of two real hyper-parameters, one for the IBP process (α), and one for the variance of the normal distribution underlying the weight matrix (σ_w). In the case of undirected networks, the matrices $\boldsymbol{Y}$ and $\mathbf{\Phi}$ are symmetric and only their upper (or lower) diagonal parts are generated. Lastly, both $\mathbf{\Theta}$ and $\mathbf{\Phi}$ are infinite matrices. In practice however, one always deals with a finite number of latent features.

4.3.2 Infinite Mixed-Membership Stochastic Blockmodel (IMMSB)

The MMSB model generates class membership distributions per node on the basis of a Dirichlet distribution. Then, for each connection between two nodes, a particular class for each node is first sampled from the class membership distribution, and the probability of connecting the two nodes is, as in the previous model, based on a Bernoulli distribution integrating the weight of the two classes.

The non-parametric version IMMSB parallels this development but considers, in lieu of the Dirichlet distribution, a Hierarchical Dirichlet Process, leading to the following generative model:

- Generate the class membership distributions $\boldsymbol{\Theta}_{N\times\infty}$:

  $$\begin{aligned} \boldsymbol{\beta} &\sim \mathrm{GEM}(\gamma) \\ \boldsymbol{\theta}_i &\sim \mathrm{DP}(\alpha_0, \beta) \quad \text{for } i \in \{1, \dots, N\} \end{aligned}$$

  where GEM (named after Griffiths, Engen and McCloskey) denotes the Stick Breaking Process distribution over the set of natural numbers and DP a Dirichlet Process (Teh et al. 2006);
- Generate a weight matrix for each latent class from i.i.d Beta distribution:

  $$\phi_{mn} \sim \mathrm{Beta}(\lambda_0, \lambda_1),\ m, n \in \mathbb{N}^{+*}$$

- For any node i and any node j, choose a class from their class membership distribution according to a Categorical distribution and generate or not a link according to a Bernoulli distribution:

  $$\begin{aligned} z_{i\rightarrow j} \sim \mathrm{Cat}(\boldsymbol{\theta}_i)\ , \quad z_{i\leftarrow j} \sim \mathrm{Cat}(\boldsymbol{\theta}_j) \\ y_{ij} \sim \mathrm{Bern}(\phi_{z_{i\rightarrow j} z_{i\leftarrow j}}) \end{aligned}$$

We have this time four real hyper-parameters, two for the Hierarchical Dirichlet Process (γ and α_0) and two for the Beta distribution underlying the weight matrix (λ_0 and λ_1). As for the previous model, in the case of undirected networks, the matrices $\boldsymbol{Y}$ and $\boldsymbol{\Phi}$ are symmetric and only their upper (or lower) diagonal parts are generated; as before again, both $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are infinite matrices.

4.3.3 Settings

In a Bayesian context, the set of hyper-parameters underlying the model considered is known. This set, denoted $\mathcal{M}_g$, respectively corresponds to α and σ_w for ILFM and to γ, α_0, λ_0 and λ_1 for IMMSB. For mixed membership models, the evidence $P(Y|\mathcal{M}_g)$ has no closed-form solution. Yet, the random graph G is exchangeable so that, for any permutation π on integers, one has:

$$P((y_{ij})_{i,j \in \mathcal{R}}|\mathcal{M}_g) = P((y_{\pi(i)\pi(j)})_{i,j \in \mathcal{R}}|\mathcal{M}_g)$$

and one can generate networks from $\mathcal{M}_g$ by following the generative processes described above for ILFM and IMMSB. In this setting, the question we ask ourselves is whether the networks generated from $\mathcal{M}_g$ comply with the homophily and preferential attachment effect.

However, the typical use of the above models corresponds to the scenario in which some observations (*i.e.* an existing network, observed till a certain time) are available and are used to estimate $\mathbf{\Theta}$ and $\mathbf{\Phi}$ from which new links are created. The estimation of $\mathbf{\Theta}$ and $\mathbf{\Phi}$ is based on:

$$P(\Theta, \Phi|\ Y, \mathcal{M}_g) = \frac{P(Y|\Theta, \Phi)P(\Theta|\mathcal{M}_g)P(\Phi|\mathcal{M}_g)}{P(Y|\mathcal{M}_g)} \tag{4.1}$$

and usually makes use of standard Gibbs sampling and Metropolis-Hastings algorithms[2].

In the remainder, we denote by $\hat{\mathbf{\Theta}}$ and $\hat{\mathbf{\Phi}}$, for both ILFM and IMMSB, the estimates of $\mathbf{\Theta}$ and $\mathbf{\Phi}$ obtained from $\mathcal{M}_g$ and Y, and furthermore set $\mathcal{M}_e = \{\hat{\mathbf{\Theta}}, \hat{\mathbf{\Phi}}\}$. Whether, from the learned parameters $\hat{\mathbf{\Theta}}$ and $\hat{\mathbf{\Phi}}$, the new links generated produce networks that comply with the preferential attachment effect is the second question we ask ourselves in this study.

We now propose a formalization of homophily and preferential attachment in social networks and answer the above questions.

[2] We do not detail the inference of $\hat{\mathbf{\Theta}}$ and $\hat{\mathbf{\Phi}}$ here and refer the interested reader to (Miller et al. 2009; Griffiths & Ghahramani 2011; Teh et al. 2006; Fan et al. 2013).

4.4 Homophily

Homophily refers to the tendency of individuals to connect to similar others; two individuals (and thus their corresponding nodes in a social network) are more likely to be connected if they share common characteristics (McPherson et al. 2001; Lazarsfeld et al. 1954). The characteristics often considered are inherent to the individuals and may represent their social status, their preferences or their interests. A related notion is the one of assortativity, that is slightly more general as it applies to any network, and not just social networks, and refers to the tendency of nodes in networks to be connected to others that are similar in some way.

A definition of homophily has been proposed in (La Fond & Neville 2010). However, this definition, which relies on a single characteristic (like age or gender), does not allow one to assess whether latent models for link prediction capture the homophily effect or not. We thus introduce a new definition of homophily:

Definition 4.4.1 (Homophily) *Let $\mathcal{M}_e$ be a probabilistic link prediction model and s a similarity measure between nodes. We say that $\mathcal{M}_e$ is homophilic under the similarity s iff, $\forall (i,j,i',j') \in V^4$:*

$$s(i,j) > s(i',j') \implies P(y_{ij} = 1 \mid \mathcal{M}_e) > P(y_{i'j'} = 1 \mid \mathcal{M}_e)$$

As one can note, this definition directly captures the effect “if two nodes are more similar, then they are more likely to be connected”.

Different similarities can be considered, as long as they are based on the proximity of the properties of the nodes considered. In stochastic mixed membership models, these properties are encoded in the latent factors. Indeed, as mentioned before, the factor matrix $\hat{\mathbf{\Theta}}$ aims at capturing some latent properties of the nodes, whereas the estimated matrix $\hat{\mathbf{\Phi}}$ captures the correlations between these latent properties. One can thus define, on their basis, a “natural” similarity between nodes as follows:

$$s_n(i,j) = \hat{\boldsymbol{\theta}}_i \hat{\mathbf{\Phi}} \hat{\boldsymbol{\theta}}_j^\top$$

It is straightforward that both ILFM and IMMSB in the setting $\mathcal{M}_e$ are homophilic with respect to s_n. Indeed, $P(y_{ij} = 1 \mid \mathcal{M}_e)$ increases with s_n for ILFM as the sigmoid function is strictly increasing (Eq. 3). Furthermore, marginalizing over the z variables in IMMSB leads to:

$$\begin{aligned}
&P(y_{ij} = 1 \mid \mathcal{M}_e) \\
&\quad = \sum_{k,k'} \hat{\phi}_{k,k'} P(z_{i\to j} = k|\mathcal{M}_e) P(z_{i\leftarrow j} = k'|\mathcal{M}_e) \\
&\quad = \sum_{k,k'} \hat{\phi}_{k,k'} \hat{\theta}_{ik} \hat{\theta}_{jk'} = \hat{\boldsymbol{\theta}}_i \hat{\boldsymbol{\Phi}} \hat{\boldsymbol{\theta}}_j^\top
\end{aligned}$$

Dropping the correlation between latent factors in the natural similarity leads to a new similarity, solely based on the latent factors and defined by $s_l(i,j) = \hat{\boldsymbol{\theta}}_i \hat{\boldsymbol{\theta}}_j^\top$ (s_l stands for latent similarity). With this similarity, however, neither ILFM nor IMMSB are homophilic. Indeed, let us first assume that $\hat{\boldsymbol{\Phi}}$ is null on the diagonal, and strictly positive elsewhere (this can be obtained for both models). For IMMSB, one has:

$$P(y_{ij} = 1 \mid \mathcal{M}_e) = \sum_{k' \neq k} \hat{\theta}_{ik} \hat{\phi}_{kk'} \hat{\theta}_{jk'}$$

as $\hat{\phi}_{kk} = 0$. Let us now consider $\hat{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\theta}}_j = (0,1,0)$ and $\hat{\boldsymbol{\theta}}_{i'} = (0.5, 0, 0.5)$ and $\hat{\boldsymbol{\theta}}_{j'} = (0,1,0)$. Then, $s_l(i,j) = 1$ and $s_l(i',j') = 0$. However, $P(y_{ij} = 1 \mid \mathcal{M}_e) = 0$ whereas $P(y_{i'j'} = 1 \mid \mathcal{M}_e) > 0$. IMMSB is thus not homophilic under s_l. The same example, replacing $\hat{\boldsymbol{\theta}}_{i'} = (0.5, 0, 0.5)$ by $\hat{\boldsymbol{\theta}}_{i'} = (1, 0, 1)$, can be used to show that ILFM is neither homophilic under s_l.

This shows that, for a model to be homophilic, it should be designed according to the similarity at the basis of the proximity between individuals. Both ILFM and IMMSB have been designed on the basis of the natural similarity s_n, and directly encode the fact that similar nodes, according to s_n, are more likely to be connected. It is furthermore possible to make these models homophilic under s_l by imposing constraints on the weight matrix $\boldsymbol{\Phi}$ (and hence its estimate $\hat{\boldsymbol{\Phi}}$); for example, considering positive, diagonal matrices with equal values on the diagonal leads to homophilic models. In that case, the latent factors can be interpreted as community indicators, each community being of equal importance. This is in line with what is done in the study presented in (Gopalan & Blei 2013) to find overlapping communities through assortativity constraints in the mixed membership stochastic block model.

4.5 Preferential attachment

As mentioned before, preferential attachement can be global, in which case nodes are connected across communities, and/or local to the network communities. Preferential attachment is reminiscent of a phenomenon called *burstiness*, studied in different contexts (Barabási 2011). We introduce here definitions for the local and global preferential attachment effects that are extensions of the definitions for burstiness proposed in (Clinchant & Gaussier 2010) for text collections. We will first study global preferential attachment for the models ILFM and IMMSB in the two contexts defined by $\mathcal{M}_g$ and $\mathcal{M}_e$. We will then turn our attention to local preferential attachment.

4.5.1 Global preferential attachment

Probabilistic models naturally lead to the following generative process for creating links between nodes in a network[3]. This process considers all possible pairs of nodes in turn and generates or not a link between them:

1. *For each node* $i \in \{1, \ldots, N\}$,

2. *For each node* $j \in \{1, \ldots, N\}$,

3. *Generate a link between* i *and* j *with probability* $P(y_{ij} = 1|\mathcal{M})$ *where* $\mathcal{M}$ *is either* $\mathcal{M}_e$ *or* $\mathcal{M}_g$.

As one can note, this process considers all nodes in turn, from node 1 to node N. An indexing, *i.e.* a mapping between nodes and integers in $\{1, \cdots, N\}$, is however arbitrary and conclusions drawn from the above process should be independent of the indexing. As we will see, the results we establish below are indeed independent of the indexing.

For a given node i at step p of the above process, p nodes, from node 1 to node p, have been considered and links from these nodes to node i generated or not. We will denote by $d_i^{(p)}$ the degree of node i, i.e. the number of links of node i, at the p^{th} step of this process. By definition:

$$d_i^{(p)} = \sum_{j=1}^{p} y_{ij} \tag{4.2}$$

[3] For simplicity in the notation, we consider that nodes can be linked to themselves. Excluding such links does not raise particular problems.

As mentioned before, preferential attachment characterizes the propensity of nodes in social networks to connect to nodes that already have a lot of connections and can be stated as *the higher the number of links a node has, the more likely it will get new links.* The following definition directly captures this idea:

Definition 4.5.1 (Global preferential attachment) *In the above setting, a probabilistic model satisfies the global preferential attachment effect iff for any indexing, for any node i, $1 \leq i \leq N$, for any p, $1 \leq p < N$, $P(d_i^{(N)} \geq n+1 |\ d_i^{(p)} = n; \mathcal{M})$ increases with n $(1 \leq n < p)$. If $P(d_i^{(N)} \geq n+1 |\ d_i^{(p)} = n; \mathcal{M})$ is independent of n, the model is said to be neutral w.r.t. the global preferential attachment effect. As before, $\mathcal{M}$ is either $\mathcal{M}_e$ or $\mathcal{M}_g$.*

Thus, a model satisfies the global preferential attachment effect if and only if the more links a node i has at some point in the process, the more likely a new link will be created with that node.

For both ILFM and IMMSB, in $\mathcal{M}_e$, the generation of links are independent of each other. The fact that n links have been created after p steps has thus no impact on the future links to a given node. In $\mathcal{M}_g$, as one first needs to generate Θ and Φ prior to generate all the links, a similar behavior is likely to be observed. Intuitively thus, both ILFM and IMMSB are neutral wrt the global preferential attachment effect. The following property formalizes this intuition.

Proposition 4.5.1 *Both ILFM and IMMSB, for both $\mathcal{M}_e$ and $\mathcal{M}_g$, are neutral wrt the global preferential attachment effect.*

Proof 4.5.1 *We first consider model $\mathcal{M}_e$. For any indexing, a node i, $i \leq i \leq N$, and a step p, $1 \leq p < N$. One has, $\forall n, 1 \leq n < p$:*

$$\begin{aligned}
P(d_i^{(N)} \geq n+1 |\ d_i^{(p)} = n, \mathcal{M}_e) &= 1 - P(d_i^{(N)} = n | d_i^{(p)} = n, \mathcal{M}_e) \\
&= 1 - P(y_{i,p+1} = 0, \ldots, y_{iN} = 0 |\ \mathcal{M}_e) \\
&= 1 - \prod_{j=p+1}^{N} P(y_{ij} = 0 |\ \mathcal{M}_e)
\end{aligned}$$

where the last equality comes from the fact that, in $\mathcal{M}_e$, links are independently *generated. Similarly:*

$$\begin{aligned} P(d_i^{(N)} \geq n+2 |\ d_i^{(p)} = n+1, \mathcal{M}_e) &= 1 - \prod_{j=p+1}^{N} P(y_{ij}=0 |\ \mathcal{M}_e) \\ &= P(d_i^{(N)} \geq n+1 |\ d_i^{(p)} = n, \mathcal{M}_e) \end{aligned}$$

which shows that both ILFM and IMMSB are neutral wrt to global preferential attachment with $\mathcal{M}_e$.

For $\mathcal{M}_g$, it suffices to observe that the above result holds for all $\mathbf{\Theta}$ and $\mathbf{\Phi}$, and not only for $\hat{\mathbf{\Theta}}$ and $\hat{\mathbf{\Phi}}$, so that:

$$P(d_i^{(N)} \geq n+1 |\ d_i^{(p)} = n, \mathcal{M}_g) = \int_{\mathbf{\Theta},\mathbf{\Phi}} P(\mathbf{\Theta}, \mathbf{\Phi} | \mathcal{M}_g) P(d_i^{(N)} \geq n+1 |\ d_i^{(p)} = n, \mathbf{\Theta}, \mathbf{\Phi})\ d\mathbf{\Theta} d\mathbf{\Phi}$$

As the models are neutral with $(\mathbf{\Theta}, \mathbf{\Phi})$, $P(d_i^{(N)} \geq n+1 |\ d_i^{(p)} = n, \mathbf{\Theta}, \mathbf{\Phi}) = P(d_i^{(N)} \geq n+2 |\ d_i^{(p)} = n+1, \mathbf{\Theta}, \mathbf{\Phi})$ and thus:

$$P(d_i^{(N)} \geq n+2 |\ d_i^{(p)} = n+1, \mathcal{M}_g) = P(d_i^{(N)} \geq n+1 |\ d_i^{(p)} = n, \mathcal{M}_g)$$

which completes the proof.

We now turn to local preferential attachment that deals with the fact that preferential attachment can be also observed within classes of nodes, as exemplified in (Leskovec et al. 2008). The classes we consider here are the latent classes of the stochastic mixed-membership models.

4.5.2 Local preferential attachment

Local preferential attachment is a restriction of global preferential attachment at the community level and aims at capturing the fact that the more links a node has in a given community, the more links it will have in the future within this community. The latent classes used in ILFM and IMMSB play the role of latent communities gathering nodes sharing unobserved properties. Local preferential attachment can thus be studied in stochastic mixed membership models by studying how preferential attachment is captured within latent classes. This nevertheless entails that the latent classes be set in one way or another, meaning that the question of whether stochastic mixed membership models comply with the local preferential attachment effect only makes sense in $\mathcal{M}_e$, and not in $\mathcal{M}_g$.

For **ILFM**, the situation wrt to local preferential attachment is very similar to the one for global preferential attachment. This is due to the fact that, in $\mathcal{M}_e$ (i.e. given Θ and Φ), a local degree can be defined in the same way as the global degree above.

Considering the same generative process as before, for $\mathcal{M}_e$ and ILFM, the local degree in class k ($0 \leq k \leq K-1$), for a node i such that $\theta_{ik} = 1$, is defined by:

$$d_{i,k}^{(p)} = \sum_{j=1, \theta_{jk}=1}^{p} y_{ij}$$

Note that if $\theta_{ik} = 0$, $d_{i,k}^{(p)} = 0$ for all p. This then leads to the following definition for local preferential attachment with ILFM.

Definition 4.5.2 (ILFM - local preferential attachment, $\mathcal{M}_e$) *We say that ILFM, in $\mathcal{M}_e$, satisfies the local preferential attachment effect iff for any indexing, for any node i, $1 \leq i \leq N$ such that $\theta_{ik} = 1$, and for any step p, $1 \leq p < N$, $P(d_{i,k}^{(N)} \geq n+1 |\ d_{i,k}^{(p)} = n, \mathcal{M}_e)$ increases with n ($1 \leq n < p$). If $P(d_{i,k}^{(N)} \geq n+1 |\ d_{i,k}^{(p)} = n, \mathcal{M}_e)$ is independent of n, the model is said to be neutral wrt to the local preferential attachment effect.*

As before, we have the following property.

Proposition 4.5.2 *ILFM, with $\mathcal{M}_e$, is neutral w.r.t. the local preferential attachment effect.*

Proof 4.5.2 *The proof is identical to the first part of the proof of Property 4.5.1.*

For **IMMSB** in $\mathcal{M}_e$, we do not have a direct access to classes, encoded in the Z variables. One can nevertheless define local random variables $y_{ij,k}$ that are 1 if a link is generated between nodes i and j within class k and 0 otherwise. One has:

$$\begin{aligned} P(y_{ij,k} = 1 |\ \mathcal{M}_e) &= P(y_{ij} = 1 |\ z_{i\to j} = z_{i\leftarrow j} = k, \Phi) P(z_{i\to j} = k|\Theta) P(z_{i\leftarrow j} = k|\Theta) \\ &= \theta_{ik}\phi_{kk}\theta_{jk} \end{aligned}$$

The local degree $d_{i,k}^{(p)}$ can then be defined as the expectation of $y_{ij,k}$ over the nodes $1, \dots, p$:

$$\begin{aligned} d_{i,k}^{(p)} &= \sum_{j=1}^{p} P(y_{ij,k} = 1 | \mathcal{M}_e) \\ &= \sum_{j=1}^{p} \theta_{ik} \phi_{kk} \theta_{jk} \end{aligned} \tag{4.3}$$

As one can note, such a local degree is not necessarily an integer and the definition of the local preferential attachment has to be adapted accordingly.

