Etude des modèles mixed-membership pour la prédiction de liens dans les réseaux sociaux

!!! No author defined !!!!

!!! No address defined !!!!

RÉSUMÉ. *Nous évaluons dans ce papier si la classe de modéle, dit mixed-membership, est adapté pour la prédiction de liens dans réseaux sociaux en étudiant leurs comportements vis-à-vis de l'homophilie et de l'attachement préférentiel. Pour étudier ces propriétés, nous introduisons les définitions de ces phénomènes, après quoi nous étudions comment les modèles sont reliés à ces définitions. Notre étude théorique révèle que les modéles mixed-membership satisfont l'homophily avec la similarité naturelle qui les sous-tend. Pour l'attachement préférentiel la situation est plus contrasté: ces modèles ne satisfont pas l'attachement préférentiel global. En revanche, ils satisfont un attachement préférentiel local si l'appartenance au classes latentes est strict ou léger, et dans ce dernier cas si la distribution sur les classes latentes est bursty ou pas. Nous illustrons ces éléments sur des réseaux réel et synthétique grâce au propriété générative de ces modéles Bayésien.*

ABSTRACT. *We assess here whether standard stochastic mixed membership models are adapted for link prediction in social networks by studying how they handle homophily and preferential attachment. To study these properties, we first introduce formal definitions of these phenomena; we then study how stochastic mixed membership models relate to these definitions. Our theoretical analysis reveals that standard stochastic mixed membership models comply with homophily with the similarity that underlies them. For preferential attachment, the situation is more contrasted: if these models do not comply with global preferential attachment, their compliance to local preferential attachment depends on whether the memberships to latent factors are hard or soft, and in the latter case on whether the underlying latent factor distribution is bursty or not. We illustrate these elements on synthetic and real networks by using the generative properties of Bayesian model.*

MOTS-CLÉS : *Modèle Bayesian Hérarchique, Graphe aléatoire, Apprentissage Automatique, Réseaux Sociaux*

KEYWORDS: *Hierachical baysian Models, Random Graph, Machine Learning, Social Networks*

DOI:10.3166/RIA.28.1-17 © 2014 Lavoisier

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). More generally, these models fall in the category of mixed-membership models who establish a common theoretical framework that encompass a wide range of models (such as admixture and topic model) that are able to learn complex pattern from structured data (Airoldi *et al.*, 2014).

Nevertheless, although drawn from a wide range of domains, real world social networks exhibit general properties and one can wonder if these models are able to capture these properties. In this work, we focus on two important properties, namely *homophily* and *preferential attachment* (Newman, 2010 ; Barabási, 2003) and assess to which extent link prediction models, as the ones mentioned above, comply with them. 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 vertex is more likely to create connections with vertices having a high degree. 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 remainder of the paper is organized as follows: in the next section (Section 2), we present the related work. In Section 3, we describe the main stochastic mixed membership models used for link prediction in social networks, relying on their non-parametric version that generalizes the parametric one. In Sections 4 and 5, we introduce formal definitions of homophily and preferential attachment and study how stochastic mixed membership models relate to them. In Section 6, we illustrate our theoretical development on two synthetic networks and two real networks, prior to restate our conclusions in Section 7.

2. Related Work

Recently, the class of stochastic mixed membership models have been successfully used for link prediction and structure discovery in social networks. 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 used it for discovering overlapping communities in large size networks having millions of nodes since a-MMSB scales well using stochastic variational inference. They constrained the weight matrix to have weights in the diagonal and a fixed small value elsewhere. 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 way, to handle non-negative weights in (Mørup *et al.*, 2011) and with a more subtle latent feature structure in (Palla *et al.*, 2012). Nevertheless, the characterization of these models with regards to the properties of the networks remains to be explored (Jacobs, Clauset, 2014).

In this article, we focus on two properties: *homophily* and *preferential attachment* (Newman, 2010 ; Barabási, 2003). The interest of these properties has been widely emphasized in previous works notably for modeling and generating networks reflecting properties of real networks, as in

the Barabàsi-Albert model (Albert, Barabási, 2002) or Buckley and Osthus model (Buckley, Osthus, 2001) that integrate a preferential attachment mechanism, in the Multiplicative Attribute Graph (MAG) model (Kim, Leskovec, 2012) that considers node affinities, or in the Dancer model that takes into consideration both properties (Largeron *et al.*, 2017). Homophily and preferential attachment have also been exploited for improving the results obtained in classical tasks such as community detection (Ciglan *et al.*, 2013 ; Zhang *et al.*, 2016) or link prediction (Aiello *et al.*, 2012 ; Zeng, 2016). That said, few theoretical works have been done to study to what extent models comply with these properties.

