Social Networks Topology with Bayesian Perspectives

***Abstract*—Burstiness, preferential attachment, Assortativity, community detection, link prediction, dyadic relation, latent class model, latent feature model. Matrix Factorization, Bayesian, uncertanty, regularisation**

I. Introduction

We provide formal definitions of fundamental properties of social networks which are design to be consistent with the probabilistic framework. We study those properties on two general models based on Bayesian nonparametric prior namely the Hierarchical Dirichlet Process (HDP) and the Indian Buffet Process (IBP). We show the relation between those properties and the models. Thus it provides a better comprehension of the models and their limitation in order to capture those properties in a learning problem. Additionally, we propose an adaptation of priors which gives a better interpretation of models in terms of assumptions on social networks and lead to better prediction performance.

II. Motivations

Recently, several complex Bayesian models based on latent variables to explain the structure of social networks have been introduced [mmsb, ilfrm, etc]. This work was mainly evaluated on prediction tasks, such as link prediction or communities detection. However, few works have been done concerning the study of the intrinsic capacity of the models to model basic properties that arise in social networks, such as the dynamics of degree distribution, known to exhibit the preferential attachment effect [barabasi, web..] or the homophily effect[ref].

(++ Indeed the most heavily studied properties in social networks was the degree distribution and the mixing pattern (homophily/assortativity) tableaux !)

(++ not clear consensus of the formalism of properties and their evaluation, and whatsoever for the homophily property, the feature the definition are usually for single attribute... We consider a general vector . (with a measure working for both latent and real features)

(++ Probabilistic models we are interested in provide two ways of representing the data or network. One fall in the paradigm of mixture models and the other in the latent feature modeling. A motivation of those two modeling paradigm is that they are consistent with two key nonparametric prior for discrete data, namely the Dirichlet process (DP) and the the Indian Buffet Process (IBP). Many baysian model can be view as equivalent to truncated models with nonparametric priors. This provide a motivation to study those models. Furthermore, they are used as priors to generate latent features, either as proposition vector (class/DP) or binary vector (feature/IBP). It is admitted that those priors gives bursty features [accounting for burstiness in topic model]. We seek to clarify why this is true and how the burstiness can propagate at the degree level.

In the next section we will, first, explain the mathematical background in a machine learning context. Secondly, we will review the models of interest for dyadic data. Then, we will introduce the formal definition of properties of interest in social networks within the Bayesian frameworks, and how this is translated in terms of assumptions within Bayesian priors. Finally, we will show empirical results (on synthetic and real datasets) to support our claims.

III. Related Work

= Prop Burstiness on topic model: Modeling Word Burstiness Using the Dirichlet Distribution (DCM) Accounting for Burstiness in Topic Models (DCMLDA) LDA bursty on topics Proposal of a-MMSB in : Scalable Inference of Overlapping Communities with high diagonal only...

to read: Stochastic blockmodels and community structure in networks

= Model Recent work on MMSB and copula: Copula Mixed-Membership Stochastic Blockmodel with Subgr oup Correlation

IV. Background

A. Relational Learning

Relational learning provides a framework for predictive task in graph based data. Even though we focus on graphical models, we seek for a general representation in this framework. Indeed most of the models have a matrix factorization representation 1. While bringing a bridge between classical matrix factorization approach, such as ICA and NMF, it represents a common frameworks for latent variables analysis to emphasize core structural similarities of observed variables. A formal introduction that review this bridge was done in [**?**], and has been generalized to Mixed Membership Models [**?**].

B. Bayesian Models

Our work rely on two concurrent models in this framework. They account for baselines in our analysis, although they are near state of the art approach [**?**], [**?**].

1) Proportion feature – MMSB: The first model we focus on falls down in the *latent class* category. In this link prediction model, each node has a latent feature vector of class proportion. For each single interaction between two nodes, one class is drawn for each one. The probability to have a link is only conditioned on the classes assignment. This model has a natural interpretation in term of soft clustering, by predicting the link structure according to which class nodes belong to in an Latent Dirichlet Allocation like generative process. Our reference for the class based model is the Mixed Membership Stochastic Blockmodel (MMSB) [**?**] and its nonparametric version using a Hierarchical Dirichlet Process prior (HDP) [**?**].