Definition 4.5.3 (IMMSB - local preferential attachment, $\mathcal{M}_e$) *We say that IMMSB, in $\mathcal{M}_e$, satisfies the local preferential attachment effect iff for any indexing, for any node i, $1 \leq i \leq N$ such that $\theta_{ik} = 1$, for any step p, $1 \leq p < N$, and for all ϵ compatible with the domain of definition of $d_{i,k}$ and x, $P(d_{i,k}^{(N)} \geq x + \epsilon | \, d_{i,k}^{(p)} \geq x, \mathcal{M}_e)$ increases with x. If $P(d_{i,k}^{(N)} \geq x + \epsilon | \, d_{i,k}^{(p)} \geq x, \mathcal{M}_e)$ is independent of x, the model is said to be neutral wrt to the local preferential attachment effect.*

This definition can be seen as the continuous counterpart of Definition 4.5.2. If ϵ is too large, the probability is null and is independent on x, hence the compatibility requirement with the domain of definition of x and $d_{i,k}$.

Because of the Hierarchical Dirichlet Process underlying the IMMSB model, $\boldsymbol{\theta}_i$ follows a Dirichlet distribution: $\boldsymbol{\theta}_i \sim \text{Dir}((\alpha_0 \beta_k + N_{ik})_{1 \leq k \leq K})$, with $\boldsymbol{\beta} \sim \text{GEM}(\gamma)$ and N_{ik} being the number of edges connecting node i through class k (see for example (Teh et al. 2006)) and K the number of latent classes obtained. The marginals θ_{ik} are thus distributed according to a Beta distribution: $\theta_{ik} \sim \text{Beta}(a_{ik}, b_{ik})$ with $a_{ik} = \alpha_0 \beta_k + N_{ik}$ and $b_{ik} = \sum_{k'=1, k' \neq k}^{K} \alpha_0 \beta'_k + N_{ik'}$.

The following property displays a sufficient condition on x, ϵ, a_{ik} and b_{ik} for IMMSB to satisfy the local preferential attachment.

Proposition 4.5.3 *Let $E_k^p = \sum_{j=1}^{p} \hat{\phi}_{kk} \hat{\theta}_{jk}$ with $1 \leq p \leq N$. In the region where x and ϵ are such that*

$$x^{a_{ik}-1} \left(a_{ik}(E_k^N - \epsilon) + x(b_{ik} - a_{ik}) \right) \geq b_{ik} (E_k^p)^{a_{ik}}$$

$P(d_{i,k}^{(N)} \geq x + \epsilon | \, d_{i,k}^{(p)} \geq x, \mathcal{M}_e)$ increases with x.

Proof 4.5.3 *We consider IMMSB in $\mathcal{M}_e$. Let $F_k^P = \sum_{j=1}^{p} \hat{\theta}_{jk}$ and $F_k^N = \sum_{j=1}^{N} \hat{\theta}_{jk}$.*

Using the change of variables $x' = \frac{x}{F_k^p \phi_{kk}}$ *and* $\epsilon' = \frac{\epsilon}{F_k^N \phi_{kk}}$, *one gets*

$$\begin{aligned} p(d_{ik}^{(N)} > x + \epsilon | d_{ik}^{(p)} > x, \mathcal{M}_e) &= p(\theta_{ik} > qx' + \epsilon' | \theta_{ik} > x') \\ &= \frac{P(\hat{\theta}_{ik} > qx' + \epsilon')}{P(\hat{\theta}_{ik} > x')} = g(x') \end{aligned}$$

where $q = \frac{F_{ik}^p}{F_{ik}^N}$. *The conditional distribution* $g(x')$ *is not trivially equal to one in the case where* $qx' + \epsilon' > x' \Leftrightarrow \epsilon' > x'(1-q)$. *Further, the survival function of* $\hat{\theta}_{ik}$ *is* $P(\hat{\theta}_{ik} > x') = 1 - \int_0^{x'} f(x')$ *where* $f(x')$ *is the density of* $\hat{\theta}_{ik}$. *One can show that the marginal distribution of* $\hat{\theta}_{ik}$ *is a Beta of the form* $f(x') = \mathrm{Beta}(a_{ik}, b_{ik})$ *where* $a_{ik} = \alpha_0\beta_k + N_{ik}$ *and* $b_{ik} = \sum_{k' \neq k} \alpha_0 \beta_{k'} + N_{ik'}$ *(this is a consequence of the form of the posterior distribution of the DP). In the following we ommit the references to* i *and* k *for the parameters of the Beta, simply noting* $\mathrm{Beta}(a, b)$ *for short. The derivative of* g *is*

$$\begin{aligned} g'(x') = (&-q(qx' + \epsilon')^{a-1}(1 - qx' - \epsilon')^{b-1} \int_{x'}^1 t^{a-1}(1-t)^{b-1} dt \\ &+ x'^{a-1}(1-x')^{b-1} \int_{qx'+\epsilon'}^1 t^{a-1}(1-t)^{b-1} dt) \frac{1}{\left(\int_{x'}^1 t^{a-1}(1-t)^{b-1} dt\right)^2} \end{aligned}$$

But one has

$$\int_{qx'+\epsilon'}^1 t^{a-1}(1-t)^{b-1} dt \geq (qx'+\epsilon')^{a-1} \int_{qx'+\epsilon'}^1 (1-t)^{b-1} dt = (qx'+a)^{a-1} \frac{(1-qx'-\epsilon')^b}{b}$$

and

$$\int_{x'}^1 t^{a-1}(1-t)^{b-1} dt \leq (1-x')^{b-1} \int_{x'}^1 t^{a-1} dt = (1-x')^{b-1} \frac{(1-x'^a)}{a}$$

Thus one can show that

$$\begin{aligned} Cg'(x') &\geq x'^{a-1} \frac{1 - qx' - \epsilon'}{b} - q\frac{1 - x'^a}{a} \\ &= \frac{1}{abF_k^N} \left[ax'^{a-1}(1-\epsilon')F_k^N + F_k^P(b(x'^a - 1) - ax'^a)\right] \end{aligned}$$

where $C = \frac{\left(\int_{x'}^1 t^{a-1}(1-t)^{b-1} dt\right)^2}{(qx'+\epsilon')^{a-1}(1-qx'-\epsilon')^{b-1}(1-x')^{b-1}}$, *is a positive constant. Thus, A sufficient* *condition for g to be increasing is that*

$$\begin{aligned}
&F_k^N a x'^{a-1}(1-\epsilon') \geq F_k^p a x'^a + b F_k^p (1 - x'^a) \\
&\Leftrightarrow x'^{a-1} a(1-\epsilon') \geq \frac{F_k^p}{F_k^N}(x'^a(a-b)+b) \\
&\Leftrightarrow x'^{a-1}\left(a(1-\epsilon') - x'(a-b)\frac{F_k^p}{F_k^N}\right) \geq b\frac{F_k^p}{F_k^N}
\end{aligned}$$

Finally, let $E_k^p = \sum_{j=1}^p \phi_{kk}\theta_{jk} = \phi_{kk}F_k^p$ *with* $1 \leq p \leq N$*, by rolling back the change of variables, one obtains*

$$x^{a-1}\left(a(E_k^N - \epsilon) - x(a-b)\right) \geq bE_k^p,$$

which concludes the proof.

As one can note, when x is large, ϵ is small and $b_{ik} - a_{ik} > 0$ (which roughly means that the class k concentrates less than half of the capacaity of the network), then the above condition is likely to be met. In this situation, IMMSB satisfies the local preferential attachment effect.

We now present an experimental illustration of the above theoretical results.

4.6 Illustration

To illustrate our theoretical results, we evaluate the predictive performance and the ability of the models to capture the preferential attachment on artificial and real networks. In order to evaluate this property we used several measures.

The measures considered to evaluate the preferential attachment rely on a goodness of fit. Indeed, it has been reported that preferential attachment leads to networks characterized by a degree distribution with heavy tail drawn from a power law (Barabási & Albert 1999). A graphical method, most often used to verify that the observations are consistent with this law consists in constructing the histogram representing the degree distribution and if the plot on doubly logarithmic axes approximately falls on a straight line, then one can assume that the distribution follows a power law. Thus, the comparison of the degree distribution in the log-log scale with a linear function gives us a qualitative measure for the preferential attachment. To obtain a second evaluation of the power law hypothesis for the degree distribution, we follow the statistical framework, introduced by (Clauset et al. 2009), for discerning and quantifying power-law behavior in empirical data. This framework combines maximum-likelihood fitting methods with goodness-of-fit tests based on the Kolmogorov-Smirnov statistic. It includes the following steps:

- Estimate the parameters α and $x_{\min}$ of the power law model. α is the scaling parameter of the law and $x_{\min}$, the lower bound for the tail. It has been fixed to the smallest value observed in the distributions evaluated, in our experiments to allow their comparisons.
- Using the Kolmogorov-Smirnov (KS) statistic, compute the distance KS_{obs} between the degree distribution obtained on the network with the theoretical distribution corresponding to the power law with the estimated parameters.
- Sample S synthetic datasets from the power law with the estimated parameters. For each sample dataset $s \in S$, compute the distance KS_s between the distribution obtained on this synthetic dataset, drawn from the power law, with the corresponding theoretical distribution using the Kolmogorov-Smirnov statistic.
- Decide how many sample dataset S to use, with a rule of thumb, based on a worst-case performance analysis of the test (Clauset et al. 2009). To obtain a precision of the p-value about ϵ, one should choose $S = \frac{1}{4}\epsilon^{-2}$.
- The p-value is defined as the fraction of the resulting statistics $KS_s, s \in \{1, \dots, S\}$ obtained on the samples larger than the value KS_{obs} computed on the network distribution.

If p-value is large (close to 1), then the difference between the data and the model can be attributed to statistical fluctuations alone; if it is small, the model is not a plausible fit for the data and we can not conclude that there is an evidence for the preferential attachment in the network. However, as mentioned in (Clauset et al. 2009) high value of the p-value should be considered with caution for at least two reasons. First, there may be other distribution that match the data equally or better. Second, a small number of samples of the data may lead to high p-value and reflect the fact that is hard to rule out a hypothesis in such a case.

For local preferential attachment, we follow the same approach as before to compute the p-value, the only difference being that the empirical data does not correspond any longer to the global adjacency matrix, but to reduced matrices for each class. The computation of the reduced adjacency matrices varies from one model to the other:

- For IMMSB, for a given class k, the reduced adjacency matrix Y^k is defined by: $y_{ij,k} = 1$ if $y_{ij} = 1, z_{i\to j} = z_{i\leftarrow j} = k$ and 0 otherwise.
- For ILFM, the reduced adjacency matrix Y^k is defined by: $y_{ij,k} = 1$ if $y_{ij} = 1, \theta_{ik} = \theta_{jk} = 1$ and 0 otherwise.

Note that all our experiments where realized in a platform that we developed and maintain in order to help reproducibility of machine learning experiments. It is available online[4] under a GNU GPL license.

4.6.1 Datasets

To illustrate the above developments, we consider two artificial and two real networks, the characteristics of which are summarized in Table 4.1.

Table 4.1: Characteristics of artificial and real networks.

Networks	nodes	edges	density
Network1	1000	3507	0.007
Network2	1000	31000	0.062
Blogs	1490	20512	0.009
Manufacturing	167	5950	0.215

The non-oriented artificial networks (Network1 and Network2) have been generated with the DANCer-Generator (Largeron et al. 2015). This generator has been chosen because it allows one to build an attributed graph having a community structure as well as known properties of real-world networks such as preferential attachment and homophily. In order to test link prediction models on different types of networks, Network1 was generated, by design, to comply with preferential attachment whereas Network2 was not.

The first real network, denoted Blogs[5], contains front-page hyperlinks between blogs in the context of the 2004 US election. A node represents a blog and an oriented link represents a hyperlink between two blogs. The second one, denoted Manufacturing[6], is an internal email communication network between employees of a mid-sized manufacturing company. Each node is associated to an employee and an oriented link represents an email sent between the two employees. One can notice that the second network is specific since it is an enterprise network in which the relationships between the employees are (professionally) constrained.

[4] https://github.com/dtrckd/pymake
[5] moreno.ss.uci.edu/data.html#blogs
[6] www.ii.pwr.edu.pl/~michalski/index.php?content=datasets#manufacturing

This means that this network is less likely to display some of the properties that occur in unconstrained social networks.

The adjacency matrices and global degree distributions of these networks are presented in Figure 4.1. This figure allows us to visualize some characteristics of the networks such as their density and their clustering patterns: as one can note, Blogs and the two artificial networks (Network1 and Network2) have a clear community structure, corresponding to the blocks of white dots on the figure, whereas Manufacturing, the denser network, does not have such a structure. Furthermore, the log-log scale plots show that Network1 and Blogs verify the global preferential attachment (the fitted line represents relatively well the data points) whereas neither Network2 nor Manufacturing verify it. This is confirmed by the p-values reported in the first section of Table 4.2 (Training Datasets): the p-value is 1 for Network1 and Blogs, whereas it is null for Network2 and Manufacturing. The parameter α reported in Table 4.2 corresponds to the parameter of the estimated power law distribution (*i.e.* the slope of the best fitting line in log-log scale).

Figure 4.2 represents the local degree distributions for all networks, each curve in each plot being associated to a different class. As the ground truth is not available for the real networks (Blogs and Manufacturing), classes have been determined with Louvain algorithm (Blondel et al. 2008) and the local distribution defined according to the obtained classes. As one can note, the plots for Network1 and Blogs are linear for the most frequent degrees, whereas the plots for Network2 and Manufacturing do not display any clear linearity, suggesting that Network1 and Blogs satisfy, at least partly, local preferential attachment whereas Network2 and Manufacturing do not. This is confirmed by the p-values reported in Table4.2: the p-value equals 1 for Network1 and Blogs, 0 for Network2 and 0.4 for Manufacturing.

4.6.2 Homophily

Figure 4.3 presents boxplots describing the distributions of the natural $s_n(i,j)$ and latent $s_l(i,j)$ similarities computed respectively on linked and non-linked pairs of nodes for IMMSB (top) and ILFM (bottom). The results have been aggregated over the four datasets. They confirm that the natural similarity is higher for pairs of nodes which are linked than for pairs of nodes which are not linked, for both models. For the latent similarity, there is no difference between the linked and non-linked pairs, indicating that the links are not homophilic. These experimental results are in line with the theoretical results presented in Section4.4 that state

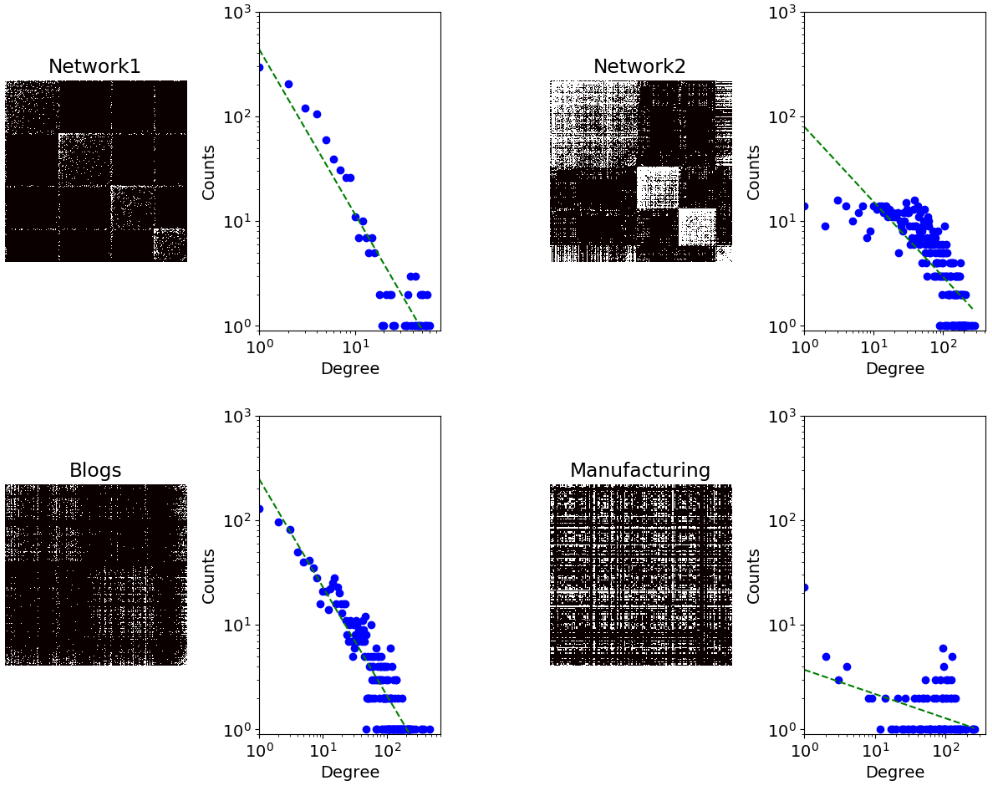

Figure 4.1: Adjacency matrices (left) and global degree distributions (right) for the four training datasets. In the adjacency matrices, a white dot corresponds to a 1 and a black dot to a 0.

that both ILFM and IMMSB are homophilic for the natural similarity but are not homophilic for the latent similarity.

4.6.3 Preferential attachment in $\mathcal{M}_e$

For each dataset, we estimated the model parameters through a Markov Chain Monte Carlo inference consisting of 200 iterations. For IMMSB, the concentration parameters of HDP were optimized using vague gamma priors $\alpha_0 \sim \text{Gamma}(1, 1)$ and $\gamma \sim \text{Gamma}(1, 1)$ following (Teh et al. 2006). The parameters for the matrix weights λ_0 and λ_1 were fixed to 0.1. For ILFM, the hyper-parameter σ_w was fixed to 1 and the IBP hyper-parameter α to 0.5. Once the models have been learned, they are used to generate links (or non-links) between the entire set of network nodes. The whole procedure is repeated 10 times and the average values are reported as final results.

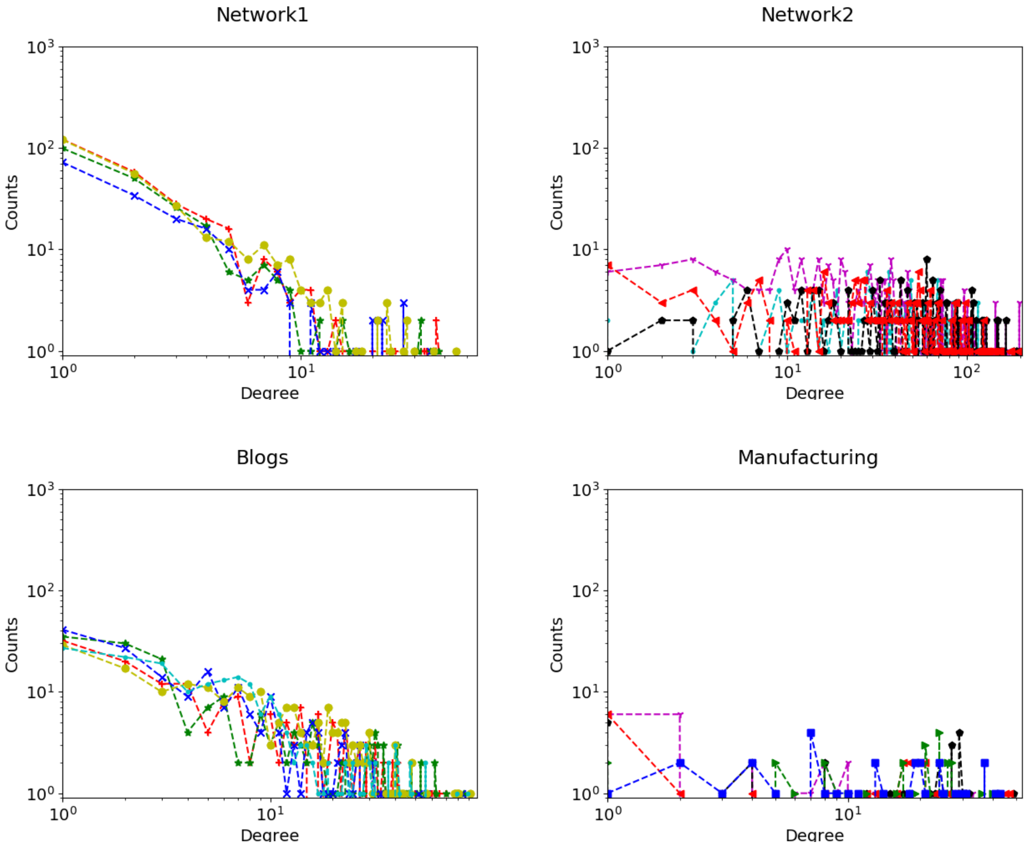

Figure 4.2: Local degree distributions for the four training datasets. For Network1 and Network2 the classes come from ground-truth. For Blogs and Manufacturing, classes are obtained with a Louvain algorithm.