Concerning preferential attachment, Orbanz and Roy (Orbanz, Roy, 2015) 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. Consequently, Lee *et al.* (Lee *et al.*, 2015) proposed a random network model in order to capture the power law typical of the degree distribution in social networks. 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).

Following these previous studies, 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 these phenomena and then study how the models behave with respect to these definitions.

3. Backround Models

Stochastic mixed membership models are generative models that rely on latent factors (also called latent *classes* or *features*) that represent hidden properties of the nodes of the graph $G = (V, E)$ associated to the social network (each node of the graph represents an individual in the social network); in the remainder, we denote by N the number of nodes in this graph ($N = |V|$). 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 $\mathbf{f}_i$, of finite dimension K in standard versions of the models, and of infinite dimension in non-parametric versions. The collection of all vectors $\mathbf{f}_i$ ($1 \leq i \leq N$) constitutes the factor matrix $\mathbf{F}$. Furthermore, a weight matrix is used to encode correlations between the latent factors.

Stochastic mixed membership models differ on the way the vectors $\mathbf{f}_i$ ($1 \leq i \leq N$) and the matrix 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 Hierachical

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.

Dirichlet Process (HDP) for `IMMSB`. All our results are nevertheless also valid for the finite versions of these models.

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 for a finite number of nodes, only a finite number of dimensions is actually active). The following steps summarize this process:

1. Generate a feature matrix $\mathbf{F}_{N\times\infty}$ representing the feature vector of each node: $\mathbf{F} \sim \text{IBP}(\alpha)$
2. Generate a weight matrix for each latent feature:
$\phi_{mn} \sim N(0, \sigma_w)$, $m, n \in \mathbb{N}^{+*}$
3. Generate or not a link between any node i and any node j according to:

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

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 $\mathbf{Y}$ the $N \times N$ matrix with elements y_{ij}. Finally, $\mathbf{f}_i$ denotes the row feature vector corresponding to the i^{th} row of $\mathbf{F}$.

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 $\mathbf{Y}$ and are symmetric and only their upper (or lower) diagonal parts are generated. Lastly, both $\mathbf{F}$ and are infinite matrices. In practice however, one always deal with a finite number of latent features. A graphical representation of this model is given in Figure 3 (left).

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:

1. Generate the class membership distributions $\mathbf{F}_{N\times\infty}$:

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

where GEM denotes the Stick Breaking Process distribution over the set of natural numbers and DP a Dirichlet Process (Teh *et al.*, 2006).

2. Generate a weight matrix for each latent class from i.i.d Beta distribution:

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

3. 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:

$$z_{i\to j} \sim \text{Cat}(\mathbf{f}_i)$$
$$z_{i\leftarrow j} \sim \text{Cat}(\mathbf{f}_j)$$
$$y_{ij} \sim \text{Bern}(\phi_{z_{i\to j} z_{i\leftarrow j}})$$

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 $\mathbf{Y}$ and are symmetric and only their upper (or lower) diagonal parts are generated; as before again, both $\mathbf{F}$ and are infinite matrices. A graphical representation of this model is given in Figure 3 (right).

Standard Gibbs sampling and Metropolis-Hastings algorithms can be used for inference in this model. We do not detail them here and refer the interested reader to (Miller *et al.*, 2009) and (Griffiths, Ghahramani, 2011).