2) Binary feature – ILFM: The second model of interest is related to the *latent feature* category. Here, each node has an associated feature vector and the model uses the features to predict the link structure. In this approach, and with binary features, ones's can see each active features (set to one) as a membership indicator for the corresponding node. Our reference for the feature based model is binary matrix factorization (BMF) [**?**] and its nonparametric version using a Indian Buffet Process (IBP) [**?**] known as the Infinite Latent Feature Model (ILFM) [**?**]. Note that we will avoid to refer to this model by the term *latent feature* because both latent class/feature model carry the notion of latent features regarding nodes of a network, either as a proportion vector or as a binary vector. (– Rename the latent feature model to the Mixed Membership Deterministic Blockmodel (MMDB) vs MMSB. (see graphical model).)

The difference between the two approaches can be expressed by the structure of the Bayesian network or Graphical Model (GM) behind the generative model, and the type of priors chosen for the random variables distributions V. Nevertheless the GM formalizes the regularization applied when fitting the model in a meaningful way, while both models can have a common interpretation in terms of matrix factorization. Thus, pursuing the approach in [**?**], it allows us to highlight a structural similarity in the general field of relational learning.

C. Social Networks

Without loss of generality, we focus on social networks with binary relationships. Our object of interest is the topology of the network representing the presence or absence of links between nodes in the graph. The network can be either directed or not. For a network with N nodes, we represent the topology by an adjacency matrix $Y \in \{0,1\}^{N\times N}$ associated to a graph $G = (V, E)$, where V is a set of nodes representing entities, $E \in V \times V$ is a set of edges who represents relationships between pairs of entities. From a probabilistic point of view, the network topology is modeled using a kernel with a Bernoulli density. The parameters of the Bernoulli is the probability to observe a link between two nodes.

We define a matrix of weight interactions $\Phi \in W^{K\times K}$ with W the space of weights, where K is the number of classes or features. Let $\Theta \in \mathcal{F}^{N\times K}$, be a matrix where each row i represents the latent feature vector associated to the node i, and $\mathcal{F}$ the latent feature space. Hence for the MMSB and ILFM, the latent feature vectors are respectively proportion vectors (who sum to one) and binary vectors. In this framework the network is generated with the following density:

$$Y \sim \mathrm{Bern}(\sigma(\Theta\Phi\Theta^T)) \tag{1}$$

where σ is a function that map values to a probability space. When σ is the identity function, the expectation of the observation reduces to a matrix factorization (bilinear) expression, and is related to Discrete Component Analysis (DCA) [**?**]:

$$E_{y\sim p(y|\Theta,\Phi)}[Y] = \Theta\Phi\Theta^T \tag{2}$$

This matrix factorization approach of the Bayesian model is in due to the likelihood of the model when applying the sum rule over the latent variables. Indeed the probability to have a link for the interaction (i, j) is:

$$\mathbf{p}(y_{ij} = 1 \mid \Theta, \Phi) = \sum_{k,k'} \mathbf{p}(y_{ij} = 1 \mid \phi_{k,k'})\mathbf{p}(k \mid \theta_i)\mathbf{p}(k' \mid \theta_j) \tag{3}$$

The questions that arise are:

- What kind of properties the model can capture or learn on networks ?
- Which constraint on the models can come with an consistent interpretation of latent variables along with the concepts of communities structure and homophily in social networks ?

In the next session we review the models of interest.

V. Models

Yet another view

Our two chosen baseline use prior distributions that fall into the two major classes of discrete nonparametric priors. The Hierarchical Dirichlet Process (HDP) that generalizes the Latent Dirichlet Allocation (LDA) for infinite mixtures models. On the other hand, the Indian Buffet Process (IBP), which is the generalization of the Beta-Bernoulli compound distribution (ie Beta Process), which generates infinite binary matrices. The nonparametric models in their truncated version are equivalent to well-known models such as LDA, widely used for text analysis, and Mixed Membership Stochastic Blockmodel which is an adaptation of the latter for relational learning.

We adopt the following notation; if a matrix has a negative index superscripted, it indicates that the values corresponding to this index are excluded. A dot **.** in the index means that we marginalize over all possible values.

A. Binary Feature

In the latent feature models, the features are distributed according to an Indian Buffet Process (IBP), and the weights interaction according to a Gaussian :

$$\Theta \sim \mathrm{IBP}(\alpha) \quad \text{is a } N \times K \text{ matrix} \tag{4}$$

$$\phi_{mn} \sim N(0, \sigma_w) \quad \text{for } m, n \in \{1, .., K\}^2 \tag{5}$$

The observation level is defined by deterministically selecting the row of Θ which is distributed according an IBP prior. Hence for a node i, we note his feature vector by $F_i = \theta_i$ and for $i, j \in V$ we have:

$$y_{ij} \sim \text{Bern}(\sigma(F_i \Phi F_j^\top)) \tag{6}$$

Finally the function $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function to map $[-\infty, +\infty]$ values to [0,1], a probability space.