4.6.3.1 Degree distributions

Table 4.2 reports the value of the power-law goodness of fit for IMMSB and ILFM in the global case (left) and in the local case (right). The precision of the test was set to $\epsilon = 0.03$. It appears that for both models, the global preferential attachment is only verified for networks generated from datasets where the property was observed, namely in Network1 with p-value equal to 0.9 for IMMSB and 1 for ILFM, and in Blogs with a p-value equal to 1 for both models; the property is not verified in Network2 and in Manufacturing, where the p-values are equal to 0. This is in accordance with Proposition 2.1 according to which both ILFM and IMMSB do not satisfy global preferential attachment. However, these models are able to capture this property if it exists in the training datasets. Moreover, one can observe that, in the local case, IMMSB complies with the preferential attachment with p-values equal or close to 1 for the four networks, while ILFM obtained low p-values for the networks that were less locally bursty (respectively 0 for Network2 and 0.3 for Manufacturing). In addition, the power-law coefficients α are significantly greater

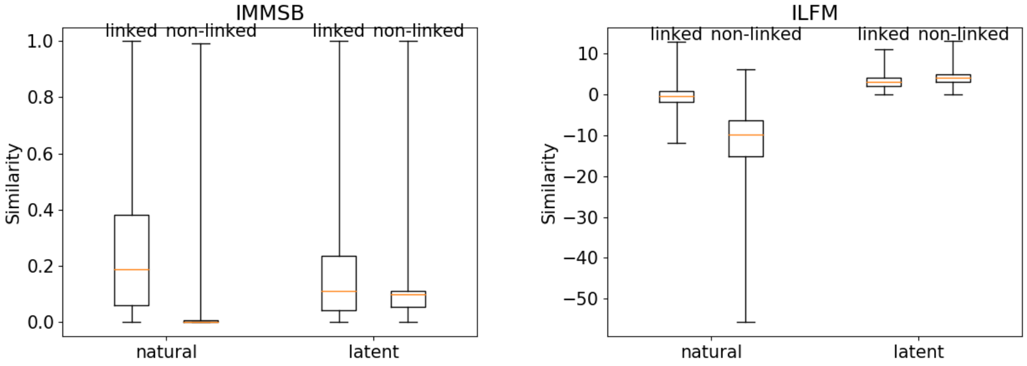

Figure 4.3: Natural and latent similarities aggregated over all datasets and computed on linked and non-linked pairs of nodes for IMMSB (left) and ILFM (right).

for IMMSB than for ILFM, and specially for the bursty networks Network1 and Blogs.

Figure 4.4 illustrates the local preferential attachment for Network1 (top) and Network2 (bottom) estimated with IMMSB (left) and ILFM (right). The shape of the local degree distributions appears more linear for IMMSB and with more fluctuations for ILFM. This illustrates the fact that ILFM does not capture local preferential attachment whereas IMMSB does, as stated in Proposition 4.5.3.

4.6.3.2 Generating process

In Figure 4.5 and 4.6 we show respectively for IMMSB and ILFM the evolution of the local preferential attachment following the definition given in section 4.5.2, for the networks Manufacturing and Networks1 and for two different values of the generating process at step p (p is given as a percentage of N in the plots). For IMMSB one can see on figure 4.5, that the probability of generating new links increases with the degree. However, for ILFM, one can observe, on figure 4.6, some classes where the preferential attachment is no true such as for the class 3 in Manufacturing where the probability to generate new links decreases with the degree or contains some plateau. For Networks1, the probability to generate new links increases in average because the model is fitted with a networks where the preferential attachment is present. However, on can see that the increase of the probability is not as clear than for IMMSB. The value of the probability fluctuate and reaches some plateau. The interpretation of these results with regard to the properties that we asses for the local preferential attachment is that IMMSB is better adapted than ILFM to capture the local preferential attachment.

Table 4.2: Preferential attachment measures for training datasets (first row group) and networks generated with fitted models (second row group for IMMSB and third row group for ILFM. For the local preferential attachment, the mean and standard deviation values of the latent classes are reported.

Training Datasets	Global		Local	
	p-value	α	p-value	α
Network1	1	2.4	1.0 ± 0.0	1.8 ± 0.03
Network2	0	1.3	0.0 ± 0.0	1.2 ± 0.01
Blogs	1	1.5	1.0 ± 0.0	1.4 ± 0.03
Manufacturing	0	1.4	0.4 ± 0.3	1.3 ± 0.05
IMMSB				
Network1	0.9	1.4	**1.0** ± 0.0	**3.5** ± 0.7
Network2	0	1.3	**0.9** ± 0.0	1.6 ± 0.2
Blogs	1	1.3	**1.0** ± 0.0	**4.3** ± 1.1
Manufacturing	0	1.2	**0.9** ± 0.01	1.6 ± 0.1
ILFM				
Network1	1	1.4	**1.0** ± 0.0	1.7 ± 0.1
Network2	0	1.2	0.0 ± 0.0	1.2 ± 0.0
Blogs	1	1.3	**0.9** ± 0.2	1.5 ± 0.1
Manufacturing	0	1.2	0.3 ± 0.3	1.3 ± 0.0

4.6.3.3 Performance evalutation

In Figure 4.7, we compare the performance of the models for predicting new links using the Area Under the Curve (AUC) measure as a function of the training set size. In the bottom plot, the y-axis gives the relative performance defined as the difference of the AUC values for IMMSB and ILFM ($AUC_{IMMSB} - AUC_{ILFM}$) whereas the x-axis indicates the percentage of links randomly removed from the datasets and used as test examples. Hence, the number of training data decreases with the x-axis and a positive value on the y-axis indicates that IMMSB outperforms ILFM. The relative performance corresponds to the difference of the best AUC values obtained for both models on the 10 inference experiences. The top plots illustrate a case where 75 percent of the data is used as test set and where IMMSB dominates ILFM on Network1 (left), and the opposite on Network2 (right).

In general, as shown in the bottom plot, ILFM obtains better performance than IMMSB. However, the relative predictive performance of IMMSB increases when the quantity of training data decreases on bursty networks, whereas for non-bursty networks the results are the opposite: the performance of ILFM increases when the size of the learning dataset decreases. This is particularly visible for Network2. The results for Manufacturing are less marked, which is certainly due to the small

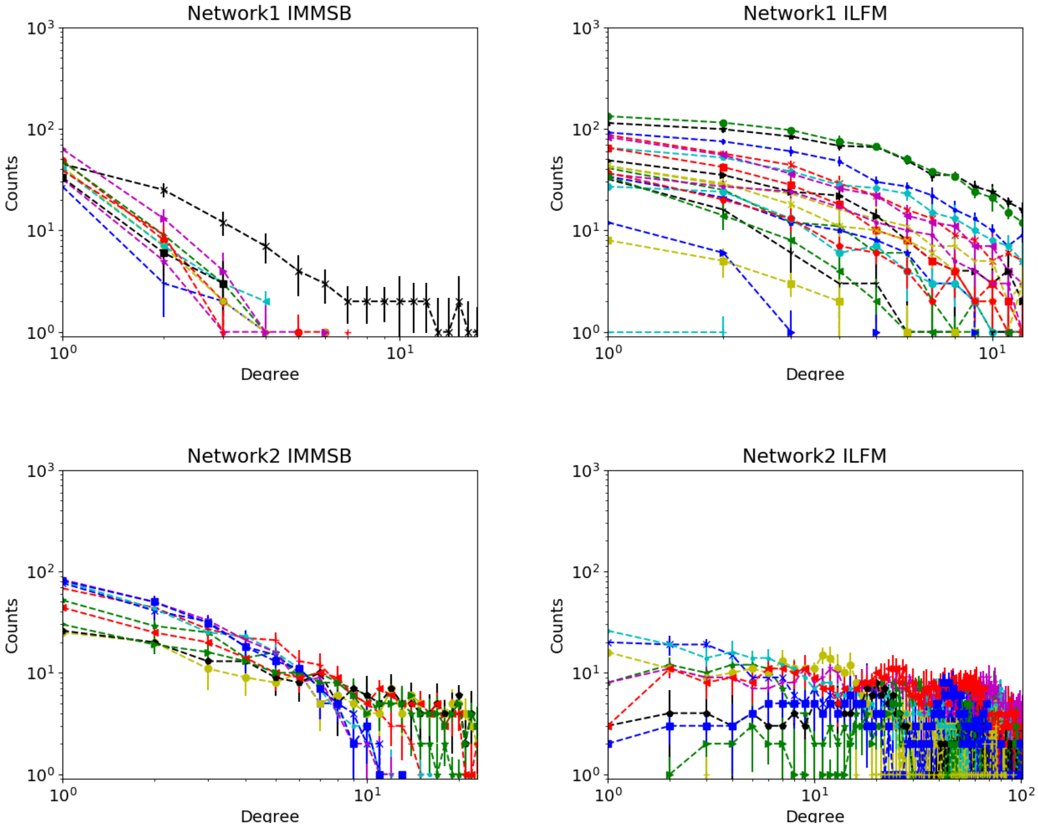

Figure 4.4: Local degree distributions for Network1 (top row) and Network2 (bottom row) generated with fitted models IMMSB (first column) and ILFM (second column).

size of this network, making the prediction less challenging.

The above behavior can be explained by the fact that IMMSB satisfies the local preferential attachment whereas ILFM does not: as links are randomly removed, one is more likely to remove links from large classes than from small ones; a model that enforces local preferential attachment on bursty networks is thus more likely to reconstruct those removed links. This is what is happening on Network1 and Blogs for IMMSB. On the contrary, for non-bursty networks, a model enforcing local preferential attachment is penalized.

4.6.4 Preferential attachment in $\mathcal{M}_g$

Illustrations in the $\mathcal{M}_g$ case are based on the simulation of the models where the parameters Θ and Φ have been marginalized out. In other words the degrees that we are going to observe are the expected degree for a large numbers (in the sense of the theory of large numbers) of generated parameters, given the hyper-parameters of the model. In order to simulate this scenario, we generated a large number (100)

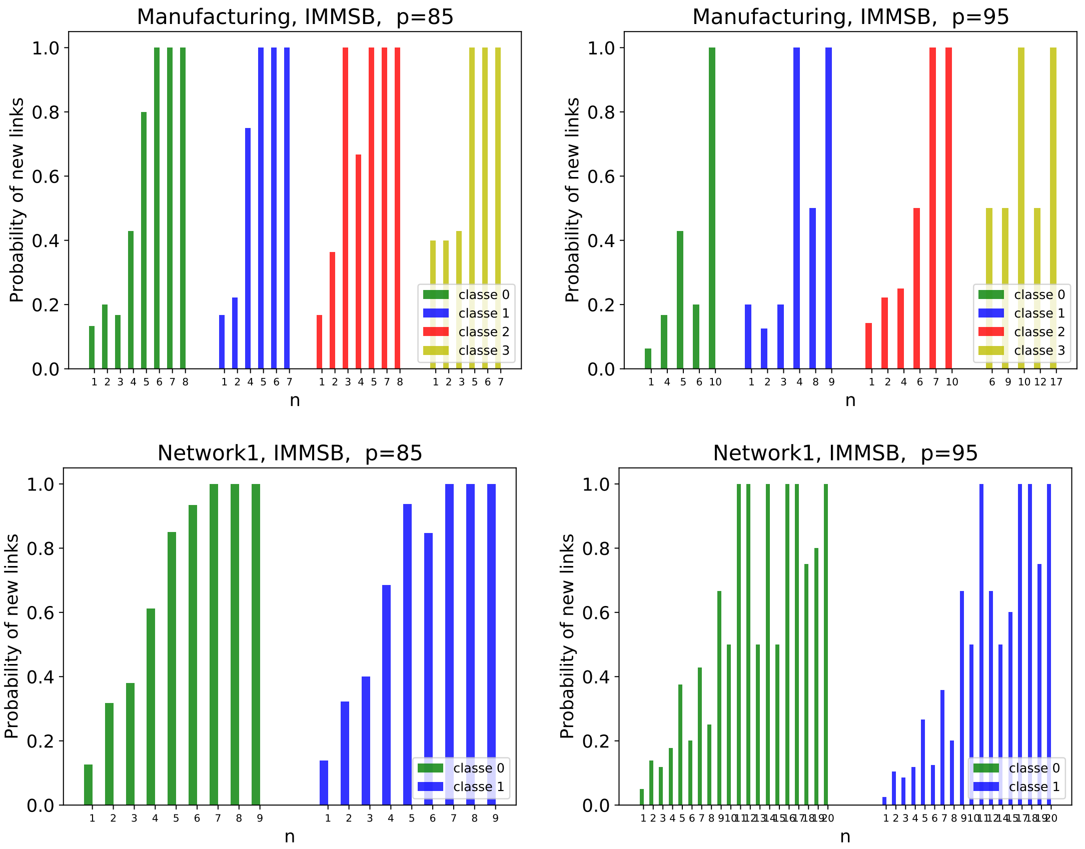

Figure 4.5: Local burstiness process for IMMSB illustrated by the probability to generate new links for degree at step p. The model is fitted with the Manufacturing and Network1 networks for respectively line 1 and 2. First row is for a value of the generating step $p = 85\%$ (percentage of total number of nodes N) and $p = 95\%$ for the second row.

of networks with a given set of hyper-parameters for the models. Then we reported the average global degree distribution in Figure 4.8 (top). For this experiments, fix the number of nodes to 1000. We went trough the generative process 100 times in order to simulate the $\mathcal{M}_g$ mode. In order to have a comparable number of classes for ILFM and IMMSB and block-block probability priors, we fix the hyper-parameters $\lambda_0 = \lambda_1 = 0.5$ for both models, $\alpha = \alpha_0 = 1$ for ILFM, and $\gamma = 0.5$ for IMMSB.

We see that the global degree distributions are not monotone, with several peaks, and that the range values of the outcome degrees are concentrated in a small segment determined by the hyper-parameters of the models. The shape of the global degree distributions shows that the global preferential attachment is not satisfied.

In the Figure 4.8 (bottom), we also reported a measure on the local preferential attachment in $\mathcal{M}_g$. An important note is that, to be able to compute the statistics

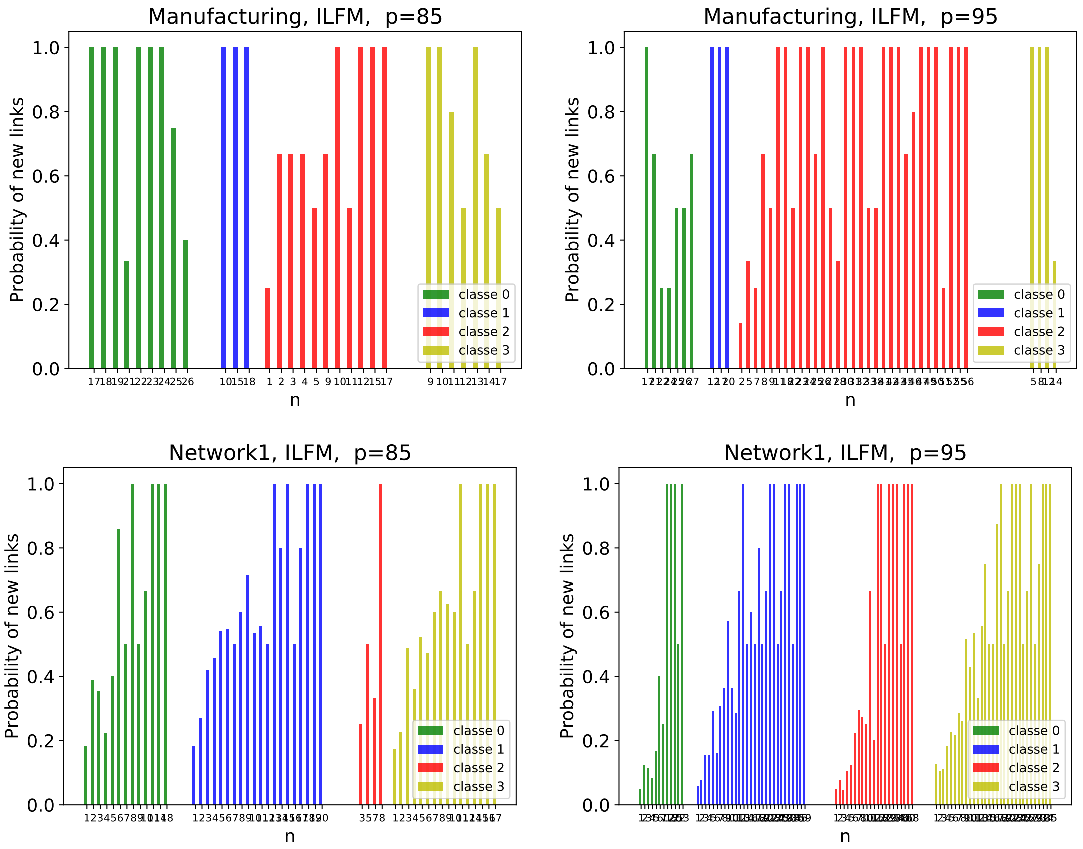

Figure 4.6: Local burstiness process for ILFM illustrated by the probability to generate new links for degree at step p. The model is fitted with the Manufacturing and Network1 networks for respectively line 1 and 2. First row is for a value of the generating step $p = 85\%$ (percentage of the total number of nodes N) and $p = 95\%$ for the second row.

for the local degree, the latent classes need to be aligned between the different epochs in order to report average values of the local degree distributions. But, the mixed membership models do not defined unique labels over the latent classes. Thus, it is not straightforward to identify the common classes between the generations of the different network realizations. Actually, as the processes are exchangeable, they is no strict correspondence between classes in two independent generative process. Nevertheless, the property of the Dirichlet Process and the Indian Buffet Process, enable to identify the classes by ordering them with their size (or concentration). For example, the stick breaking process interpretation of the DP provides a natural class ordering with a descending (or ascending) order of the class representations. While the IBP generates a row-exchangeable feature matrix, it is possible to reorder the rows to obtain a Θ matrix where the size of the classes keeps the same descending (or ascending) order.

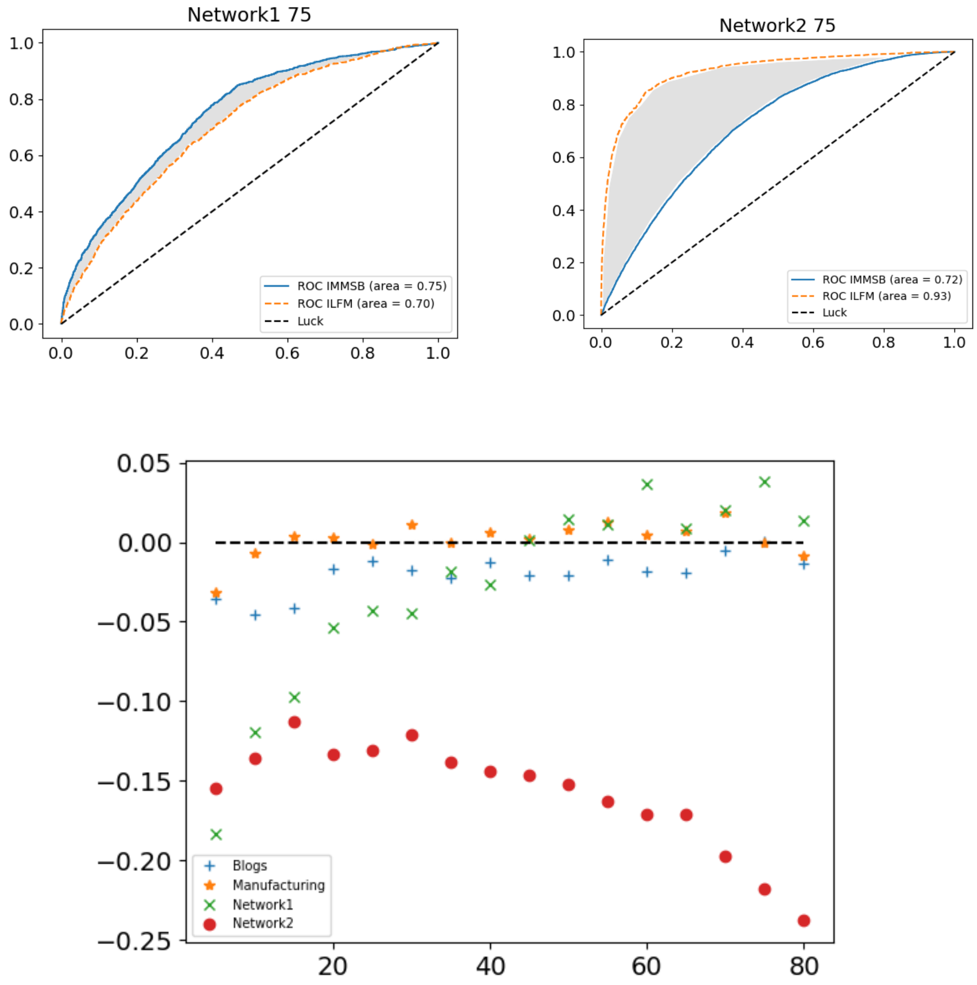

Figure 4.7: Top: AUC-ROC curves for Network1 (left) and Network2 (right) with 75 percent of data used for learning that compares the performance of models. Bottom: Relative performance of IMMSB and ILFM according to the percentage of data used for testing, the rest being used for learning.