In the typical use of the above models for link prediction, some observations (*i.e.* an existing network, observed till a certain time) are available and are used to estimate $\mathbf{F}$ and , from which new links are predicted. In the remainder, we denote by $\widehat{\mathbf{F}}$ and $\widehat{\ }$ the estimates of $\mathbf{F}$ and , that can be obtained through standard (collapsed) Gibbs sampling and Metropolis-Hastings algorithms. We do not detail them here and refer the interested reader to (Miller *et al.*, 2009 ; Griffiths, Ghahramani, 2011 ; Teh *et al.*, 2006 ; Fan *et al.*, 2013). We furthermore denote by $\mathcal{M}_e$ the version of both `ILFM` and `IMMSB` models in which $\mathbf{F}$ and are assumed known and fixed to $\widehat{\mathbf{F}}$ and $\widehat{\ }$. We now investigate whether, from the learned parameters $\widehat{\mathbf{F}}$ and $\widehat{\ }$, the new links generated produce a network that comply with homophily and preferential attachment.

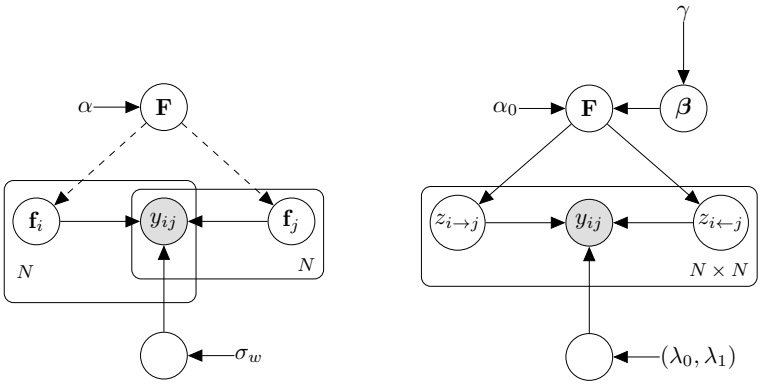

Figure 1. The two graphical representations of (left) the latent feature model and (right) the latent class model. The difference between the two graphical structures of models lies in the way representations are associated to nodes: a fixed representation is used in the case of the latent feature model, whereas the representation in the latent class model varies according to the link considered.

4. Homophily: *"Birds of a feather flock together"*

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:

D FINITION 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 $\widehat{\mathbf{F}}$ aims at capturing some latent properties of the nodes, whereas the estimated matrix $\widehat{\ }$ captures the correlations between these latent properties. One can thus define, on their basis, a "natural" similarity between nodes as follows:

$$s_n(i,j) = \widehat{\mathbf{f}}_i \widehat{\ } \widehat{\mathbf{f}}_j^\top$$

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:

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

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

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

as $\hat{\phi}_{kk} = 0$. Let us now consider $\widehat{\mathbf{f}}_i = \widehat{\mathbf{f}}_j = (0,1,0)$ and $\widehat{\mathbf{f}}_{i'} = (0.5,0,0.5)$ and $\widehat{\mathbf{f}}_{j'} = (0,1,0)$. 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 $\widehat{\mathbf{f}}_{i'} = (0.5,0,0.5)$ by $\widehat{\mathbf{f}}_{i'} = (1,0,1)$, 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 Ω (and hence its estimate $\widehat{\Omega}$); 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.

5. Preferential attachment: *"The rich get richer"*

Preferential attachment, sometimes referred to as the *rich get richer* rule, is a mechanism according to which each node is connected to an existing node with a probability that increases with the number of links of the chosen node [2]. However, as noted in Leskovec *et al.*, usually, in social networks, entities do not have a global knowledge of the network. The preferential attachment model is thus more likely to be local, and to be specific to communities (Leskovec *et al.*, 2008).

Preferential attachment relates to a general phenomenon known as *burstiness* [3] which describes the fact that some events appear in bursts, *i.e.* once they appear, they are more likely to appear again. Burstiness has been studied in different fields, in particular in computational linguistics and information retrieval to characterize word occurrences (Church, Gale, 1995). In these domains, simple definitions of burstiness have been proposed (Clinchant, Gaussier, 2008 ; 2010), for both discrete and continuous probability distributions. They directly capture the fact that a probability distribution is bursty if the probability of generating a new occurrence of an event increases with the number of occurrences of this event. We adapt here these definitions for preferential attachment in social networks.