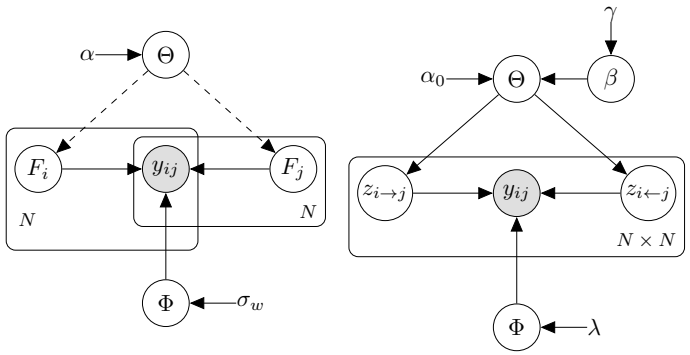

Fig. 1. The two Graphical model for (left) the latent feature model and (right) the latent class model. The conceptual difference between the two model are the actualization of latent variables for each interaction. The latent feature model uses the same latent features vector for each node's interaction, while the latent class model draws new variables at each interactions.

1) MCMC Updates for Posterior Inference: The update for the latent features are obtained by the following Gibbs updates:

$$P(f_{ik} = 1 \mid F^{-ik}) = \frac{m^{-ik}}{N} \tag{7}$$

$$P(f_{ik} = 0 \mid F^{-ik}) = 1 - \frac{m^{-ik}}{N} \tag{8}$$

Where m^{-ik} represents the number of active features k for all entities excluding entity i, hence $m^{-ik} = \sum_{j \neq i} f_{jk}$.

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

$$P(\phi_{mn} \mid Y, F, \phi_{-mn}, \sigma_w) \propto P(Y \mid F, \Phi) P(\phi_{mn} \mid \sigma_w) \tag{9}$$

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

$$J(\phi^*_{mn} \mid \phi_{mn}) = \mathcal{N}(\phi_{mn}, \eta) \tag{10}$$

with η a parameter letting us controlling the acceptance ratio.

The acceptance ratio of ϕ^*_{mn} is thus:

$$r_{\phi_{mn} \to \phi^*_{mn}} = \frac{P(Y \mid F, \Phi^*) P(\phi^*_{mn} \mid \sigma_w) J(\phi_{mn} \mid \phi^*_{mn})}{P(Y \mid F, \Phi) P(\phi_{mn} \mid \sigma_w) J(\phi^*_{mn} \mid \phi_{mn})} \tag{11}$$

B. Proportion Feature

In the latent class model, the rows of the feature matrix Θ are Dirichlet distributed, and the weights interaction are Beta distributed :

$$\theta_i \sim \text{Dirichlet}(\alpha_0) \quad \text{for } i \in \{1, .., N\} \tag{12}$$

$$\phi_{mn} \sim \text{Beta}(\lambda) \quad \text{for } m, n \in \{1, .., K\}^2 \tag{13}$$

The observation level is defined by multinomial draws to obtain class assignments for each nodes. The likelihood for a links depends only on the class assignments of the nodes and for $i, j \in V$, we have:

$$z_{i \to j} \sim \text{Mult}(\theta_i) \tag{14}$$

$$z_{i \leftarrow j} \sim \text{Mult}(\theta_j) \tag{15}$$

$$y_{ij} \sim \text{Bern}(z_{i \to j} \Phi z_{i \leftarrow j}^\top) \tag{16}$$

In this special case, the mapping function σ is the identity.

a) Altenartive Description ! (wich one we keep ?):

An alternate view, maybe more consistent with the IBP representation is to say that:

$$Z \sim CRF(\alpha_0, \gamma) \tag{17}$$

$$\phi_{mn} \sim \text{Beta}(\lambda) \quad \text{for } m, n \in \{1, .., K\}^2 \tag{18}$$

$$y_{ij} \sim \text{Bern}(\Phi_{z_{i \to j}, z_{i \leftarrow j}}) \tag{19}$$

In the (not printed here) corresponding graphical model, the arrow between $z_{\to}$ and Z are dashed as for the IBP representation. Note that $Z \in \{1, .., K\}^{N \times N \times 2}$.