For the local degree, one can see that, for IMMSB, the shape of the distributions is characteristic of the preferential attachment effect (linear decrease in a log-log space) while it is not the case for ILFM. This experiment is interesting as it show that, for IMMSB, the local preferential attachment property in $\mathcal{M}_e$ seems to holds also in $\mathcal{M}_g$.

4.7 Conclusion

We have studied whether stochastic mixed membership models, such as ILFM and IMMSB, can generate new links while satisfying important properties frequently observed in real social networks, namely homophily and preferential attachment. To

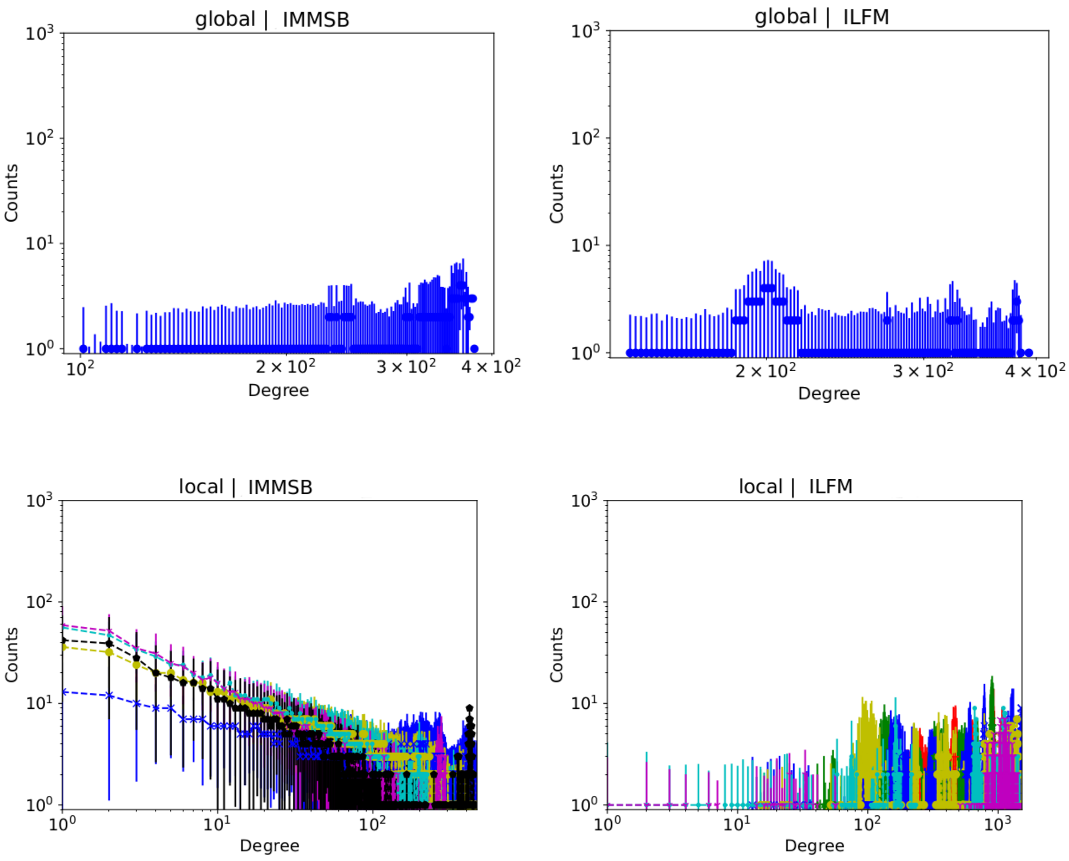

Figure 4.8: Global degree distribution (top) and local degree distribution (bottom) from IMMSB (left) and ILFM (right) in the generative mode $\mathcal{M}_g$.

do so, we have introduced formal definitions adapted for these properties in a global and a local context where edges are either considered across the full network or inside communities. We have analyzed how these models behave according to those definitions. We have shown, in particular, that both models are homophilic with the natural similarity that underlies them. Concerning the preferential attachment, we have shown that stochastic mixed membership models do not comply with global preferential attachment. The situation is however more contrasted when the property is considered at the local level: IMMSB enforces a partial local preferential attachment whereas ILFM does not.

These findings have been validated experimentally on two real and two artificial networks that have different degrees of global and local preferential attachment. An important, practical finding of our study is that IMMSB, usually considered of lesser "quality" than ILFM, can indeed yield better results on bursty networks (*i.e.* networks with preferential attachment) when the number of training data is limited.

There are many directions to extend this work with the motivation of improving our theoretical understanding of graphical models for link prediction in complex networks. A interesting extension is to examine the relation between the local preferential attachment and the dynamic of the latent classes.

An other direction of interest in the line of this work is to study how the preferential attachment relates to the sparsity of a network and how the exchangeability assumption should be relaxed in order to have models that naturally comply with both the preferential attachment and the sparsity properties.

Chapter 5

Weighted Mixed Membership Stochastic Block Model and Scalable Inference

5.1 Introduction

Most of the real networks exhibit a topology more complex than just binary relationship between nodes. Instead, the relations can be weighted and dynamic. For example, co-authorship networks can be constructed such that the edge covariates correspond to the number of collaborations between the underlying authors (M. Newman 2001). In a communication network, the weight can be the number of messages sent from the sender to the receiver. In the web, documents are connected with hyperlinks where the counts of those are for example used to construct the PageRank algorithm. Finally, in a linguistic network, a network of words can be built where the weight between two words is the number of times where they follow each other. Another useful case where weighted networks can be useful is temporal networks. For instance, in communication networks, messages are sent at a specific time, thus taking into account the number of messages sent during a period of time allows to represent the strength of the relation over the time.

In order to capture this information and to alleviate the bias that often consist of thresholding the weighted networks to binary ones, weighted versions of the Stochastic Block Model have been proposed (Karrer & Newman 2011; Mariadassou et al. 2010; Aicher et al. 2014; Peixoto 2018). Those models however suffer from the same drawback as standard stochastic block models, namely the fact that a node can belong to only one class, which is not realistic for many networks. Mixed-membership block models were specifically designed to overcome this limitation and we propose here a new mixed-membership block model adapted to weighted networks. One important aspect in designing a generative model for networks is to develop a scalable inference method so that the model can be applied on large networks. We rely in this study on collapsed variational inference coupled with stochastic variational inference to do so.

The remainder of the chapter is organized as follows: Section 5.2 describes related work; Section 5.4 presents the weighted mixed-membership models and Section 5.5 their inference; Section 5.6 illustrates the behavior of the proposed models on several real-world networks. Finally, Section 5.7 concludes the study.

5.2 Related work

The original MMSB model was proposed in (Airoldi et al. 2009) with a variational inference scheme. The inference process was later extended with stochastic variational inference in (Gopalan & Blei 2013) and structured variational inference in (Kim et al. 2013) for scalability purposes. Stochastic variational inference has been applied with a collapsed variational objective for the latent Dirichlet allocation model (Foulds et al. 2013). To our knowledge, it is the first time that stochastic and collapsed variational inference are coupled in the context of stochastic block models.

In the other hands, the original Stochastic Block model has been extended to the case of weighted network with Poisson law, with maximum likelihood algorithm in (Karrer & Newman 2011), a variational EM approach in (Mariadassou et al. 2010) and with a Variationnal Bayes method in (Aicher et al. 2014). More recently, an other weighted version of the stochastic block model has been proposed in (Peixoto 2018) compatible with different possible kernels depending on the types of weights and with an efficient MCMC inference method. The weighted Stochastic Block Model can be seen as a special case of the WMMSB model proposed in this chapter in which nodes are constrained to belong to only one latent class. If this type of models is interesting, it nevertheless relies again on the assumption that a node belongs to only one class, which may be inappropriate for real-world networks. Furthermore, unlike MMSB models, the lack of a hierarchical prior structure does not allow one to rely on efficient non-parametric extensions (hence the use of costly model selection techniques for non-parametric versions).

Similar to our model, count processes with Poisson distributions and Gamma conjugate priors have been studied by different authors (Zhou & Carin 2012; Zhou & Carin 2015). The relation of such processes with Negative Binomial processes is well-known and is highlighted by these authors. Such processes can be used for topic modeling, as the Beta-Gamma-Gamma-Poisson model of (Zhou et al. 2012) that relies on MCMC inference. The main difference between this model and WMMSB is that the former factorizes counts as Poisson variables of a sum of latent factors as in (Zhou 2015), while for WMMSB, counts are factorized as a convex sum of Poisson variable depending on class memberships.

The main theoretical contribution of this chapter is two-fold: firstly, we propose a mixed-membership stochastic block model, called WMMSB-bg, for weighted networks allowing nodes to belong to several classes, and secondly we show how to efficiently learn this model on large networks with a stochastic collapsed variational inference algorithm.

5.3 Weighted networks and the Poisson law

In this study, we consider the weighted relations as a measure for the number of times each node has interacted. Thus, a natural prior for count edge covariate is the Poisson distribution. Furthermore, it has several nice properties:

- Additivity: If $K_1 \sim \text{Poi}(\alpha_1)$ and $K_2 \sim \text{Poi}(\alpha_2)$ then,

$$K_1 + K_2 = \text{Poi}(\alpha_1 + \alpha_2)$$

- Thinning: The number of successes in a Poisson number of coin flips is Poisson, namely if $K \sim \text{Poi}(\alpha)$ and $X_1, \dots, X_K \sim \text{Bern}(p)$ then,

$$\sum_{i=1}^{K} X_i = \text{Poi}(p\alpha)$$

These two properties justify to build weighted networks datasets from sequence of either weighted graphs or binary graphs to feed a Poisson based model. This is convenient to exploit the network datasets that are often provided as a time sequence of binary or weighted interactions by summing up all node pairs interactions.

As usual, we will consider that a network is represented by a graph $G = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of nodes such that $N = |\mathcal{V}|$ and $\mathcal{E}$ the set of edges. We consider the adjacency matrix $Y = (y_{ij})_{ij \in N^2}$ such that $y_{ij} = 0$ if $(i, j) \notin \mathcal{E}$ and $y_{ij} > 0$ otherwise.

5.4 Mixed-Membership Stochastic Block Models and (un)weighted graphs

Mixed-membership stochastic block (MMSB) models extend stochastic block models (Airoldi et al. 2009) by allowing nodes to "belong" to several blocks (or classes) through a given (usually Dirichlet) probability distribution. Prior to generate a link between two nodes, a particular class is selected for each node. The link is then generated according to a probability distribution F, sometimes referred to as the *kernel* distribution, that depends on the selected classes. The generative process behind such models can be summarized as:

1. For each node i, draw
$$\theta_i \sim \text{Dir}(\alpha),$$
where θ_i and α are K-dimensional vectors, K denoting the number of classes considered;
2. Generate two sets of latent class memberships for each possible interactions,
$$Z_{\rightarrow} = \{z_{i \rightarrow j} \sim \text{Cat}(\theta_i), 1 \leq i, j \leq N\}$$
and
$$Z_{\leftarrow} = \{z_{i \leftarrow j} \sim \text{Cat}(\theta_j), 1 \leq i, j \leq N\},$$
with categorical draws;
3. Generate or not a link between two nodes (i, j) according to
$$y_{ij} \sim F(\phi_{z_{i \rightarrow j} z_{i \leftarrow j}}),$$
where F is a distribution in the exponential family and $\phi_{z_{i \rightarrow j} z_{i \leftarrow j}}$ an associated (usually drawn from a conjugate distribution) parameter that represents the relations between classes. For unweighted graphs, F is usually Bernoulli and ϕ its conjugate Beta distribution.

As mentionned, many real networks nevertheless rely on graphs in which edges are naturally weighted. the number of collaborations in a co-authorship networks, the number of messages sent from the sender to the receiver in a communcation network, etc. In all these cases, weights are integers that can naturally be modeled with Poisson distributions. Relying on its conjugate Gamma distribution for ϕ, one finally obtains the following models, denoted MMSB for unweighted graphs and WMMSB for weighted graphs:

$$\theta_i \sim \mathrm{Dir}(\alpha),\ z_{i\to j} \sim \mathrm{Cat}(\theta_i),\ z_{i\leftarrow j} \sim \mathrm{Cat}(\theta_j)$$

and:

$$\begin{aligned} y_{ij} &\sim \mathrm{Bern}(\phi_{z_{i\to j}z_{i\leftarrow j}}), \quad & \phi_{kk'} &\sim \mathrm{Beta}(\lambda_0, \lambda_1), & \textit{for unweighted graphs} \\ y_{ij} &\sim \mathrm{Poi}(\phi_{z_{i\to j}z_{i\leftarrow j}}), \quad & \phi_{kk'} &\sim \mathrm{Gamma}(r, \frac{p}{1-p}), & \textit{for weighted graphs} \end{aligned}$$

The choice made here for the Poisson and Gamma distributions in WMMSB allows one to represent overdispersed count data as one has (Zhou et al. 2012)

$$y_{ij} \sim \mathrm{NB}(r, p)$$

where NB denotes the negative binomial distribution. Furthermore, the above models are valid for both directed and undirected graphs, the matrix $\Phi = (\phi_{kk'})_{k,k' \in \{1,\dots,K\}^2}$ being symmetric in the latter case.

5.4.1 Beta-Gamma augmentation

The generative process for WMMSB defined above assumes that the parameters of the Poisson distributions used to generate links are drawn from the same Gamma distribution. Having a unique prior over these parameters however limits the ability of the model to capture the variance in the relations between the latent classes. Hierarchical extensions can be used here to have a better representation of the classes and the relations between them. Following the Beta-Gamma-Gamma-Poisson model (Zhou et al. 2012) and the Gamma-Negative Binomial process (Zhou & Carin 2015), we model here the rate parameter of the Gamma distribution used in WMMSB with a Beta prior and its shape parameter with another Gamma distribution of the form:

$$r_{kk'} \sim \text{Gamma}(c_0 r_0, 1/c_0) \qquad p_{kk'} \sim \text{Beta}(c\epsilon, c(1-\epsilon))$$
$$\phi_{kk'} \sim \text{Gamma}(r_{kk'}, \frac{p_{kk'}}{1-p_{kk'}})$$

The variable y_{ij} is again distributed according to a negative binomial distribution, of the form:

$$y_{ij}|Z \sim \text{NB}(r_{z_{i\to j}z_{i\leftarrow j}}, p_{z_{i\to j}z_{i\leftarrow j}}).$$

As one can note, and contrary to WMMSB, the parameters of the negative binomial distribution depend this time on the classes selected for each node, meaning that classes now play a prominent role in the model. We will denote this model as WMMSB-bg.

As for most hierarchical Bayesian model, exact inference is intractable and one must resort to approximate inference. In the next section we propose a stochastic collapsed variational inference algorithm for the above models (MMSB, WMMSB, WMMSB-bg).

5.5 Inference

Standard inference method for MMSB models rely either on Gibbs sampling or variational approach (Airoldi et al. 2009). The former approach give generally better results than the latter as sampling methods approximate the true posterior distribution while variational ones makes stronger assumptions on the posterior distribution that leads to a biased estimation. On the other hand variational approach usually allow faster convergence due to its deterministic form.

Collapsed variational Bayes inference presents the advantage, over standard variational inference, to rely on weaker assumptions and has proven to be efficient on the latent Dirichlet allocation model (Teh et al. 2007). Recent advances in stochastic variational inference (Hoffman et al. 2013), notably based on well-designed sampling techniques (Gopalan & Blei 2013; Kim et al. 2013), have furthermore shown that it is possible to speed-up (collapsed) variational inference with online updates based on minibatches. Coupling collapsed and stochastic variational inference thus leads here to an efficient inference method that can be used on large networks.

We first provide below the results obtained through collapsed variational inference for MMSB and its weighted counterparts. A detailed derivation of these results is given in the appendix 7. We then detail how stochastic variational inference is used on these models.

5.5.1 Collapsed Variational Inference

In the remainder, we use the notation n^{-ij} to indicate that the superscript ij is excluded from the underlying count variable, and $n_{.}$ to indicate a sum over the dotted subscript index. Furthermore, Π will denote the model parameters ($\Pi = (\Theta, \Phi, Z)$ for MMSB and WMMSB and $\Pi = (\Theta, \Phi, Z, R, P)$ for WMMSB-bg) and Ω the hyper-parameters ($\Omega = (\alpha, \lambda_0, \lambda_1)$ for MMSB, $\Omega = (\alpha, r, p)$ for WMMSB and $\Omega = (\alpha, c_0, r_0, c, \epsilon)$ for WMMSB-bg).

From Jensen's inequality, for any distribution q, one has:

$$\log p(Y|\Omega) \geq \mathbb{E}_q[\log p(Y, \Pi \ |\Omega)] + \mathrm{H}[q(\Pi)]$$

where H denotes the entropy. The goal of variational inference is then to find q that maximizes the right-hand side of the above inequality, usually referred to as the Evidence Lower BOund (ELBO). In its collapsed version, following (Teh et al. 2007), one weakens the mean-field assumption made over the variational distribution, leading to, for MMSB and WMMSB:

$$q(\Pi) = q(\Theta, \Phi|Z)q(Z)$$

with $q(z_{i\rightarrow j}, z_{i\leftarrow j}|\gamma_{ij})$ being multinomial with parameter γ_{ij}. The evidence is then lower bounded by:

$$\log p(Y|\Omega) \geq \underbrace{\mathbb{E}_q[\log p(Y, Z)] + \mathrm{H}[q(Z)]}_{\mathcal{L}_Z}$$

Maximizing $\mathcal{L}_Z$ w.r.t $\gamma_{ijkk'}$ under a zero order Taylor expansion and a Gaussian approximation, following (Teh et al. 2007; Asuncion et al. 2009), yields the following updates:

$$\gamma_{ijkk'} \propto (N^{\Theta^{-j}}_{\rightarrow ik} + \alpha_k)(N^{\Theta^{-i}}_{\leftarrow jk} + \alpha_{k'})p(y_{ij}|\ Y^{-ij}, Z^{-ij}, z_{i\rightarrow j} = k, z_{i\leftarrow j} = k', \Omega) \quad (5.1)$$

where the elements N^{Θ} are defined in Eqs 5.2. Depending on the model considered, the predictive link distribution takes the following form:

$$p(y_{ij}|Y^{-ij}, Z^{-ij}, z_{i\rightarrow j} = k, z_{i\leftarrow j} = k', \Omega) =$$
$$\begin{cases} \left(\frac{N^{\Phi^{-ij}}_{1kk'}+\lambda_1}{N^{\Phi^{-ij}}_{.kk'}+\lambda_.}\right)^{y_{ij}} \left(1 - \frac{N^{\Phi^{-ij}}_{1kk'}+\lambda_1}{N^{\Phi^{-ij}}_{.kk'}+\lambda_.}\right)^{1-y_{ij}} & \textit{for MMSB} \\ \mathrm{NB}\left(y_{ij}; N^{Y^{-ij}}_{kk'} + r, \frac{p}{p\, N^{\Phi^{-ij}}_{.kk'}+1}\right) & \textit{for WMMSB} \end{cases}$$

The different count statistics N^* are estimated from the variational parameters $\gamma_{ijkk'}$ by:

$$N^{\Theta}_{\rightarrow ik} = \sum_{j,k'} \gamma_{ijkk'} \qquad N^{\Theta}_{\leftarrow jk'} = \sum_{i,k} \gamma_{ijkk'}$$
$$N^{\Phi}_{xkk'} = \sum_{ij:y_{ij}=x} \gamma_{ijkk'} \qquad N^{Y}_{kk'} = \sum_{ij} y_{ij}\gamma_{ijkk'} \tag{5.2}$$

In this inference scheme, γ_{ij} are the *local* parameters while the count statistics N^* represent the *sufficient* statistics (or global counts).