In the context of social networks, the notion of preferential attachment amounts to the fact that the more links a node has (*i.e.* the higher its degree), the more likely it will be linked to new nodes. As mentioned before, this phenomenon however appears at different levels: globally for the whole network, and locally within classes. For global preferential attachment, the degree of a node is a known integer; for local preferential attachment, in most models, the exact degree is not known, and one has to rely on an estimate of it, as done in the following definition:

Definition 2 (Preferential attachment). — *Let i be a node in a social network and let d_i denote its degree.*

(1) Global Preferential Attachment*: we say that a probabilistic link prediction model $\mathcal{M}_e$ satisfies the global preferential attachment iff, for any node i, $P(d_i \geq n+1 \mid d_i \geq n, \mathcal{M}_e)$ increases with $n \in \{0, ..., N-1\}$;*

(2) Local Preferential Attachment*: we say that a probabilistic link prediction model $\mathcal{M}_e$ satisfies the local preferential attachment iff, for any node i, denoting $d_{i,k}$ the degree of node i in class k, $\forall \epsilon \in [0,1]$, $P(d_{i,k} \geq x+\epsilon \mid d_{i,k} \geq x, \mathcal{M}_e)$ increases with $x \in [0, N[$. Furthermore, $d_{i,k}$ is defined as the sum of the expectations, over all nodes in the network, of forming a link through latent factor k.*

As one can note, these definitions directly translate the fact that "the more connections a node has, the more likely it is to be connected to new nodes". The only difference between the local

2. This property is well captured by a power law distribution, hence the claim often made that preferential attachment translates as a power law for the node degrees distribution.

3. A.L. Barabási, for example, uses the term *preferential attachment* in (Barabási, Albert, 1999), and *burstiness* in (Barabási, 2011).

and global cases is that the degree is usually unknown in the local case, and is here estimated through its expectation.

For global preferential attachment, the degree d_i directly corresponds to the number of outgoing links of node i. Exploiting the fact that the observations are independent given $\widehat{\mathbf{F}}$ and $\widehat{\ }$, one has:

$$\begin{aligned} &P(d_i \geq n+1 \mid d_i \geq n, \mathcal{M}_e) \\ &\quad = 1 - \prod_{j \notin \mathcal{V}(i)} p(y_{ij} = 0 \mid d_i \geq n, \mathcal{M}_e) \\ &\quad = 1 - \prod_{j \notin \mathcal{V}(i)} (1 - p(y_{ij} = 1 \mid d_i \geq n, \mathcal{M}_e)) \end{aligned}$$

where $\mathcal{V}(i)$ denotes the set of nodes connected to node i. Let $c = \min_{j \in V}(1 - p(y_{ij} = 1 \mid d_i \geq n, \mathcal{M}_e))$. One has:

$$0 \leq P(d_i \geq n+1 \mid d_i \geq n, \mathcal{M}_e) \leq (1 - c^{N-1-n})$$

As $c < 1$, $(1 - c^{N-1-n})$ decreases with n and is 0 when $n = N - 1$. We thus have the following property.

Proposition 3. — *Both `ILFM` and `IMMSB` do not satisfy global preferential attachment.*

For local preferential attachment, the situation is slightly more complex:

Proposition 4. — *`IMMSB` satisfies local preferential attachment whereas `ILFM` does not.*

Proof (sketch) Let $y_{ij,k}$ be the binary random variable that is 1 if nodes i and j are linked through the latent factor k and 0 otherwise. Then, $d_{i,k} = \sum_{j \in V} P(y_{ij,k} = 1 | \mathcal{M}_e)$. For `IMMSB`, this leads to $d_{i,k} = \sum_{j \in V} \hat{f}_{ik} \hat{\Phi}_{kk} \hat{f}_{jk} = \hat{f}_{ik} \sum_{j \in V} \hat{\Phi}_{kk} \hat{f}_{jk}$. The positive reinforcement effect of the Dirichlet Process (Teh *et al.*, 2006) at the basis of `IMMSB` corresponds to a burstiness phenomenon and directly translates, for any i and any k, as: $P(\hat{f}_{ik} \geq x' + \epsilon' \mid \hat{f}_{ik} \geq x', \mathcal{M}_e)$ increases with x' (for all ϵ' and x' chosen according to the domain of definition of $\hat{f}_{ik}$). Setting $x = x'(\sum_{j \in V} \hat{\Phi}_{kk} \hat{f}_{jk})$ and $\epsilon = \epsilon'(\sum_{j \in V} \hat{\Phi}_{kk} \hat{f}_{jk})$ and exploiting the fact that $\sum_{j \in V} \hat{\Phi}_{kk} \hat{f}_{jk}$ is positive and independent of i leads to: $P(d_{i,k} \geq x + \epsilon \mid d_{i,k} \geq x, \mathcal{M}_e)$ increases with x, which proves that `IMMSB` satisfies the local preferential attachment effect. For `ILFM`, let $C_{i,k} = |\{j \in V, \hat{f}_{jk} = \hat{f}_{ik} = 1\}|$. As the factor matrix is binary, one has:

$$d_{i,k} = \sum_{j \in V} \sigma(\hat{f}_{ik} \hat{\Phi}_{kk} \hat{f}_{jk}) = C_{i,k}(\sigma(\hat{\Phi}_{kk}) - 0.5) + \frac{N}{2}$$

As $\hat{f}_{ik}$ is binary, there is no positive reinforcement effect: $C_{i,k}$ does not increase if $\hat{f}_{ik} = 1$, thus `ILFM` does not satisfy local preferential attachment. □

The above propositions show that both models are deficient in the sense that they do not guarantee that the networks they generate will comply with the global (and local in case of `ILFM`) preferential attachment phenomena, which are inherent properties of the probability distributions underlying the models. This does not mean however that `ILFM` and `IMMSB` are not able to well model social networks during the learning phase, even according to the underlying degree distribution. Indeed, the Gibbs updates for both models will assign higher weight to nodes and factors that have been encountered more often during the learning phase. Provided there

is enough training data, both models will likely reproduce the degree distributions observed in the training data. We will observe that in the following section, devoted to the illustration of the properties we have established.

6. Illustration

To illustrate our theoretical results, we evaluate the predictive performance and the ability of the models to capture homophily and preferential attachment on artificial and real networks. For homophily, we simply compare the distributions of the natural and latent similarities on linked and non-linked pairs of nodes. For global preferential attachment, we use plots of the degree distribution and its corresponding best fitting line in log-log scale. In addition, we use the measure developed in (Clauset *et al.*, 2009) for assessing whether empirical data behaves according to a power law (as mentioned before, power laws are the standard bursty distributions in social networks (Barabási, Albert, 1999)). This framework combines maximum-likelihood methods with goodness-of-fit tests based on the Kolmogorov-Smirnov statistics to compute a p-value. If the obtained p-value is large (close to 1), then the data is likely to be distributed according to a power law and the associated network displays preferential attachment; on the other hand, if it is small, the data is likely not distributed according to a power law and the associated network does not display preferential attachment.

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 $y_{ij} = 1, f_{ik} = f_{jk} = 1$ 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.

6.1. Datasets and model parameters

To illustrate the above developments, we consider two artificial and two real networks, the characteristics of which are summarized in Table 1.

Tableau 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

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

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. 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 6.1. The adjacency matrices enable 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 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 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 6.1 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 Table 2: the p-value equals 1 for Network1 and Blogs, 0 for Network2 and 0.4 for Manufacturing.

For each dataset, we estimate the model parameters through 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 hyperparameter σ_w was fixed to 1 and the IBP hyperparameter α to 0.5 in order to have comparable number of classes with `IMMSB`. 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.

6.2. Homophily

Figure 6.2 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

5. moreno.ss.uci.edu/data.html#blogs
6. www.ii.pwr.edu.pl/ michalski/index.php?content=datasets#manufacturing

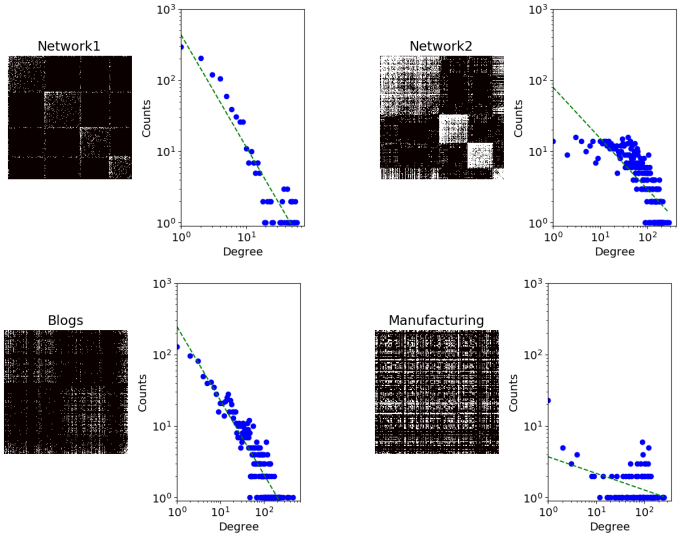

Figure 2. Adjacency matrices (left) and global degree distributions (right) for the four training datasets. In the adjacency matrices, a white dot corresponds to a 1 and a black dot to a 0.