@ERIC: This representation would justify to explain our interpretation for the Chinese Restaurant Franchise (CRF), because we see it explicitly here...

1) MCMC Updates for Posterior Inference:

In the latent class model, the collapsed gibbs sampling approach allow us to only sample the classes couple for each observations:

$$\mathbf{p}(z_{ji} = k, z_{ij} = k' \mid .) \propto \mathbf{p}(z_{ji} = k \mid .) \mathbf{p}(z_{ij} = k' \mid .) f^{-ji}_{(k,k')}(y_{ji}) \tag{20}$$

The class assignment updates for the node couple are:

$$\mathbf{p}(z_{i \to j} = k \mid Z^{-ij}) \propto N^{-ij}_{ik} + \alpha_0 \tag{21}$$

$$\mathbf{p}(z_{i \leftarrow j} = k \mid Z^{-ij}) \propto N^{-ij}_{jk} + \alpha_0 \tag{22}$$

And the likelihood of a link given the couple $c = (k, k')$ is:

$$f^{-ij}_{(k,k')}(y_{ij} = r) = \frac{C^{-ij}_{(k,k')r} + \lambda_r}{C^{-ij}_{(k,k')\cdot} + \sum_r \lambda_{r'}} \tag{23}$$

Finally Θ and Φ can be reconstructed from the count matrices with the following equations:

$$\theta_{ik} = \frac{N_{ik} + \alpha_0}{N_{i\cdot} + K\alpha_0} \tag{24}$$

$$\phi_{rc} = \frac{C_{cr} + \lambda_r}{C_{c\cdot} + \lambda_\cdot} \tag{25}$$

C and N are counts matrices and are describes in section VII-B. We refer to the supplementary material for detail of derivations for MCMC updates.

C. Comparison of models

class proportion vs feature vector
Class strength vs feature correlation/metric ?
HDP vs IBP
complexity $O((E^2K^2)$ vs $O(NK^3)$
property yes/no

VI. Homophily

Birds of a feather flock together

Homophily describes the fact that two nodes are more likely to be connected if they share common characteristics [**?**], [**?**]. So, the more similar the nodes, the more likely is it to be connected. In their research on dynamic attributed networks where the attributes are discrete (e.g. age, gender etc.).Some definition has been proposed [**?**], but they consider unimodal feature and lack of generality to use them in a Bayesian framework. We proposed a more general definition of homophily at the model level:

Definition VI.1 (Homophily)**.** *Given a model $\mathcal{M}_s = \{\Theta, \Phi, s\}$, defined by its parameters and a function. The parameters define respectively the latent features of nodes and a matrix of weight controlling the strength between the features. The function define a similarity measure $s(.; \Theta, \Phi)$ on $V \times V$ and parametrized by the model parameters. We consider that the model exhibits homophily if:*

if $s(i,j) > s(i',j')$ then $\mathbb{E}_\Phi[\mathbf{p}(y_{ij} = 1 \mid \Theta)] > \mathbb{E}_\Phi[\mathbf{p}(y_{i'j'} = 1 \mid \Theta)]$
$\forall (i,j,i',j') \in V^4$

A model which verifies this condition is said to be homophilic.

The similarity aims at measuring the closeness between two node, given the (latent) features. The homophily evaluate if the order is preserved between the similarity and the probability to create a link between two nodes when averaging on the weight matrix of the model.

In both models, we can assess a natural similarity which arises from the Bernoulli parameter. That is to say that the probability to have a link for the couple (i,j) is defined as the bilinear form such as:

$$s(i,j) = \theta_i \Phi \theta_j^\top \tag{26}$$

Proposition VI.1. *Let's suppose a model $\mathcal{M}_s = \{\Theta, \Phi, s\}$ with the similarity measure defined as $s(i,j) = \theta_i \Phi \theta_j^\top$, then the ILFM and MMSB models are homophilic.*

Proof: Suppose that for two couple of nodes $(i,j,i',j') \in V^4$ we have $s(i,j) > s(i',j')$ with $s(i,j) = \theta_i \Phi \theta_j^\top$, and σ a either strictly increasing function or the identity function. We have:

$$\mathbb{E}_\Phi[\mathbf{p}(y_{ij} = 1 \mid \Theta)] - \mathbb{E}_\Phi[\mathbf{p}(y_{i'j'} = 1 \mid \Theta)] \tag{27}$$

$$= \int_\Phi \mathbf{p}(y_{ij} = 1 \mid \Theta, \Phi)\mathbf{p}(\Phi)d\Phi - \int_\Phi \mathbf{p}(y_{i'j'} = 1 \mid \Theta, \Phi)\mathbf{p}(\Phi)d\Phi \tag{28}$$

$$= \int_\Phi (\mathbf{p}(y_{ij} = 1 \mid \Theta, \Phi) - \mathbf{p}(y_{i'j'} = 1 \mid \Theta, \Phi))\mathbf{p}(\Phi)d\Phi \tag{29}$$

$$= \int_\Phi (\sigma(\theta_i \Phi \theta_j^T) - \sigma(\theta_{i'} \Phi \theta_{j'}^T))\mathbf{p}(\Phi)d\Phi \tag{30}$$