Finally, the model parameters can be recovered from their estimates as follows:

$$\hat{\theta}_{ik} = \frac{N^{\Theta}_{\rightarrow ik} + N^{\Theta}_{\leftarrow ik} + \alpha_k}{2N + \alpha_.} \qquad \hat{\phi}_{kk'} = \begin{cases} \frac{N^{\Phi}_{1kk'}+\lambda_1}{N^{\Phi}_{.kk'}+\lambda_.} & \textit{for MMSB} \\ \frac{p(N^{Y}_{kk'}+r)}{N^{\Phi}_{.kk'}-p+1} & \textit{for WMMSB} \end{cases}$$

5.5.1.1 Beta-Gamma augmentation

For WMMSB-bg model, we consider the following collapsed variational distribution:

$$q(\Pi) = q(\Theta, \Phi|Z, R, P)q(Z)q(R)q(P)$$

with $R = (r_{kk'}), P = (p_{kk'})$, $1 \leq k, k' \leq K$. As before, $q(z_{i\rightarrow j}, z_{i\leftarrow j}|\gamma_{ij})$ is multinomial with parameter γ_{ij}.

The same development as above applies for the parameters $\gamma_{ijkk'}$, given here also by Eq. 5.1. Furthermore, the predictive link probability now take the form:

$$p(y_{ij}|Y^{-ij}, Z^{-ij}, z_{i\to j} = k, z_{i\leftarrow j} = k', \Omega) \sim \\ \mathrm{NB}\left(y_{ij}; N_{kk'}^{Y^{-ij}} + \mathbb{E}_q[r_{kk'}], \frac{\mathbb{E}_q[p_{kk'}]}{\mathbb{E}_q[p_{kk'}]\, N_{.kk'}^{\Phi^{-ij}} + 1}\right)$$

and the block-block probability estimation takes the following form:

$$\hat{\phi}_{kk'} = \frac{\mathbb{E}_q[p_{kk'}](N_{kk'}^{Y} + \mathbb{E}_q[r_{kk'}])}{N_{.kk'}^{\Phi} - \mathbb{E}_q[p_{kk'}] + 1}$$

Setting $q(P) = p(P|Y, Z, \Omega)$ where p is the true distribution and exploiting the conjugacy of the Beta and the negative binomial distributions leads to a Beta distribution for $p_{kk'}$:

$$p_{kk'} \sim \mathrm{Beta}(c\epsilon + N_{kk'}^{Y}, c(1-\epsilon) + N_{kk'}^{\Phi}\mathbb{E}_q[r_{kk'}]) \tag{5.3}$$

so that:

$$\mathbb{E}_q[p_{kk'}] = \frac{c\epsilon + N_{kk'}^{Y}}{c\epsilon + N_{kk'}^{Y} + c(1-\epsilon) + N_{kk'}^{\Phi}\mathbb{E}_q[r_{kk'}]}$$

Lastly, as for its true distribution, the variational distribution for $r_{kk'}$ is taken in the Gamma family: $q(r_{kk'}) \sim \mathrm{Gamma}(a_{kk'}, b_{kk'})$. Even though $a_{kk'}$ can not be estimated explicitly, one only needs to have access to the expectation of $r_{kk'}$, that takes the following form:

$$\mathbb{E}_q[r_{kk'}] = \frac{r_0 c_0 + N_{kk'}^{Y}}{c_0 - N_{kk'}^{\Phi}\log(1 - p_{kk'})} \tag{5.4}$$

5.5.2 Stochastic Variational Inference with Stratified Sampling

Stochastic variational inference aims at optimizing ELBO through noisy yet unbiased estimates of its natural gradient computed on sampled data points. Different sampling strategies (Gopalan & Blei 2013; Kim et al. 2013) can be used. Following the study in (Gopalan & Blei 2013), we rely here on stratified sampling that allows one to control the number of links and non-links considered at each step of the inference process. For each node i, $1 \le i \le N$, one first constructs a set, denoted s_1^i, containing all the nodes to which i is connected to as well as M sets of equal size, denoted $s_0^{i,m}$, $1 \leq m \leq M$, each containing a sample of the nodes to which i is not connected to[1]. We will denote by S_0^i the set of all $s_0^{i,m}$ sets. Furthermore, we will denote by S_0 the union of all non-links set and S_1 the union of all links set. The sets thus obtained, for all nodes, constitute minibatches that can be sampled and used to update the global counts in Eq. 5.2. The combined scheme is summarized below:

1. Sample a node i uniformly from all nodes in the graph; with probability $\frac{1}{2}$, either select s_1^i or any set from S_0^i (in the latter case, the selection is uniform over the sets in S_0^i). We will denote by s_i the set selected and by $|s_i|$ its cardinality.

2. For each node $j \in s_i$, compute $\gamma_{ijkk'}$ through Eq. 5.1 and intermediate global counts acc. to:

$$\hat{N}^{\Theta}_{\leftarrow jk'} = \frac{1}{Cg(s_i)} \sum_k \gamma_{ijkk'}$$
$$\hat{N}^{\Theta}_{\rightarrow ik} \mathrel{+}= \frac{1}{|s_i|} \frac{1}{Cg(s_i)} \sum_{k'} \gamma_{ijkk'}$$
$$\hat{N}^{\Phi}_{.kk'} \mathrel{+}= \frac{1}{|s_i|} \frac{1}{Cg(s_i)} \gamma_{ijkk'}$$
$$\hat{N}^{Y}_{kk'} \mathrel{+}= \frac{1}{|s_i|} \frac{1}{Cg(s_i)} \gamma_{ijkk'} y_{ij}$$

   where C is a constant that is 2 for undirected graphs and 1 for directed graphs and $g(s_i) = \frac{1}{Nm}$ if $s_i \in S_0^i$ and $\frac{1}{N}$ otherwise. Note that $Cg(s_i)$ correspond to the probability to observe the node i depending on either s_i belongs to S_0 or S_1.

3. Update of the global counts (online version of Eq. 5.2):

$$N^{\Theta}_{\rightarrow ik} \leftarrow (1 - \rho_t^{i,\Theta}) N^{\Theta}_{\rightarrow ik} + \rho_t^{i,\Theta} \hat{N}^{\Theta}_{\rightarrow ik}$$
$$N^{\Theta}_{\leftarrow jk'} \leftarrow (1 - \rho_t^{i,\Theta}) N^{\Theta}_{\leftarrow jk'} + \rho_t^{i,\Theta} \hat{N}^{\Theta}_{\leftarrow jk'}$$
$$N^{\Phi}_{.kk'} \leftarrow (1 - \rho_t^{\Phi}) N^{\Phi}_{.kk'} + \rho_t^{\Phi} \hat{N}^{\Phi}_{.kk'}$$
$$N^{Y}_{kk'} \leftarrow (1 - \rho_t^{Y}) N^{Y}_{kk'} + \rho_t^{Y} \hat{N}^{Y}_{kk'}$$

4. $\rho_t^* = \frac{1}{(\tau+t)^\kappa}$ with $\kappa \in (0.5, 1]$.

[1] The sampling is here uniform over the nodes not connected to i with replacement; sampling without replacement led to poorer results in our experiments.

5. Go back to step 1 till convergence.

As one can note, the intermediate global counts correspond to a restriction, on minibatches, of the complete computation given in Eq. 5.2. The value of C is due to the fact that in undirected networks, each edge can be seen twice. The terms $\frac{1}{|s_i|}$ and $\frac{1}{Cg(s_i)}$ serve as a normalization in the gradient-like updates of the global counts (as there are more non-links than links, each non-link minibatch, representing a smaller fraction of the non-links, leads to more conservative updates). The "gradient steps" ρ^* are discussed below (Robbins-Monro condition).

For IMMSB, the procedure is silghtly different. The parameter N^Y does not exist for this model and the update coresponding to the count $N^{\Phi}_{.kk'}$ is replaced by updates of $N^{\Phi}_{xkk'}$ where $x = 1$ if the current point observed is a link as $y_{ij} = 1$ and $x = 0$ if it is a non-link as $y_{ij} = 0$.

Lastly, to be able to efficiently compute such quantities as $N^{\Phi^{-ij}}$ used for the computation of the link probability, one needs to store in memory, for each pair of nodes (i, j), a $K \times K$ matrix, which is not feasible for large networks. Thus, following (Foulds et al. 2013), we replace here $N^{\Phi^{-ij}}$ by N^{Φ} (and as well for N^Y and N^{Θ}), which amounts to assume that the contribution of each individual pair of nodes is negligible compared to all other pairs, a reasonable assumption when the network is large.

5.5.2.1 Robbins-Monro condition and implementation remarks

The convergence of stochastic variational inference is guaranteed under the Robbin-Monro condition (Robbins & Monro 1951) that imposes constraints on the gradient step, $\sum \rho_t = \infty$ and $\sum \rho_t^2 < \infty$ which can be obtained with $\rho_t = \frac{1}{(\tau+t)^{\kappa}}$ with $\kappa \in (0.5, 1]$. Thus, we maintain a gradient step for each of the global counts ρ^{Φ} and ρ^{Y} accounting respectively for N^{Φ} and N^Y. For N^{Θ}, we maintain individual gradient steps ρ^{Θ}_i for $1 \leq i \leq N$, following (Miller et al. 2009); this improved both convergence and prediction performance. Furthermore, to increase the speed of the inference, we update the global count N^{Φ} and N^Y only after a minibatch round. For the global count N^{Θ}, we update it after a burn-in period T_{burnin} such that $T_{burnin} \leq |S|$. This heuristic provides a trade-off between updating the global statistics after each observation, which slows down the inference and may result in bad local optima, and updating them only after minibatches that are potentially large (proportional to the number of nodes).

Within a Stratified sampling scheme, the network dataset is divided into $N(1+m)$ minibatches. The sampling uniformly choose between the minibatches of links S_1 and non-links S_0, with the parameters m controling the size of the non-links minibatch. The distribution of a minibatch S has the following distribution:

$$S \sim h(S; m) = \frac{1}{2N}\delta_{S_1} + \frac{1}{2Nm}\delta_{S_0}$$

where δ is the dirac operator, and $\delta_{S_1} = 1$ if $S \in S_1$ and 0 otherwise. One can see that the number of the non-link minibatches observed is in average m times lower that the number of the link minibatches. This is particularly interesting for sparse networks where the number of non-links is predominant over the number of links. As, the model is update after each minibatch, one could expect that the inference converge much before the total number of minibatches is reached which represents a great interest to scale the inference process to large networks[2]. This is in fact what we observed in our experiments where the model converge in general after observing a small proportion of the total of minibatches (5.6).

Our SCVI algorithm is summarized in the pseudo-code 1.

5.6 Experimental validation

We experimented our models on several real-world networks, directed and undirected. Their statistics and properties are summarized in Table 5.1 and detailed descriptions are available in the online Koblenz network collection[3]. For both astro-ph and hep-ph datasets, we used the cleaned version available in the graph-tool framework.

5.6.1 Experimental setup

As standard in social network analysis, the evaluation of the models is based on the missing link prediction task using the AUC-ROC score. For weighted models, we consider the probability that an edge exists between two unobserved nodes (i,j)

[2] If all the $N(1+m)$ minibatches are observed during the inference, the time complexity will be in O(Nˆ2).

[3] http://konect.uni-koblenz.de/networks/

Algorithm 1: SCVI pseudo-code.

```
Input: Random initialization of $N^{\Theta}, N^{\Phi}, N^{Y}$.
Output: $\hat{\Theta}, \hat{\Phi}$.
begin
    $t \leftarrow 0$
    while Convergence criteria not met do
        Sample a minibatch $S$ from $h(S; m)$.
        foreach $i, j \in S$ do
            Maximize local parameters $\gamma_{ij}$ from 5.1.
            if burn-in finished then
                Compute intermediate gradient $\hat{N}^{\Theta}$ from 5.2.
                Update global statistic $N^{\Theta}$.
                Update gradient step $\rho_t^{\Theta}$.
        Compute intermediate gradient $\hat{N}^{\Phi}$ and $\hat{N}^{Y}$ from 2.
        Update global statistics $N^{\Phi}$ and $N^{Y}$ from 3.
        Update gradient steps $\rho_t^{\Phi}$ and $\rho_t^{Y}$.
        Sample $P$ and $R$ from 5.3 and 5.4.
        $t \leftarrow t + 1$ .
```

belonging to the test set, namely:

$$p(y_{ij} \geq 1 | \hat{\Theta}, \hat{\Phi}) = 1 - \sum_{kk'} \hat{\theta}_{ik} \hat{\theta}_{jk'} e^{-\hat{\phi}_{kk'}}$$

For all the datasets, we built a test set by extracting randomly 20 percent of the edges of the network and about the same amount of non-links. The remaining data constitutes the “full” training set. Then, in order to assess how the models behave when few training data is available, we sub-sampled this full training set in order to obtain smaller sub-training sets (subgraphs) containing different proportions of the edges (i.e 1%, 5%, 10%, 20%, 30%, 50%, and 100%). Note that we ensure that all the sub-training sets are inclusive. We repeated this sampling 10 times with different seeds to cross validate our results. The average values (and standard deviations) computed on the ten sub-training sets are reported, for each proportion, as final results.

For deciding when to stop the inference process, 10% of the training set used serves as a validation set on which the log-likelihood is computed after each minibatch iteration. When the increase of the log-likelihood, averaged over the last 20 measures, is less than 0.001, the inference is stopped. The log-likelihood of a

Table 5.1: Network datasets used in the experiments. Type A is for co-authorship, type C is for communication (e.g. email exchange), type H is for hyperlinks, type L is for lexical network and I for interaction network (e.g money loan).

Datasets	Nodes	Edges	Density	Directed	Diameter	Weights			type
			$\times 10^{-3}$			mean	std	max	
astro-ph[1]	16,706	121,251	0.87	False	14	1.8	3.3	306	A
hep-th[2]	8,361	15,751	0.45	False	1	5.2	16	1226	A
moreno_names[3]	1,773	9,131	5.81	False	8	1.8	3.0	100	L
fb_uc[4]	1,899	20,296	5.63	True	4	2.8	4.7	98	C
digg_reply[5]	30,398	85,247	0.09	True	11	2.0	0.2	26	C
slashdot[6]	51,083	130,370	0.05	True	11	2.1	0.3	18	C
enron[7]	87,273	320,154	0.04	True	15	3.4	12.4	3904	C
wiki-link[8]	100,312	887,426	0.09	True	14	1.7	3.0	185	H
prosper-loans[9]	89,269	3,330,225	0.42	True	2	2.0	0.2	16	I

[1] http://konect.uni-koblenz.de/networks/ca-AstroPh. We used the cleaned version available in the graph-tool framework.
[2] hhttp://konect.uni-koblenz.de/networks/ca-cit-HepTh. We used the cleaned version available in the graph-tool framework.
[3] hhttp://konect.uni-koblenz.de/networks/moreno_names
[4] hhttp://konect.uni-koblenz.de/networks/opsahl-ucsocial
[5] hhttp://konect.uni-koblenz.de/networks/munmun_digg_reply
[6] hhttp://konect.uni-koblenz.de/networks/slashdot-threads
[7] hhttp://konect.uni-koblenz.de/networks/enron
[8] hhttp://konect.uni-koblenz.de/networks/link-dynamic-simplewiki
[9] hhttp://konect.uni-koblenz.de/networks/prosper-loans

given set of observations $\mathcal{D}_{set}$ is given by:

$$\log p(\mathcal{D}_{set}) = \sum_{i,j \in \mathcal{D}_{set}} \log p(y_{ij}|\hat{\phi}_{kk'})p(k|\hat{\theta}_i)p(k'|\hat{\theta}_j)$$

For all our models, the gradient step parameters τ and κ were fixed respectively to 1024 and 0.5, the burn-in period T_{burnin} to 150; for stratified sampling, M was set to 50, the size of $s_0^{i,m}$, $1 \leq m \leq M$ being equal to the number of nodes to which i is not connected to divided by M. For MMSB, the hyper-parameters λ_0 and λ_1 were set to 0.1. For WMMSB, the shape and scale parameters r et p were fixed to 1 and for WMMSB, the beta-gamma hyper-parameters were fixed to $c_0 = 10$, $r_0 = 1$, $c = 100$ and $\epsilon = 10^{-6}$. The number of latent classes K was fixed to 10 for all models and the latent-class hyper-parameters α_k to $\frac{1}{K}$. Our implementation is available online[4]. In addition, we consider here two standard link prediction models, the stochastic block model, referred to as SBM, and its weighted extension, referred to as WSBM. For these two models, the microcanonical stochastic block model implementation of (Peixoto 2018) has been used since it integrates an efficient MCMC inference method for the stochastic block model family. The number of classes was also set to $K = 10$.

Variational inference, used here for MMSB models, and MCMC, used for SBM

[4] https://github.com/dtrckd/pymake

models, lead to different performance, the latter usually yielding better models than the former (Asuncion et al. 2009). Indeed, despite the fact that the MMSB models considered here rely on more realistic assumptions regarding the distribution of nodes over latent classes, the approximations made on the likelihood for scalable inference purposes penalize MMSB models when it comes to prediction accuracy. This said, the strong averaging step of the stochastic gradient descent allows for faster convergence so that, as the models are more realistic, they may yield better performance when the amount of training data is limited. This is indeed what we observe in practice.

5.6.2 Results

Figure 5.1 gives the AUC/ROC scores for the different models when using 1%, 5%, 10%, 20%, 30% and 50% of the training data, for 6 networks (the complete results, over all training set size and networks are given in the appendix 7). As one can note, MMSB models outperform the other models when the amount of training data is limited. Among these models, WMMSB-bg is the best performing one, which highlights the importance of the Beta and Gamma priors used. The poor performance of MMSB on some networks can be explained by the fact that the convergence of the model is very sensitive to the sampling choices done during the online inference, as illustrated by the high variance in the results. When the amount of training data is sufficient (which depends on the network considered), SBM models tend to be better. As discussed before, we attribute this to the MCMC method used in SBM models. Surprisingly, and contrary to what is happening for MMSB models, WSBM does not really outperform SBM; this model does not seem to be able to make a good use of the edge covariates.

Table 5.2, which displays the results of MMSB, WMMSB-bg, SBM and WSBM for all networks when using 10% and 100% of the training data, confirms these elements. As one can note, using all training data, SBM outperforms WSBM on 5 datasets. Interestingly, there is an important degradation for SBM models when only 10% of the training set is used. MMSB models are more stable in this aspect, showing that the stochastic variational inference used in MMSB models allows one to learn a correct model with few data.

Finally, it is worth mentioning that on the dataset prosper-loans, the only network classified as “Interaction” in the Konnect repository, most models fail to learn the topology. In particular, MMSB barely exceeds a random classifier. Only

Table 5.2: Comparison of MMSB, WMMSB-bg, SBM and WSBM in terms of AUC-ROC when using 10% and 100% of the training data.

	10%				100%			
	MMSB	WMMSB-bg	SBM	WSBM	MMSB	WMMSB-bg	SBM	WSBM
astro-ph	**708** ± 3	700 ± 30	594 ± 16	586 ± 9	**716** ± 11	710 ± 18	701 ± 6	705 ± 5
hep-th	**617** ± 11	579 ± 12	480 ± 9	482 ± 26	675 ± 8	676 ± 8	**779** ± 10	714 ± 7
moreno_names	680 ± 72	**707** ± 29	571 ± 29	594 ± 30	738 ± 33	739 ± 7	**862** ± 7	859 ± 11
fb_uc	732 ± 127	**827** ± 8	726 ± 20	788 ± 18	784 ± 140	850 ± 20	**902** ± 2	896 ± 2
digg_reply	485 ± 178	**651** ± 127	551 ± 47	582 ± 35	482 ± 204	**744** ± 15	728 ± 26	717 ± 17
slashdot	519 ± 193	**820** ± 6	721 ± 66	732 ± 81	634 ± 181	791 ± 11	830 ± 16	**834** ± 12
enron	459 ± 289	875 ± 14	870 ± 80	**923** ± 14	529 ± 256	835 ± 8	799 ± 20	**853** ± 63
wiki-link	491 ± 242	739 ± 73	848 ± 4	**850** ± 4	432 ± 185	785 ± 8	**925** ± 2	915 ± 3
prosper-loans	548 ± 284	**752** ± 11	466 ± 57	455 ± 44	434 ± 274	**727** ± 30	500 ± 4	504 ± 6

the weighted MMSB models, WMMSB and WMMSB-bg, succeed in predicting new edges, with a performance above 0.75 when using only 10% of the training data.

5.6.3 Convergence analysis

Figure 5.2 shows the evolution of the log-likelihood for the MMSB-based models on a validation set composed of 20% of links and non-links for each network. We used three different sets for the hyper-parameters shape r and scale p of WMMSB. Regardless of the values of these hyper-parameters, one can observe that the augmented model WMMSB-bg is less prone to overfitting, usually converges to a better solution and only needs a small proportion of the total number N^2 of edges to do so.