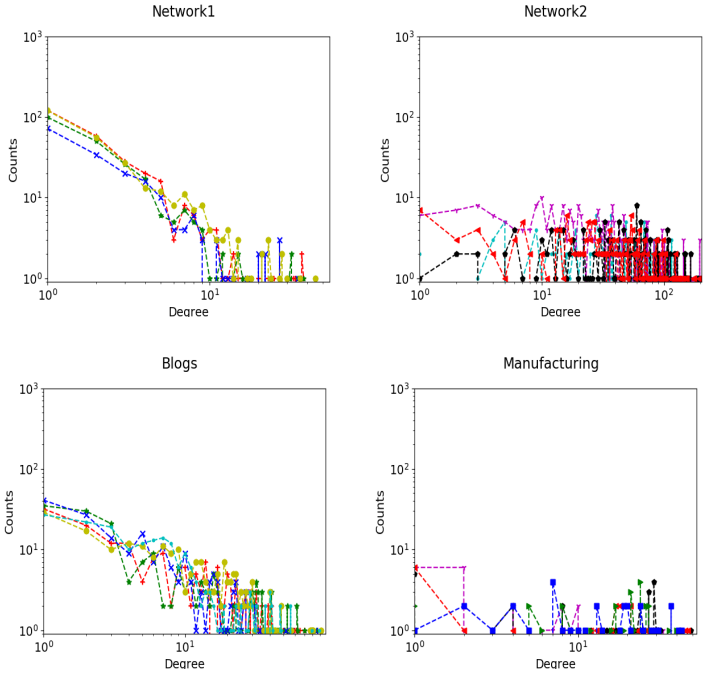

Figure 3. Local degree distributions for the four training datasets. For Network1 and Network2 the classes come from ground-truth. For Blogs and Manufacturing, classes are obtained by Louvain algorithm.

are in line with the theoretical results presented in Section 4 that state that both ILFM and IMMSB are homophilic for the natural similarity but are not homophilic for the latent similarity.

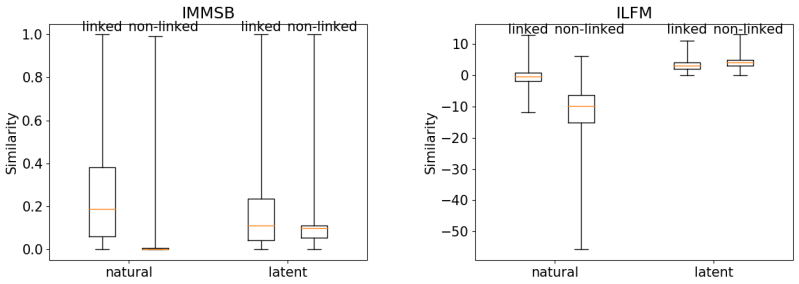

Figure 4. Natural and latent similarities aggregated over all datasets and computed on linked and non-linked pairs of nodes for IMMSB (top) and ILFM (bottom).

6.3. Preferential attachment

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

Figure 6.3 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 2.2.

Lastly, Figure 6.3 compares 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 ($AUC_{\texttt{IMMSB}} - AUC_{\texttt{ILFM}}$) 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 MAX 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 size of this network, making the prediction less challenging.