Since σ are strictly increasing (or the identity), the left hand part and the right hand are strictly positive function we have $\mathbb{E}_\Phi[\mathbf{p}(y_{ij} = 1 \mid \Theta)] > \mathbb{E}_\Phi[\mathbf{p}(y_{i'j'} = 1 \mid \Theta)]$.

■

In this case, the homophily effect holds in both models as it is the parameter of the Bernoulli. In other words, it is the probability to observe a link between i and j. Thus the metric for the network topology and the node similarity are the same. Nevertheless it is not obvious to state the role that plays latent variables for the homophily because of the weak constraints applied on the interactions matrix Φ. In other word we loose the intuition behind the similarity.

But It is important to note that the metric for the topology of the network and the node similarity has not to be same, as it is a very restrictive assumption. Especially for an real networks, node features may be accessible and without a priori knowledge a natural metric that can be applied on both real network with real features and the model with latent features is cosinus based measure.

A natural constraint for the *similarity metric* is to use some cosinus based measure. Furthermore those measure only rely on the latent/real feature of nodes. It thus make them applicable to model and real networks. We will consider next the dot product as our similarity based measure; $s(i,j) = \theta_i \theta_j^\top$.

A natural constraint for the *topology metric*, would be to stay consistent with the notion of communities in networks. In this case nodes belonging to the same community would be more likely to be linked. One way to encode this believe is to have strong values in the diagonal for the interaction matrix Φ. Moreover by defining decreasing weights from the diagonal, we allow inter-communities links and define a hierarchies (ie an order) between the possible interactions between the communities. We develop this idea with constraint prior for the weight interaction matrix.

Proposition VI.2. *Let's suppose a model $\mathcal{M}_s = \{\Theta, \Phi, s\}$ with the similarity measure defined as $s(i,j) = \theta_i \theta_j^\top$ and σ a strictly increasing function or the identity. Let's suppose a $K \times K$ matrix normal prior over weight matrix such as $\Phi \sim \mathcal{N}_K(M, U, V)$ and the mean matrix M to be a diagonal matrix. Then homophily hold for $\mathcal{M}$ if $M = \Lambda I_K$ with $\Lambda > 0$.*

Proof: Suppose that for two couple of nodes $(i,j,i',j') \in V^4$ we have $s(i,j) > s(i',j')$, with $s(i,j) = \theta_i \theta_j^\top$. We first examine the difference of expectations:

$$\mathbb{E}_\Phi[\mathbf{p}(y_{ij} = 1 \mid \Theta)] - \mathbb{E}_\Phi[\mathbf{p}(y_{i'j'} = 1 \mid \Theta)] \tag{31}$$

$$= \mathbb{E}_\Phi[\sigma(\theta_i \Phi \theta_j)] - \mathbb{E}_\Phi[\sigma(\theta_{i'} \Phi \theta_{j'})] \tag{32}$$

$$= \mathbb{E}_\Phi[\sigma(\theta_i \Phi \theta_j) - \sigma(\theta_{i'} \Phi \theta_{j'})] \tag{33}$$

Because σ is stricly increasing (or the identity), we have the following equivalence:

$$\mathbb{E}_\Phi[\sigma(\theta_i \Phi \theta_j) - \sigma(\theta_{i'} \Phi \theta_{j'})] > 0 \iff \mathbb{E}_\Phi[\theta_i \Phi \theta_j - \theta_{i'} \Phi \theta_{j'}] > 0 \tag{34}$$

Then according to matrix normal property we have:

$$\mathbb{E}_\Phi[\theta_i \Phi \theta_j] = \mathbb{E}_\Phi[\mathcal{N}_K(\theta_i M \theta_j, \theta_i U \theta_i, \theta_j V \theta_j)] \tag{35}$$

$$= \theta_i M \theta_j \tag{36}$$

$$= \Lambda \theta_i \theta_j \tag{37}$$

It follows that:

$$\mathbb{E}_\Phi[\theta_i \Phi \theta_j] - \mathbb{E}_\Phi[\theta_{i'} \Phi \theta_{j'})] > \Lambda(\theta_i \theta_j - \theta_{i'} \theta_{j'}) \tag{38}$$

Homophily is satisfied for $\Lambda > 0$. ■

Proposition VI.2 highlight a generalization of the ILFM model where the correlation between feature depend on their proximity. This enable the interpretation of features as communities because node who belongs to the same membership have a higher probability to binds.