5.7 Conclusion

We exposed in this chapter a new model, from the mixed-membership stochastic block model family, to deal with (directed or undirected) weighted networks. We furthermore showed that this model can be efficiently learned through a stochastic collapsed variational approach that couples collapsed variational and stochastic inference, so that the model can be deployed on networks comprising millions of edges. Experiments conducted on several networks showed that the proposed model can successfully predict the topology of various real-world networks and that it outperforms the standard mixed-membership stochastic block model (with the same, scalable inference). Another interesting property of this model is the fact that it outperforms other stochastic block models when the mount of training data is limited.

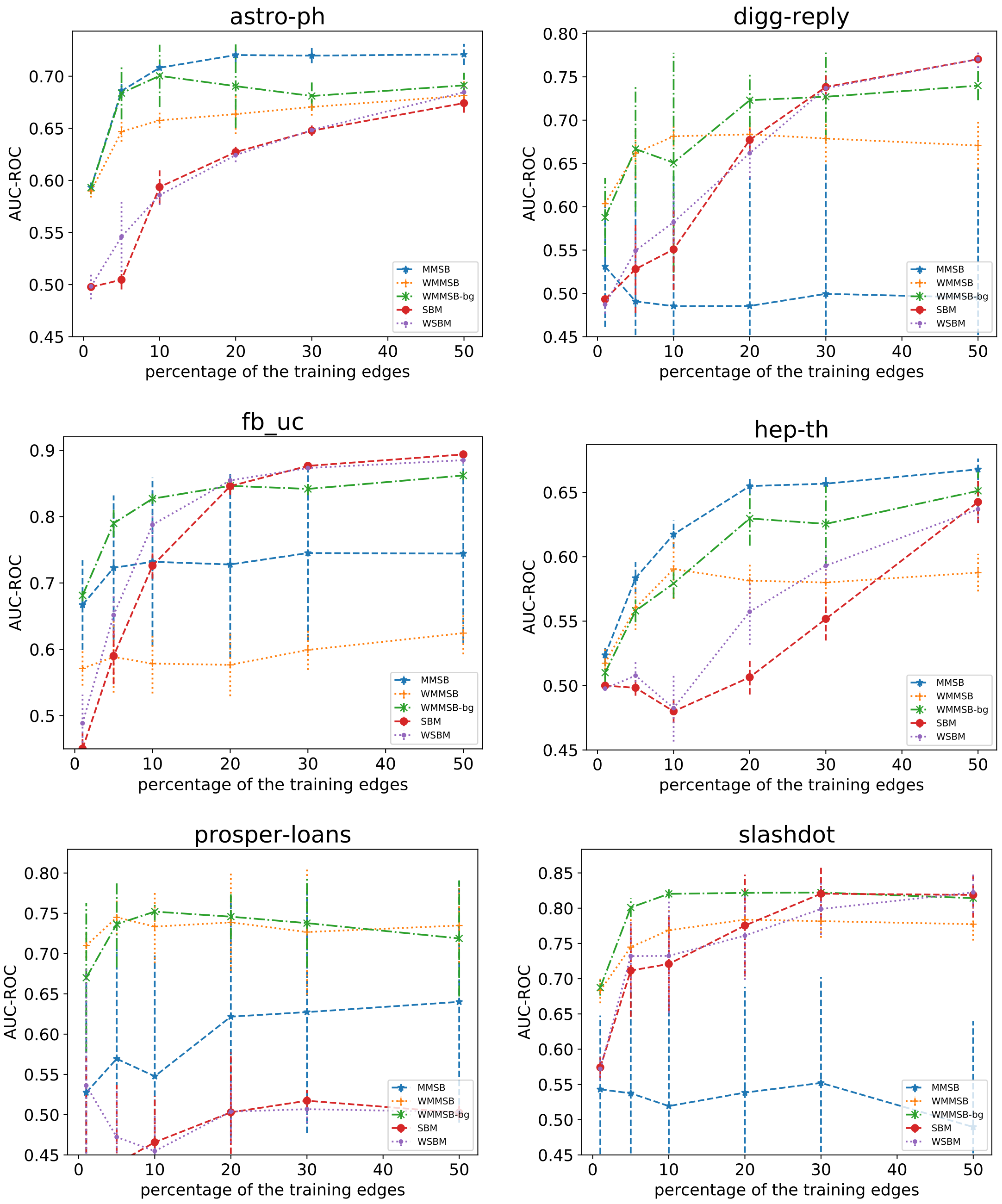

Figure 5.1: Comparison of models in terms of AUC-ROC scores according to the percentage of edges used to train the models (from 1 to 50%).

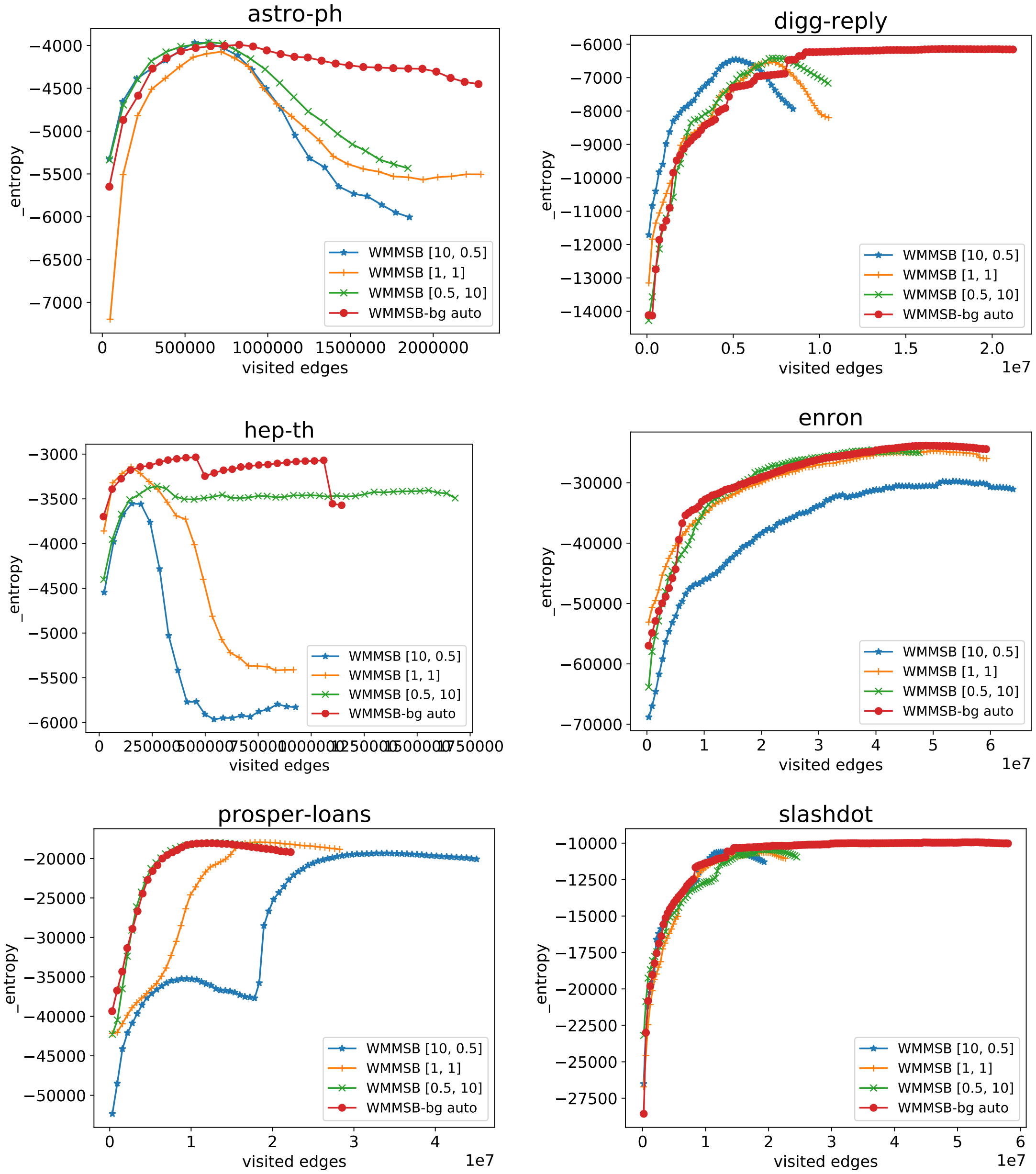

Figure 5.2: Log-likehood convergence for WMMSB and WMMSB-bg models on a test set containing 20% of the edges of the networks. Three different set of hyperparmeters are used for WMMSB.

Chapter 6

Conclusion

Our world is composed with all kinds of objects that are interconnected. They can be represented by networks which capture the relational information between those objects. Among them, social and information networks exhibit common properties that are used to characterize them. These networks can be analyzed with probabilistic models to infer their structures and make predictions. In particular, the class of mixed-membership models offers a rich framework to capture overlapping community structures with shared components in a probabilistic setting. We studied two characteristic nonparametric models in this class, namely the IMMSB and ILFM that respectively represents latent class models and latent feature models. More precisely, we ask whether those models comply with other important properties found in real-world networks. We showed that those models can be configured to learn latent variables that either capture a homophilic mixing pattern or not, and in consequence the models are flexible with regard to the detection of community with the regular equivalence or the structural equivalence. Moreover, we show that those models do not satisfy the preferential attachment effect in the global case, but that IMMSB can satisfy a local preferential attachment when the degree are considered within a community only. We conducted experiments that empirically emphasize that IMMSB may exhibit better performance for real-world networks when few data are observed. We further extended MMSB for weighted networks, to make it compatible with a wider range of large-scale networks, and empirically showed that it constitutes a good choice for real-world networks, again when few data are observed.

Our work could be extended in several directions. First, the models can be extended in order to be trained with more general types of networks. Indeed, slight modifications of the kernel likelihood in the mixed-membership models are necessary to model multi-relational graphs, or even better multi-relational weighted graphs. This could be particularly interesting for information networks such as lexical networks and ontologies whereas the type of the edges can often be described with categorical data (one can think about the grammatical relation between words). A more profound research direction concerns the relaxation of the exchangeability assumption of graphs. An important limitation of this assumption is that the graphs generated by such models are either dense or empty. This question is especially interesting because it is related to temporal graphs modeling and how we can relax the exchangeability assumptions in order to model sparse graphs. Indeed, when the data distribution is dependent on the order on which we observe edges and nodes, the model is no more exchangeable and one can interpret this as a temporal dependency. A crucial question in this direction is to find the form of the invariants, or the symmetry groups, that are relevant to temporal networks as well as the corresponding model representations.

Appendix 1: WMMSB inference derivations and complete results

6.1 Derivation of the collapsed variational updates

The derivation of the collapsed variational updates is first obtained by maximizing the ELBO w.r.t $\gamma_{ijkk'}$ with:

$$\begin{aligned}\frac{\partial \mathcal{L}_Z}{\partial \gamma_{ijkk'}} = \frac{\partial}{\partial \gamma_{ijkk'}} \sum_{Z^{-ij}} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} q(Z^{-ij}) \gamma_{ijk_1k_2} (\log p(Y, Z^{-ij}, z_{i\to j} = k_1, z_{i\leftarrow j} = k_2|\Omega) + \\ \log q(Z^{-ij}, z_{i\to j} = k_1, z_{i\leftarrow j} = k_2)) \\ = E_{q(Z^{-ij})}[p(Y, Z^{-ij}, z_{i\to j} = k, z_{i\leftarrow j} = k'|\Omega))] + H[Z^{-ij}] - \log(\gamma_{ijkk'}) + 1\end{aligned}$$

By equating this derivative to zero, one obtains the following update:

$$\gamma_{ijkk'} \propto \exp E_{q(Z^{-ij})}[\log P(z_{i\to j} = k, z_{i\leftarrow j} = k'|Y^{-ij}, Z^{-ij}, \Omega)] \qquad (6.1)$$

with $P(z_{i\to j} = k, z_{i\leftarrow j} = k'|Y^{-ij}, Z^{-ij}, \Omega)$ being the collapsed Gibbs update of WMMSB, of the form:

$$\begin{aligned}P(z_{i\to j} = k, z_{i\leftarrow j} = k'|Y^{-ij}, Z^{-ij}, \Omega) \propto \\ (n^{\Theta^{-j}}_{\to ik} + \alpha_k)(n^{\Theta^{-i}}_{\leftarrow jk} + \alpha_{k'}) \mathrm{NB}\left(y_{ij}; n^{Y^{-ij}}_{kk'} + r, \frac{p}{p\, n^{\Phi^{-ij}}_{.kk'} + 1}\right)\end{aligned}$$

with count statistics given by the following equations:

$$
\begin{aligned}
n^{\Theta}_{\rightarrow ik} &= \sum_j \delta(z_{i\rightarrow j} = k) \\
n^{Y}_{kk'} &= \sum_{ij} y_{ij}\delta(z_{i\rightarrow j} = k, z_{i\leftarrow j} = k') \\
n^{\Phi}_{.kk'} &= \sum_{ij} \delta(z_{i\rightarrow j} = k, z_{i\leftarrow j} = k')
\end{aligned}
$$

By applying a first order Taylor expansion on Eq.(6.1), following (Teh et al. 2007), one obtains:

$$
\begin{aligned}
\gamma_{ijkk'} \propto & (E_{q(Z^{-ij})}[n^{\Theta^{-j}}_{\rightarrow ik}] + \alpha_k)(E_{q(Z^{-ij})}[n^{\Theta^{-i}}_{\leftarrow jk}] + \alpha_{k'}) \\
& \times \mathrm{NB}\left(y_{ij}; E_{q(Z^{-ij})}[n^{Y^{-ij}}_{kk'}] + r, \frac{p}{p\, E_{q(Z^{-ij})}[n^{\Phi^{-ij}}_{.kk'}] + 1}\right)
\end{aligned}
$$

Finally, using a Gaussian approximation (as in *e.g.* (Asuncion et al. 2009)), one can estimate the expectations $E_{q(Z^{-ij})}[n^{\Theta^{-j}}_{\rightarrow ik}]$, $E_{q(Z^{-ij})}[n^{\Theta^{-i}}_{\leftarrow jk}]$ and $E_{q(Z^{-ij})}[n^{\Phi^{-ij}}_{.kk'}]$ with the counts defined in Eq. 5.2.

6.2 Beta-Gamma updates

In the WMMSB-bg model, the collapsed variational distribution takes the form:

$$
q(\Pi) = q(\Theta, \Phi | Z, R, P) q(Z) q(R) q(P)
$$

The variational distribution for $r_{kk'}$ is taken in the Gamma family: $q(r_{kk'}) = \mathrm{Gamma}(a_{kk'}, b_{kk'})$ for $1 \leq k, k' \leq K$. The collapsed ELBO can thus be rewritten as:

$$
\begin{aligned}
\log p(Y) \geq \mathcal{L}_{Z,R,P} &= \mathbb{E}_q[\log p(Y, Z, R, P|\Omega)] + \mathrm{H}[q(Z)] + \mathrm{H}[q(R)] + \mathrm{H}[q(P)] \\
&= \mathbb{E}_q[\log p(Y, Z)] + \mathrm{H}[q(Z)] \\
&\quad + \mathbb{E}_q[\log p(R|Y, Z, P)] + \mathrm{H}[q(R)] \\
&\quad + \mathbb{E}_q[\log p(P|Y, Z)] + \mathrm{H}[q(P)]
\end{aligned}
$$

6.2.1 Optimizing $\gamma_{ijkk'}$

In the Beta-Gamma augmentation, the parameters p and r are marginalized in the update given by Eq. (6.1):

$$\gamma_{ijkk'} \propto \exp \mathbb{E}_{q(Z^{-ij})}[\log \mathbb{E}_{q(r_{kk'})}[E_{q(p_{kk'})}[P(z_{i\to j}=k, z_{i\leftarrow j}=k'|Y^{-ij}, Z^{-ij}, \Omega)]]]$$

By using a first order Taylor expansion, one obtains:

$$\gamma_{ijkk'} \propto (N^{\Theta^{-j}}_{\to ik} + \alpha_k)(N^{\Theta^{-i}}_{\leftarrow jk'} + \alpha_{k'})\mathrm{NB}\left(y_{ij}; N^{Y^{-ij}}_{kk'} + \mathbb{E}_q[r_{kk'}], \frac{\mathbb{E}_q[p_{kk'}]}{\mathbb{E}_q[p_{kk'}]\, N^{\Phi^{-ij}}_{.kk'} + 1}\right)$$

6.2.2 Optimizing $r_{kk'}$

We isolate the part of the ELBO than depends only on $r_{kk'}$ parameters ($a_{kk'}$ and $b_{kk'}$). Thus, we consider only the links that have been generated within the classes k, k', denoted by $Y^{(kk')}$. Furthermore, as $y_{ij} \sim NB(r_{kk'}, p_{kk'})$ if i is in class k and j in class k', one has:

$$\mathcal{L}_{[r_{kk'}]} = \mathbb{E}_{q(r_{kk'})}[\log p(r_{kk'}|Y^{(kk')}, Z^{(kk')}, p_{kk'})] + \mathrm{H}[q(r_{kk'})]$$

By applying Bayes rules and dropping the normalizing term that does not depend on $r_{kk'}$, one gets:

$$\begin{aligned}
\mathcal{L}_{[r_{kk'}]} &= \mathbb{E}_{q(r_{kk'})}[\log \left(p(Y^{(kk')}|Z^{(kk')}, r_{kk'}, p_{kk'})p(r_{kk'}])\right)] + \mathrm{H}[q(r_{kk'})] \\
&= \mathbb{E}_{q(r_{kk'})}[\log \left(\prod_{ij\in Y^{(kk')}} \binom{r_{kk'} + y_{ij} - 1}{y_{ij}} (1-p_{kk'})^{r_{kk'}} p_k^{y_{ij}} p(r_{kk'})\right)] + \mathrm{H}[q(r_{kk'})] \\
&= \mathbb{E}_{q(r_{kk'})}[\log \left((1-p_{kk'})^{r_{kk'} N^{\Phi}_{kk'}} p_{kk'}^{N^Y_{kk'}} p(r_{kk'}) \prod_{ij\in Y^{(kk')}} \frac{\Gamma(r_{kk'} + y_{ij})}{\Gamma(r_{kk'})\Gamma(y_{ij}+1)}\right)] + \mathrm{H}[q(r_{kk'})]
\end{aligned}$$

If $y_{ij} = 0$, then $\frac{\Gamma(r_{kk'}+y_{ij})}{\Gamma(r_{kk'})\Gamma(y_{ij}+1)} = 1$, whereas if $y_{ij} \neq 0$, then $\frac{\Gamma(r_{kk'}+y_{ij})}{\Gamma(r_{kk'})\Gamma(y_{ij}+1)} = \frac{1}{B(r_{kk'}, y_{ij})y_{ij}}$. Furthermore, in this latter case:

$$B(r_{kk'}, y_{ij}) = \int_0^1 t^{r_{kk'}-1}(1-t)^{y_{ij}-1}dt \leq \int_0^1 t^{r_{kk'}-1}dt = \frac{1}{r_k}$$

so that:

$$\log \prod_{ij \in Y^{(kk')}} \frac{\Gamma(r_{kk'} + y_{ij})}{\Gamma(r_{kk'})\Gamma(y_{ij}+1)} \geq N^Y_{kk'} \log(r_{kk'}) + \text{cst}$$

with $N^Y_{kk'} = \sum_{ij \in Y^{(kk')}} y_{ij}$. Furthermore, from the model definitions, one has

$$\log p(r_{kk'}) = (r_0 c_0 - 1)\log(r_{kk'}) - r_{kk'} c_0 + \text{cst} \ ,$$

and

$$\text{H}[q(r_{kk'})] = a_{kk'} + \log(b_{kk'}) + \log\Gamma(a_{kk'}) + (1 - a_{kk'})\Psi(a_{kk'}) \ .$$

Hence:

$$\begin{aligned}\mathcal{L}_{[r_{kk'}]} \geq{} & N^\Phi_{kk'} a_{kk'} b_{kk'} \log(1 - p_{kk'}) + (r_0 c_0 - 1)(\Psi(a_{kk'}) + \log(b_{kk'})) - c_0 a_{kk'} b_{kk'} \\ & + N^Y_{kk'}(\Psi(a_{kk'}) + \log(b_{kk'})) + a_{kk'} + \log(b_{kk'}) + \log\Gamma(a_{kk'}) + (1 - a_{kk'})\Psi(a_{kk'})\end{aligned}$$