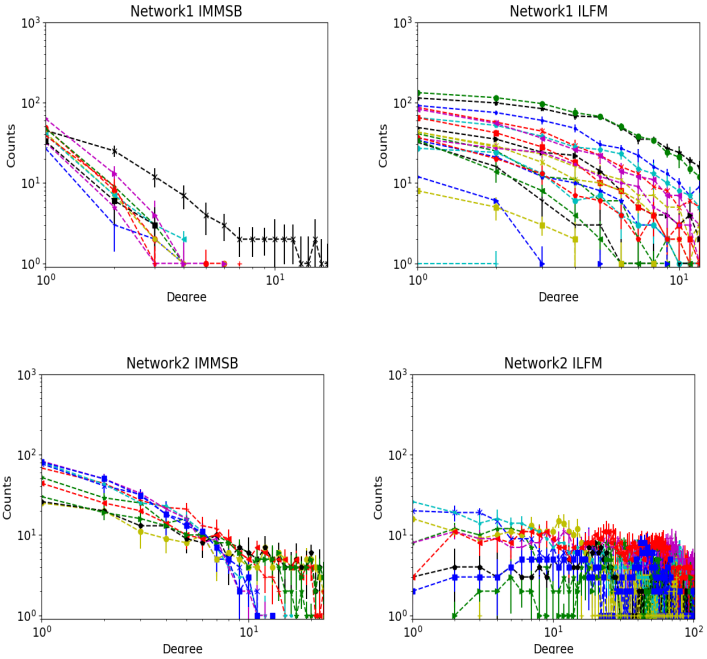

Figure 5. Local degree distributions for Network1 (top row) and Network2 (bottom row) generated with fitted models `IMMSB` *(first column) and* `ILFM` *(second column).*

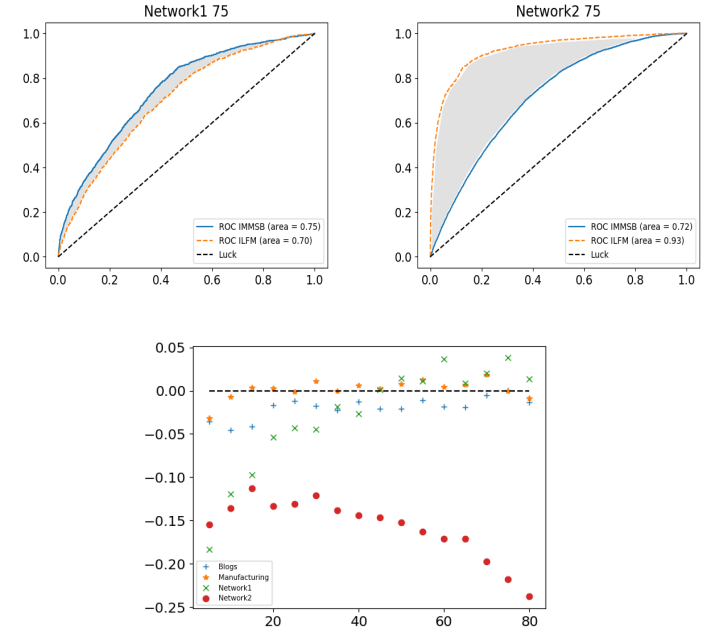

Figure 6. Top: AUC-ROC curves for Network1 (left) and Network2 (right) with 75 percent of data used for learning that compares the performance of models. Bottom: Relative performance of `IMMSB` *and* `ILFM` *according to the percentage of data used for testing, the rest being used for learning.*

Tableau 2. Preferential attachment measures for training datasets and networks generated with fitted models.

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

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.

7. Conclusion

We have studied whether stochastic mixed membership models, such as `ILFM` and `IMMSB` can generate new links while satisfying properties frequently verified in real social networks, namely homophily and preferential attachment. To do so, we have introduced formal definitions of these properties and 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 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 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 straightforward extension is to examine the relation between the local preferential attachment and the dynamic of the latent classes. For instance, a fundamental result is the Aldous-Hoover theorem, which implies that exchangeable random graphs cannot be sparse (Orbanz, Roy, 2015). It seems that the sparsity is related in some way to the preferential attachment in a network. Thus, the following question arises: would it be realistic to assume the exchangeability hypothesis for the local

case but not for the global case, and how this fact impacts the burstiness of the global degree distribution and the sparsity of the graph.

We believe that answering to those questions open a way to develop and design Bayesian models able to better capture the fundamental properties of real social networks.

Remerciements

Acknowledgments
Ce travaille a été partiellement suporté par la Région Rhône-Alpes.

Bibliographie

Aiello L. M., Barrat A., Schifanella R., Cattuto C., Markines B., Menczer F. (2012). Friendship prediction and homophily in social media. *ACM Trans. Web*, vol. 6, n° 2, p. 1–33.

Airoldi E. M., Blei D., Erosheva E. A., Fienberg S. E. (2014). *Handbook of mixed membership models and their applications*. CRC Press.

Airoldi E. M., Blei D. M., Fienberg S. E., Xing E. P. (2009). Mixed membership stochastic blockmodels. In *Advances in neural information processing systems*, p. 33–40.

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

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*, vol. 286, n° 5439, p. 509–512.

Blondel V. D., Guillaume J., Lambiotte R., Lefebvre E. (2008). Fast unfolding of community in large networks. *Journal of statistical mechanics: theory and experiment*, vol. 10, p. P10008.

Buckley P. G., Osthus D. (2001). Popularity based random graph models leading to a scale-free degree sequence. *Discrete Mathematics*, vol. 282, p. 53-68.

Church K. W., Gale W. A. (1995). Poisson mixtures. *Natural Language Engineering*, vol. 1, n° 02, p. 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*, p. 1007–1015.

Clauset A., Shalizi C. R., Newman M. E. (2009). Power-law distributions in empirical data. *SIAM review*, vol. 51, n° 4, p. 661–703.

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

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*, p. 234–241.

Fan X., Cao L., Xu R. Y. D. (2013). Dynamic infinite mixed-membership stochastic blockmodel. *CoRR*, vol. abs/1306.2999.

Goldenberg A., Zheng A. X., Fienberg S. E., Airoldi E. M. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, vol. 2, n° 2, p. 129–233.

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

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

Hasan M. A., Zaki M. J. (2011). A survey of link prediction in social networks. In C. C. Aggarwal (Ed.), *Social network data analytics*, p. 243–275. Springer. Consulté sur 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*, p. 657–664.

Jacobs A. Z., Clauset A. (2014). A unified view of generative models for networks: models, methods, opportunities, and challenges. *arXiv preprint arXiv:1411.4070*.

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

Koutsourelakis P.-S., Eliassi-Rad T. (2008). Finding mixed-memberships in social networks. In *Aaai spring symposium: Social information processing*, p. 48–53.

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*, p. 601–610.

Largeron C., Mougel P. N., Benyahia O., Zaïane O. R. (2017). Dancer: dynamic attributed networks with community structure generation. *Knowl Inf Syst*, p. 1-43.

Largeron C., Mougel P.-N., Rabbany R., Zaïane O. R. (2015, 04). Generating attributed networks with communities. *PLoS ONE*, vol. 10, n° 4, p. e0122777. Consulté sur http://dx.doi.org/10.1371%2Fjournal.pone.0122777

Lazarsfeld P. F., Merton R. K. *et al.* (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, vol. 18, n° 1, p. 18–66.

Lee J., Zaheer M., Günnemann S., Smola A. J. (2015). Preferential attachment in graphs with affinities. In *Proceedings of the eighteenth international conference on artificial intelligence and statistics, AI-STATS 2015, san diego, california, usa, may 9-12, 2015.* Consulté sur http://jmlr.org/proceedings/papers/v38/lee15b.html

Leskovec J., Backstrom L., Kumar R., Tomkins A. (2008). Microscopic evolution of social networks. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining*, p. 462–470. ACM.

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.

Liben-Nowell D., Kleinberg J. M. (2007). The link-prediction problem for social networks. *JASIST*, p. 1019–1031.

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

Meeds E., Ghahramani Z., Neal R. M., Roweis S. T. (2006). Modeling dyadic data with binary latent factors. In *Advances in neural information processing systems*, p. 977–984.

Miller K., Jordan M. I., Griffiths T. L. (2009). Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*, p. 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*, p. 1–6.

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

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

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.*

Teh Y. W., Jordan M. I., Beal M. J., Blei D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, vol. 101, nº 476, p. 1566-1581.

Zeng S. (2016). Link prediction based on local information considering preferential attachment. *Physica A: Statistical Mechanics and its Applications*, vol. 443, p. 537-542.

Zhang H., Zhao T., King I., Lyu M. R. (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*, p. 3938–3944.