Maximizing the right-hand term of the above inequality with respect to $b_{kk'}$ yields:

$$b_{kk'} = \frac{r_0 c_0 + N^Y_{kk'}}{a_{kk'}(c_0 - N^\Phi_{kk'} \log(1 - p_{kk'}))}$$

As $r_{kk'} \sim \text{Gamma}(a_{kk'}, b_{kk'})$, one finally obtains:

$$\mathbb{E}_q[r_{kk'}] = a_{kk'} b_{kk'} = \frac{r_0 c_0 + N^Y_{kk'}}{c_0 - N^\Phi_{kk'} \log(1 - p_{kk'})}$$

6.2.3 Optimizing $p_{kk'}$

In oder to maximize the ELBO w.r.t $p_{kk'}$, one can let $q(p_{kk'}) = p(p_{kk'}|Y, Z) = E_q(r_{kk'})[p(p_{kk'}|Y^{(kk')}, Z^{(kk')}, r_{kk'})]$. As the negative binomial and Beta distributions are conjugate, a closed-form expression can be obtained:

$$\begin{aligned}p(p_{kk'}|Y^{(kk')}, Z^{(kk')}, r_{kk'}) &\propto p(Y^{(kk')|Z^{(kk')}, r_{kk'}} p(r_{kk'}) \\ &\propto (1 - p_{kk'})^{r_{kk'} N^\Phi_{kk'}} p_{kk'}^{N^Y_{kk'}} p_{kk'}^{c\epsilon - 1} (1 - p_{kk'})^{c(1-\epsilon)-1} \\ &\propto p_{kk'}^{c\epsilon + N^Y_{kk'} - 1} (1 - p_{kk'})^{c(1-\epsilon) + N^\Phi_{kk'} r_{kk'} - 1} \\ &= \text{Beta}(c\epsilon + N^Y_{kk'}, c(1 - \epsilon) + N^\Phi_{kk'} r_{kk'})\end{aligned}$$

Finally, by resorting again to a first order Taylor expansion, one obtains:

$$p_{kk'} \sim \text{Beta}(c\epsilon + N^{Y}_{kk'}, c(1-\epsilon) + N^{\Phi}_{kk'} E_q[r_{kk'}])$$

6.3 Experimentation (full results)

We provide here the complete set of results for the AUC-ROC scores evaluations for the full range of training sets proportions (1%, 5%, 10%, 20%, 30%, 50% and 100%) in Figure 6.1 for all the datasets. The log-likelihood convergence of the inference for all the datasets are given in Figure 6.2,

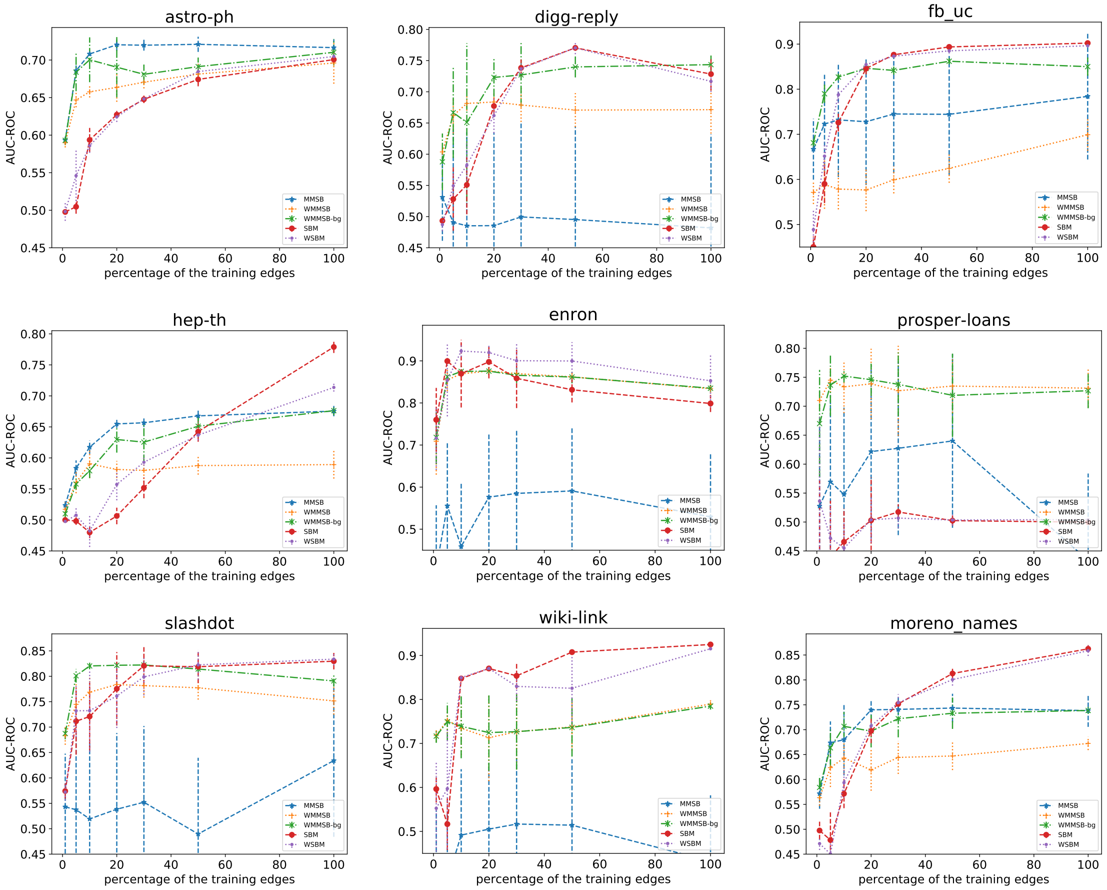

Figure 6.1: Comparison of models in terms of AUC-ROC scores according to the percentage of edges used to train the models (from 1 to 100%).

6.3.1 Reproducible Research

We published our implementation within a platform that aims to ease the development of reproducible complex experiments. We are maintaining this platform that

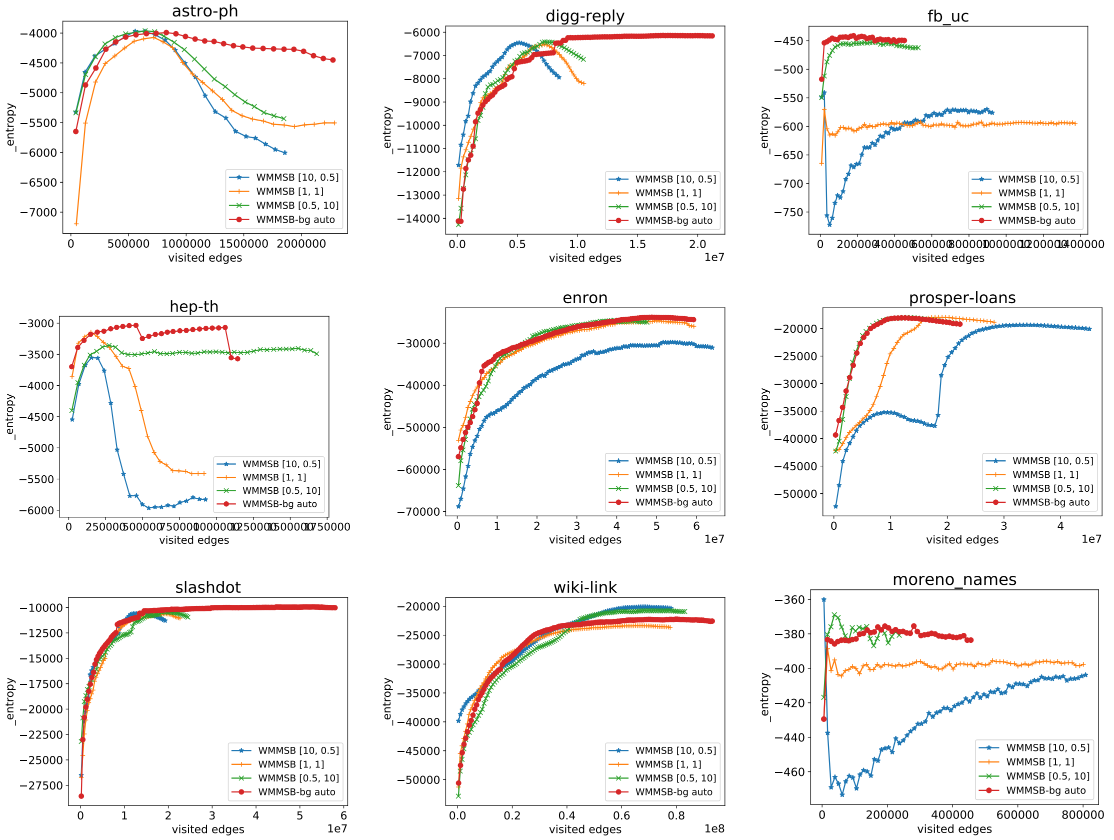

Figure 6.2: Log-likehood convergence for WMMSB and WMMSB-bg models on a test set containing 20% of the edges of the networks. Three different sets of hyperparmeters are used for WMMSB.

we released under open-source license.

To reproduce our results, one can proceed as follows: * Install the Pymake project:

```
$ git clone https://github.com/dtrckd/pymake
$ cd pymake && make install
```

- Fit all the models on all the corpus, and save the results:

```
$ pmk online_roc -x fit -w --repeat 0 1 2 3 4 5 6 7 8 9
```

- Parallelization can be obtained by adding the options ——cores NUMBER_OF_CORES,
- Figures can be plotted with the command:

```
$ pmk online_roc -x roc_evolution2 --repeat 0 1 2 3 4 5 6 7 8 9
```

Chapter 7

Appendix 2: Gibbs updates for ILFM

We change the variable name from Θ, Φ to F, W in order to have consistent notations with the original ILFM paper.

The goal of the inference is to recover the posterior densities $P(W, F \mid Y) \propto P(Y \mid F, W)P(F)P(W)$.

We use a MCMC sampling approach to learn the hidden variables of the model F and W. Namely we use a Gibbs sampling for features matrix F and a Metropolis-Hasting for sampling the new features from the IPB as well as the weight matrix W since it is not in a conjugate of the likelihood. The sampling procedure is summarized in Algorithm 2.

7.1 Feature updates

The learning of the feature matrix F is computed in 2 stages. For each entity we need to update each non-unique features [1]. Then, for the unique features of this entity we need to add possibly k^{new} features.

The sampling of each f_{ik} is given by the following conditional posterior:

$$P(f_{ik} = 1 \mid Y, F_{-(ik)}, W) = \frac{P(Y \mid W, F_{-(ik)}, f_{ik} = 1)P(f_{ik} = 1 \mid F_{-(ik)})}{\sum_{f_{ik}} P(Y \mid W, F_{-(ik)}, f_{ik})P(f_{ik} \mid F_{-(ik)})} \tag{7.1}$$

[1] Non-unique means that feature belongs to more than one entity.

The result from the IPB give us the following results for the Gibbs update for the feature k:

$$P(f_{ik} = 1 \mid F_{-(ik)}) = \frac{m_{-i,k}}{N} \tag{7.2}$$

$$P(f_{ik} = 0 \mid F_{-(ik)}) = 1 - \frac{m_{-i,k}}{N} \tag{7.3}$$

where $m_{-i,k}$ represents the number of active features k for all entities excluding entity i, hence $m_{-i,k} = \sum_{j \neq i} f_{jk}$. After sampling k features for a particular entity i, we need to evaluate the new features who should be associated with this entity i. The probability to have k_i^{new} features is proportional to this density:

$$P(k_i^{new} \mid Y, F, \alpha) \propto P(Y \mid F^{new}) P(k_i^{new} \mid \alpha) \tag{7.4}$$

The probability of having k_i^{new} features is drawn from a Poisson$(\frac{\alpha}{\mathrm{N}})$ distribution in the IPB process. However, we also need to sample the $\mathbf{w}_i$ weights associated with those features, and in our case, the density of weights are not conjugate of the likelihood. In consequence, we cannot integrate them out as explicitly assumed in the equation (7.4). We follow the approach of (Meeds et al. 2006) to jointly sample the new features and the weights using a Metropolis-Hasting method. Thanks to the exchangeability of the IPB we only need to consider features unique to entity i to either propose or reject new features. We reference by a superscript B the set of model parameters corresponding to unique features. A convenient choice of jumping distribution is:

$$J(k_i^{new}, \mathbf{w}_i^{new} \mid k_i^B, \mathbf{w}_i^B) = \mathrm{Poisson}(k_i^{new} \mid \alpha) \mathcal{N}(\mathbf{w}_i^{new} \mid \sigma_w) \tag{7.5}$$

The acceptance ratio can thus be written as:

$$r_{B \to new} = \frac{P(k_i^{new}, \mathbf{w}_i^{new} \mid Y, W, F, \alpha) J(k_i^B, \mathbf{w}_i^B \mid k_i^{new}, \mathbf{w}_i^{new})}{P(k_i^B, \mathbf{w}_i^B \mid Y, W, F, \alpha) J(k_i^{new}, \mathbf{w}_i^{new} \mid k_i^B, \mathbf{w}_i^B)} \tag{7.6}$$

When replacing by equation (7.4) and (7.5), the acceptance ratio simplify to the ratio of data likelihoods:

$$r_{B \to new} = \frac{P(Y \mid F^{new}, W^{new})}{P(Y \mid F, W)} \tag{7.7}$$

7.2 Weight updates

The learning of the weight matrix W is approximated using a Metropolis-Hasting algorithm. Thus, we sample sequentially each weight corresponding to non-zeros features interaction.

$$P(w_{kl} \mid Y, F, W_{-kl}, \sigma_w) \propto P(Y \mid F, W)P(w_{kl} \mid \sigma_w) \tag{7.8}$$

We choose a jumping distribution in the same family of our prior on weight centered around the previous sample:

$$J(w^*_{kl} \mid w_{kl}) = \mathcal{N}(w_{kl}, \eta) \tag{7.9}$$

with η a parameter letting us controlling the acceptance ratio.

The acceptance ratio of w^*_{kl} is thus:

$$r_{w_{kl} \to w^*_{kl}} = \frac{P(Y \mid F, W^*)P(w^*_{kl} \mid \sigma_w)J(w_{kl} \mid w^*_{kl})}{P(Y \mid F, W)P(w_{kl} \mid \sigma_w)J(w^*_{kl} \mid w_{kl})} \tag{7.10}$$

7.3 Hyper-parameter optimization

Optimization rules for α is given in 3.3.6.1.

Algorithm 2: Parameters sampling of ILFM for one iteration step.

```
Input: $Y$, $\alpha$, $\sigma_w$, $\eta$
Initialize: $F$, $W$, randomly
foreach entities $i \in \{1, .., N\}$ do
    foreach represented features $k \in \{1, .., K^+\}$ do
        if $m_{-i,k} > 0$ then
            Update $f_{ik}$ using equantion (7.1)
    Draw candidate for $k_i^{new}$ and $\mathbf{w}_i^{new}$ using equation (7.5)
    Accept candidate with probability $\min(1, r_{B \to new})$
foreach weights $w_{kl} \in W$ do
    Draw candidate for $w^*_{kl}$ using equation (7.9)
    Accept candidate with probability $\min(1, r_{w_{kl} \to w^*_{kl}})$
Sample $\alpha$ from eq. (3.7)
```

Chapter 8

References

Aaron Clauset, E.T. & Sainz, M., 2016. The colorado index of complex networks. Available at: https://icon.colorado.edu/.

Ackley, D.H., Hinton, G.E. & Sejnowski, T.J., 1985. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1), pp.147–169.

Adamic, L.A. & Glance, N., 2005. The political blogosphere and the 2004 us election: Divided they blog. In *Proceedings of the 3rd international workshop on link discovery*. ACM, pp. 36–43.

Aggarwal, C.C. ed., 2011. *Social network data analytics*, Springer.

Aicher, C., Jacobs, A.Z. & Clauset, A., 2014. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2), pp.221–248.

Aiello, L.M. et al., 2012. Friendship prediction and homophily in social media. *ACM Trans. Web*, 6(2), pp.1–33.

Aiello, W., Chung, F. & Lu, L., 2001. A random graph model for power law graphs. *Experimental Mathematics*, 10(1), pp.53–66.

Airoldi, E.M. et al., 2014. *Handbook of mixed membership models and their applications*, CRC Press.

Airoldi, E.M. et al., 2009. Mixed membership stochastic blockmodels. In *Advances in neural information processing systems*. pp. 33–40.

Albert, R. & Barabási, A.-L., 2002. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1), p.47.

Aldous, D.J., 1981. Representations for partially exchangeable arrays of random variables. *Journal* *of Multivariate Analysis*, 11(4), pp.581–598.

Al Hasan, M. & Zaki, M.J., 2011. A survey of link prediction in social networks. In *Social network data analytics*. Springer, pp. 243–275.

Amari, S.-i. & Nagaoka, H., 2007. *Methods of information geometry*, American Mathematical Soc.

Antoniak, C.E., 1974. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pp.1152–1174.

Asta, D.M. & Shalizi, C.R., 2015. Geometric network comparisons. In *Proceedings of the thirty-first conference on uncertainty in artificial intelligence, UAI 2015, july 12-16, 2015, amsterdam, the netherlands*. pp. 102–110.

Asuncion, A. et al., 2009. On smoothing and inference for topic models. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. pp. 27–34.

Barabási, A., 2003. *Linked - how everything is connected to everything else and what it means for business, science, and everyday life*, Plume.

Barabási, A.-L., 2011. *Bursts: The hidden patterns behind everything we do, from your e-mail to bloody crusades*, Plume, Penguin Book, USA.

Barabási, A.-L. & Albert, R., 1999. Emergence of scaling in random networks. *Science*, 286(5439), pp.509–512.

Barabási, A.-L. & others, 2016. *Network science*, Cambridge university press.

Batagelj, V. & Mrvar, A., 2006. Pajek datasets. Available at: http://vlado.fmf.uni-lj.si/pub/networks/data/.

Bell, R.M. & Koren, Y., 2007. Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, 9(2), pp.75–79.

Bickel, P.J. & Chen, A., 2009. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50), pp.21068–21073.

Blackwell, D., MacQueen, J.B. & others, 1973. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2), pp.353–355.

Blei, D.M., Kucukelbir, A. & McAuliffe, J.D., 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), pp.859–877.

Blei, D.M., Ng, A.Y. & Jordan, M.I., 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), pp.993–1022.

Blondel, V.D. et al., 2008. Fast unfolding of community in large networks. *Journal of statistical mechanics: theory and experiment*, 10, p.P10008.

Bollacker, K. et al., 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 acm sigmod international conference on management of data*. AcM, pp. 1247–1250.

Bollobás, B., 1998. Random graphs. In *Modern graph theory*. Springer, pp. 215–252.

Bollobás, B., 1981. The diameter of random graphs. *Transactions of the American Mathematical Society*, 267(1), pp.41–52.

Bollobás, B. ela et al., 2001. The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18(3), pp.279–290.

Bollobás, B. & Riordan, O., 2011. Sparse graphs: Metrics and random models. *Random Structures & Algorithms*, 39(1), pp.1–38.

Bollobás, B. & Riordan, O., 2004. The diameter of a scale-free random graph. *Combinatorica*, 24(1), pp.5–34.

Borgatti, S.P. & Everett, M.G., 1992. Notions of position in social network analysis. *Sociological methodology*, pp.1–35.

Borgs, C. et al., 2014. An L^p theory of sparse graph convergence i: Limits, sparse random graph models, and power law distributions. *arXiv preprint arXiv:1401.2906*.

Breiger, R.L., Boorman, S.A. & Arabie, P., 1975. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of mathematical psychology*, 12(3), pp.328–383.

Buckley, P.G. & Osthus, D., 2001. Popularity based random graph models leading to a scale-free degree sequence. *Discrete Mathematics*, 282, pp.53–68.

Buntine, W.L. & Jakulin, A., 2005. Discrete component analysis. In *Subspace, latent structure and feature selection, statistical and optimization, perspectives workshop, SLSFS 2005, bohinj, slovenia, february 23-25, 2005, revised selected papers*. pp. 1–33. Available at: http://dx.doi.org/10.1007/11752790_1.

Burke, R., 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4), pp.331–370.

Callaway, D.S. et al., 2000. Network robustness and fragility: Percolation on random graphs. *Physical review letters*, 85(25), p.5468.

Cappé, O. & Moulines, E., 2009. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pp.593–613.

Caron, F. & Fox, E.B., 2017. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5), pp.1295–1366.

Chen, M., Kuzmin, K. & Szymanski, B.K., 2014. Community detection via maximization of modularity and its variants. *IEEE Transactions on Computational Social Systems*, 1(1), pp.46–65.

Church, K.W. & Gale, W.A., 1995. Poisson mixtures. *Natural Language Engineering*, 1(02), pp.163–190.

Ciglan, M., Laclavík, M. & Nørvåg, K., 2013. On community detection in real-world networks and the importance of degree assortativity. In *Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining.* pp. 1007–1015.

Clauset, A., Shalizi, C.R. & Newman, M., 2009. Power-law distributions in empirical data. *SIAM review*, 51(4), pp.661–703.

Clinchant, S. & Gaussier, E., 2010. Information-based models for ad hoc ir. In *Proceedings of the 33rd international acm sigir conference on research and development in information retrieval.* pp. 234–241.

Clinchant, S. & Gaussier, E., 2008. The bnb distribution for text modeling. In *European conference on information retrieval.* pp. 150–161.

Cohen, R. & Havlin, S., 2003. Scale-free networks are ultrasmall. *Physical review letters*, 90(5), p.058701.

Coscia, M., Giannotti, F. & Pedreschi, D., 2011. A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(5), pp.512–546.

Del Genio, C.I., Gross, T. & Bassler, K.E., 2011. All scale-free networks are sparse. *Physical review letters*, 107(17), p.178701.

Diaconis, P., Ylvisaker, D. & others, 1979. Conjugate priors for exponential families. *The Annals of statistics*, 7(2), pp.269–281.

Dulac, A., Gaussier, E. & Largeron, C., 2017. A study of stochastic mixed membership models for link prediction in social networks. In *Data science and advanced analytics (dsaa), 2017 ieee international conference on.* IEEE, pp. 706–715.

Dunne, J.A., Williams, R.J. & Martinez, N.D., 2002. Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20), pp.12917–12922.

Easley, D. & Kleinberg, J., 2010. Networks, crowds and markets: Reasoning about a highly connected world. In Cambridge University Press, pp. 85–118.

E.J. Newman, M. & Girvan, M., 2004. Finding and evaluating community structure in networks. *Physical review E*, 69(2), p.026113.

ERDdS, P. & R&WI, A., 1959. On random graphs i. *Publ. Math. Debrecen*, 6, pp.290–297.

Erdos, P. & Rényi, A., 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1), pp.17–60.

Erdős, P. & Rényi, A., 1961. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1-2), pp.261–267.

Escobar, M.D. & West, M., 1995. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430), pp.577–588.

Everett, M.G. & Borgatti, S.P., 1994. Regular equivalence: General theory. *Journal of mathematical sociology*, 19(1), pp.29–52.

Fan, X., Cao, L. & Xu, R.Y.D., 2013. Dynamic infinite mixed-membership stochastic blockmodel. *CoRR*.

Feller, W., 1968. *An introduction to probability theory and its applications, vol. I*, Wiley, New York.

Ferguson, T.S., 1973. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pp.209–230.

Ferguson, T.S., Phadia, E.G. & Tiwari, R.C., 1992. Bayesian nonparametric inference. *Lecture Notes-Monograph Series*, 17, pp.127–150.

Flake, G.W. et al., 2002. Self-organization and identification of web communities. *Computer*, 35(3), pp.66–70.

Fong, B., 2013. Causal theories: A categorical perspective on bayesian networks. *arXiv preprint arXiv:1301.6201*.

Fortunato, S., 2010. Community detection in graphs. *Physics reports*, 486(3-5), pp.75–174.

Fortunato, S. & Barthelemy, M., 2007. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1), pp.36–41.

Fortunato, S. & Hric, D., 2016. Community detection in networks: A user guide. *Physics Reports*, 659, pp.1–44.

Foulds, J. et al., 2013. Stochastic collapsed variational bayesian inference for latent dirichlet allocation. In *Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining.* pp. 446–454.

Garfinkel, I., Glei, D. & McLanahan, S.S., 2002. Assortative mating among unmarried parents: Implications for ability to pay child support. *Journal of Population Economics*, 15(3), pp.417–432.

Getoor, L. & Diehl, C.P., 2005. Link mining: A survey. *Acm Sigkdd Explorations Newsletter*, 7(2), pp.3–12.

Geyer, C., 2011. Introduction to markov chain monte carlo. *Handbook of markov chain monte carlo*, 20116022, p.45.

Ghahramani, Z., 1995. Factorial learning and the em algorithm. In *Advances in neural information* *processing systems*. pp. 617–624.

Ghahramani, Z., 2015. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553), pp.452–459.

Girvan, M. & Newman, M.E., 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12), pp.7821–7826.

Goh, K.-I. & Barabási, A.-L., 2008. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4), p.48002.

Goldenberg, A. et al., 2010. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2), pp.129–233.

Gopalan, P.K. & Blei, D.M., 2013. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36), pp.14534–14539.

Görür, D., Jäkel, F. & Rasmussen, C.E., 2006. A choice model with infinitely many latent features. In *Proceedings of the 23rd international conference on machine learning*. pp. 361–368.

Griffiths, T.L. & Ghahramani, Z., 2011. The indian buffet process: An introduction and review. *The Journal of Machine Learning Research*, 12, pp.1185–1224.

Guelzim, N. et al., 2002. Topological and causal structure of the yeast transcriptional regulatory network. *Nature genetics*, 31(1), p.60.

Halmos, P.R., Savage, L.J. & others, 1949. Application of the radon-nikodym theorem to the theory of sufficient statistics. *The Annals of Mathematical Statistics*, 20(2), pp.225–241.

Harary, F. & Norman, R.Z., 1953. *Graph theory as a mathematical model in social science*, University of Michigan, Institute for Social Research Ann Arbor.

Hasan, M.A. & Zaki, M.J., 2011. A survey of link prediction in social networks. In C. C. Aggarwal, ed. *Social network data analytics*. Springer, pp. 243–275. Available at: http://dx.doi.org/10.1007/978-1-4419-8462-3_9.

Hoff, P., 2008. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in neural information processing systems*. pp. 657–664.

Hoffman, M.D. et al., 2013. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1), pp.1303–1347.

Holland, P.W., Laskey, K.B. & Leinhardt, S., 1983. Stochastic blockmodels: First steps. *Social networks*, 5(2), pp.109–137.

Holme, P. & Huss, M., 2005. Role-similarity based functional prediction in networked systems: Application to the yeast proteome. *Journal of the Royal Society Interface*, 2(4), pp.327–333.

Jacobs, A.Z. & Clauset, A., 2014. A unified view of generative models for networks: Models, methods, opportunities, and challenges. *CoRR*, abs/1411.4070. Available at: http://arxiv.org/

abs/1411.4070.

Jäger, M. et al., 2009. Analysis of single-molecule fluorescence spectroscopic data with a markov-modulated poisson process. *ChemPhysChem*, 10(14), pp.2486–2495.

Jeong, H. et al., 2000. The large-scale organization of metabolic networks. *Nature*, 407(6804), p.651.

Jordan, M.I., 2010. Bayesian nonparametric learning: Expressive priors for intelligent systems. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, 11, pp.167–185.

Kalapala, V., Sanwalani, V. & Moore, C., 2003. The structure of the united states road network. *Preprint, University of New Mexico.*

Kallenberg, O., 2006. *Probabilistic symmetries and invariance principles*, Springer Science & Business Media.

Karinthy, F., 1929. Chains. *Everything is different, Budapest.*

Karrer, B. & Newman, M.E., 2011. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1), p.016107.

Kauffman, A.S., 1992. *The origins of order: Self-organization and selection in evolution*, World Scientific.

Kemp, C. et al., 2006. Learning systems of concepts with an infinite relational model. In *AAAI.* p. 5.

Khan, B.S. & Niazi, M.A., 2017. Network community detection: A review and visual survey. *CoRR.*

Kim, D.I. et al., 2013. Efficient online inference for bayesian nonparametric relational models. In *Advances in neural information processing systems.* pp. 962–970.

Kim, K. & Altmann, J., 2017. Effect of homophily on network formation. *Communications in Nonlinear Science and Numerical Simulation*, 44, pp.482–494.

Kim, M. & Leskovec, J., 2012. Multiplicative attribute graph model of real-world networks. *Internet Mathematics*, 8(1-2), pp.113–160.

Kingman, J.F. & others, 1978. Uses of exchangeability. *The Annals of Probability*, 6(2), pp.183–197.

Kleinberg, J.M., 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5), pp.604–632.

Klimt, B. & Yang, Y., 2004. *The enron corpus: A new dataset for email classification research*, Springer.

Koutsourelakis, P.-S. & Eliassi-Rad, T., 2008. Finding mixed-memberships in social networks. In *AAAI spring symposium: Social information processing.* pp. 48–53.

Krackhardt, D., 1999. The ties that torture: Simmelian tie analysis in organizations. *Research in the Sociology of Organizations*, 16(1), pp.183–210.

Kunegis, J., 2013. Konect: The koblenz network collection. *Proceedings of the 22nd International Conference on World Wide Web*, pp.1343–1350.

Lafferty, K.D., Dobson, A.P. & Kuris, A.M., 2006. Parasites dominate food web links. *Proceedings of the National Academy of Sciences*, 103(30), pp.11211–11216.

La Fond, T. & Neville, J., 2010. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th international conference on world wide web*. WWW '10. ACM, pp. 601–610.

Lahiri, M. & Berger-Wolf, T.Y., 2007. Structure prediction in temporal networks using frequent subgraphs. In *Computational intelligence and data mining, 2007. CIDM 2007. IEEE symposium on*. pp. 35–42.

Lambiotte, R., Tabourier, L. & Delvenne, J.-C., 2013. Burstiness and spreading on temporal networks. *The European Physical Journal B*, 86(7), p.320.

Largeron, C. et al., 2017. DANCer: Dynamic attributed networks with community structure generation. *Knowl Inf Syst*, pp.1–43.

Largeron, C. et al., 2015. Generating attributed networks with communities. *PLoS ONE*, 10(4), p.e0122777. Available at: http://dx.doi.org/10.1371%2Fjournal.pone.0122777.

Latouche, P., Birmele, E. & Ambroise, C., 2012. Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1), pp.93–115.

Lawrence, N.D. & Moore, A.J., 2007. Hierarchical gaussian process latent variable models. In *Proceedings of the 24th international conference on machine learning*. ACM, pp. 481–488.

Lazarsfeld, P.F., Merton, R.K. & others, 1954. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18(1), pp.18–66.

Le, C.M., Levina, E. & Vershynin, R., 2015. Sparse random graphs: Regularization and concentration of the laplacian. *arXiv preprint arXiv:1502.03049*.

Leacock, C. & Chodorow, M., 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2), pp.265–283.

Lee, J. et al., 2015. Preferential attachment in graphs with affinities. In *Proceedings of the eighteenth international conference on artificial intelligence and statistics, AISTATS 2015, san diego, california, usa, may 9-12, 2015*. Available at: http://jmlr.org/proceedings/papers/v38/lee15b.html.

Leicht, E.A., Holme, P. & Newman, M.E., 2006. Vertex similarity in networks. *Physical Review E*, 73(2), p.026120.

Leskovec, J. et al., 2008. Microscopic evolution of social networks. In *Proceedings of the 14th* *acm sigkdd international conference on knowledge discovery and data mining*. KDD '08. pp. 462–470.

Leskovec, J., Kleinberg, J. & Faloutsos, C., 2007. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), p.2.

Leskovec, J. & Krevl, A., 2014. SNAP Datasets: Stanford large network dataset collection. Available at: http://snap.stanford.edu/data.

Ley, M., 2002. The dblp computer science bibliography: Evolution, research issues, perspectives. In *International symposium on string processing and information retrieval*. Springer, pp. 1–10.

Li, W.-J., Yeung, D.-Y. & Zhang, Z., 2011. Generalized latent factor models for social network analysis. In *Proceedings of the 22nd international joint conference on artificial intelligence (ijcai), barcelona, spain*. p. 1705.

Li, Y., Liu, B. & Sarawagi, S. eds., 2008. *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, las vegas, nevada, usa, august 24-27, 2008*,

Liben-Nowell, D. & Kleinberg, J., 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7), pp.1019–1031.

Liljeros, F. et al., 2001. The web of human sexual contacts. *Nature*, 411(6840), p.907.

Lorrain, F. & White, H.C., 1971. Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1(1), pp.49–80.

Lü, L. & Zhou, T., 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6), pp.1150–1170.

Mackey, L.W., Weiss, D.J. & Jordan, M.I., 2010. Mixed membership matrix factorization. In *ICML*. pp. 711–718.

Mariadassou, M. et al., 2010. Uncovering latent structure in valued graphs: A variational approach. *The Annals of Applied Statistics*, 4(2), pp.715–742.

Maslov, S. & Sneppen, K., 2002. Specificity and stability in topology of protein networks. *Science*, 296(5569), pp.910–913.

McPherson, M., Smith-Lovin, L. & Cook, J.M., 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1), pp.415–444.

Meeds, E. et al., 2006. Modeling dyadic data with binary latent factors. In *Advances in neural information processing systems*. pp. 977–984.

Menon, A.K. & Elkan, C., 2010. A log-linear model with latent features for dyadic prediction. In *Data mining (icdm), 2010 ieee 10th international conference on*. IEEE, pp. 364–373.

Menon, A.K. & Elkan, C., 2011. Link prediction via matrix factorization. In *Joint european* *conference on machine learning and knowledge discovery in databases*. pp. 437–452.

Merton, R.K., 1968. The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810), pp.56–63.

Michalski, R.S., Carbonell, J.G. & Mitchell, T.M., 2013. *Machine learning: An artificial intelligence approach*, Springer Science & Business Media.

Miller, K., Jordan, M.I. & Griffiths, T.L., 2009. Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*. pp. 1276–1284.

Mørup, M., Schmidt, M.N. & Hansen, L.K., 2011. Infinite multiple membership relational modeling for complex networks. In *Machine learning for signal processing (mlsp), 2011 ieee international workshop on*. pp. 1–6.

Neal, R.M., 1993. Probabilistic inference using markov chain monte carlo methods.

Nelson, R., 2013. *Probability, stochastic processes, and queueing theory: The mathematics of computer performance modeling*, Springer Science & Business Media.

Newman, M., 2010. *Networks: An introduction*, New York, NY, USA: Oxford University Press, Inc.

Newman, M., 2001. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical review E*, 64(1), p.016132.

Newman, M.E., 2004. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6), p.066133.

Newman, M.E., 2006. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3), p.036104.

Newman, M.E., 2003. Mixing patterns in networks. *Physical Review E*, 67(2), p.026126.

Newman, M.E., 2005. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5), pp.323–351.

Newman, M.E., 2003. The structure and function of complex networks. *SIAM Review*, 45(2), pp.167–256.

Newman, M.E., 2001. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2), pp.404–409.

Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401), pp.9–27.

Onnela, J.-P. et al., 2012. Taxonomies of networks from community structure. *Physical Review E*, 86(3), p.036104.

Orbanz, P., 2009. *Functional conjugacy in parametric bayesian models*, Technical report, Univ. Cambridge.

Orbanz, P. & Roy, D.M., 2015. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2), pp.437–461.

Page, L. et al., 1999. *The pagerank citation ranking: Bringing order to the web*, Stanford InfoLab.

Palla, K., Ghahramani, Z. & Knowles, D.A., 2012. An infinite latent attribute model for network data. In *Proceedings of the 29th international conference on machine learning*. ACM, pp. 1607–1614.

Palla, K., Knowles, D.A. & Ghahramani, Z., 2012. An infinite latent attribute model for network data. In *Proceedings of the 29th international conference on machine learning, ICML 2012, edinburgh, scotland, uk, june 26 - july 1, 2012*.

Pei, S. & Makse, H.A., 2013. Spreading dynamics in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(12), p.P12002.

Peixoto, T.P., 2017. Nonparametric bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95(1), p.012317.

Peixoto, T.P., 2018. Nonparametric weighted stochastic block models. *Physical Review E*, 97(1), p.012306.

Peixoto, T.P., 2014. The graph-tool python library. *figshare*. Available at: http://figshare.com/articles/graph_tool/1164194 [Accessed September 10, 2014].

Preusse, J. et al., 2013. Structural dynamics of knowledge networks. *ICWSM*, 17, p.18.

Rasmussen, C.E., 2004. Gaussian processes in machine learning. In *Advanced lectures on machine learning*. Springer, pp. 63–71.

Robbins, H. & Monro, S., 1951. A stochastic approximation method. *The annals of mathematical statistics*, pp.400–407.

Rossi, R.A. & Ahmed, N.K., 2015. Role discovery in networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(4), pp.1112–1131.

Rosvall, M. et al., 2017. Different approaches to community detection. *arXiv preprint arXiv:1712.06468*.

Schaub, M.T. et al., 2012. Markov dynamics as a zooming lens for multiscale community detection: Non clique-like communities and the field-of-view limit. *PloS one*, 7(2), p.e32210.

Schwartz, M. & Wood, D.C., 1992. Discovering shared interests among people using graph analysis of global electronic mail traffic. *Communications of the ACM*, 36, pp.78–89.

Sola Pool, I. de & Kochen, M., 1978. Contacts and influence. *Social networks*, 1(1), pp.5–51.

Sporns, O., 2002. Network analysis, complexity, and brain function. *Complexity*, 8(1), pp.56–60.

Stelling, J. et al., 2002. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912), p.190.

Sun, D. et al., 2009. Information filtering based on transferring similarity. *Physical Review E*, 80(1), p.017101.

Sutton, C., McCallum, A. & others, 2012. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4), pp.267–373.

Teh, Y.W., 2010. Dirichlet Process. In *Encyclopedia of machine learning.* Springer, pp. 280–287.

Teh, Y.W. & Jordan, M.I., 2010. Hierarchical bayesian nonparametric models with applications. *Bayesian nonparametrics*, 1, pp.158–207.

Teh, Y.W. et al., 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476), pp.1566–1581.

Teh, Y.W., Newman, D. & Welling, M., 2007. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in neural information processing systems.* pp. 1353–1360.

Thompson, R.M. & Townsend, C.R., 2000. Is resolution the solution?: The effect of taxonomic resolution on the calculated properties of three stream food webs. *Freshwater Biology*, 44(3), pp.413–422.

Travers, J. & Milgram, S., 1967. The small world problem. *Phychology Today*, 1(1), pp.61–67.

Tutorial, N., 2012. Tutorials on bayesian nonparametrics. *Journal of Mathematical Psychology*, 56, pp.1–12.

Veitch, V. & Roy, D.M., 2015. The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099.*

Wang, P. et al., 2015. Link prediction in social networks: The state-of-the-art. *Science China Information Sciences*, 58(1), pp.1–38.

Wasserman, S. & Faust, K., 1994. *Social network analysis: Methods and applications*, Cambridge university press.

Watts, D.J., 2004. *Six degrees: The science of a connected age*, WW Norton & Company.

Watts, D.J. & Strogatz, S.H., 1998. Collective dynamics of small-world networks. *Nature*, 393(6684), pp.440–442.

White, D. & P. Reitz, K., 1983. Graph and semigroup homomorphisms on networks of relations., 5, pp.193–234.

White, H.C., Boorman, S.A. & Breiger, R.L., 1976. Social structure from multiple networks. I. Blockmodels of roles and positions. *American journal of sociology*, 81(4), pp.730–780.

Xu, Z. et al., 2006. Learning infinite hidden relational models. *Uncertainity in Artificial Intelligence (UAI2006).*

Yang, J. & Leskovec, J., 2015. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1), pp.181–213.

Zeng, S., 2016. Link prediction based on local information considering preferential attachment. *Physica A: Statistical Mechanics and its Applications*, 443, pp.537–542.

Zhang, H. et al., 2016. Modeling the homophily effect between links and communities for overlapping community detection. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence, IJCAI*. pp. 3938–3944.

Zhang, Z.-K. et al., 2016. Dynamics of information diffusion and its applications on complex networks. *Physics Reports*, 651, pp.1–34.

Zhou, M., 2015. Infinite edge partition models for overlapping community detection and link prediction. In *Artificial intelligence and statistics*. pp. 1135–1143.

Zhou, M. & Carin, L., 2012. Augment-and-conquer negative binomial processes. In *Advances in neural information processing systems*. pp. 2546–2554.

Zhou, M. & Carin, L., 2015. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2), pp.307–320.

Zhou, M. et al., 2012. Beta-negative binomial process and poisson factor analysis. *Journal of Machine Learning Research*.

Zipf, G.K., 2016. *Human behavior and the principle of least effort: An introduction to human ecology*, Ravenio Books.