A. Statistical Evaluation

To evaluate the homophily on a given networks (an instance), we propose a statistical test. We seek for both a qualitative as well a quantitative evaluation of the homophily.

VII. Burstiness

The rich get richer and the poor get poorer

The preferential attachment states that a node is more likely to create connections with nodes having a high degree. To take into account this behavior, in the BarabasiAlbert (BA) model, each node is connected to an existing node with a probability proportional to the number of links of the chosen node. This leads to scale-free networks, characterized by a degree distribution with a heavy tail which can be approximated by a power law distribution such that the fraction of nodes $\mathbf{p}(d)$ having a degree d follows $\mathbf{p}(d) \sim d^{-\gamma}$ where γ ranges typically between 2 and 3 [**?**]. An equivalent notion is the burstiness, studied by [**?**], which conveys the same idea : rich get richer or the more you have, the more you will get. In [**?**], a formalized definition has been proposed. According to the authors:

Definition VII.1 (Burstiness)**.** *A discrete distribution* $\mathbf{p}$ *is bursty if and only if for all integers* (n, n')*,* $n \geq n'$ *:*

$$\mathbf{p}(d \geq n' + 1 \mid d \geq n') > \mathbf{p}(d \geq n + 1 \mid d \geq n) \tag{39}$$

A distribution which verifies this condition is said to be bursty.

In [**?**], this definition has been generalized to the continuous case but, in the sequel, we will retain this first definition since we focus on discrete distributions.

The burstiness can appear for different various variable in a model. In this paper we consider three different schemes at the node and feature level, that constitute some basic topology assumptions on networks:

Proposition VII.1.
for all $i, j \in V^2$ *and* $k \in \{1, .., K\}$*, we have:*

- *Preferential Attachment: the distribution of degree* d_i *is bursty iff* f_i *is a stricly increasing function of n with:*

$$f_i(n) = \mathbf{p}(y_{ij} = 1 \mid d_i = n)$$

- *Local Preferential Attachment: Given a class couple* $c = \{k, k'\} \in \{1, .., K\}^2$*, and a degree restricted to nodes who belongs to this interaction couple* d_{ic}*, the distribution of degree* d_{ic} *is bursty iff* $f_{i,c}$ *is a stricly increasing function of n with:*

$$f_{i,c}(n) = \mathbf{p}(y_{ij} = 1 \mid d_i = n, c)$$

- *Feature burstiness (block/class burstiness ?:s): the distribution over the number of membership of each class* $\theta_{.k}^{-ik}$ *is bursty iff* f_k *is a stricly increasing function of n with:*

$$f_k(n) = \mathbf{p}(\theta_{ik} \mid \theta_{.k}^{-ik} = n)$$

We justify this approach in the supplementary materials X-B. One can see that the approach makes the link with the classical definition of preferential attachment. Furthermore one can see that the similaraty between th functions we track $(f_i, f_{i,c}, f_k)$ and the typical Gibbs updates. The difference is that we want to assess the generative model given the data and parameters of the model (ie the model has converged). Hence we are looking if the topological property of burstiness can be handle by the model once it learned from the data.

A. Burstiness for ILFM

In this model, the weight interaction matrix Φ are not conjugate of the likelihood. Thus it can not be integrated out into a closed form expression. As a matter of simplicity we consider this parameter as known, and omit it the following conditional distributions.

Let $F = \Theta$ and $W = \Phi$, each node i has a fixed feature vector noted F_i and a weighed interactions matrix W. In this case, the function f_i is:

$$\mathbf{p}(y_{ij} = 1 \mid d_i, F^{-i\cdot}) = \sum_{F_i} \mathbf{p}(y_{ij} = 1 \mid d_i, F^{-i\cdot} F_i,) \mathbf{p}(F_i \mid d_i, F^{-i\cdot}) \tag{40}$$

$$= \sum_{F_i} \sigma(F_i W F_j^T) \frac{\mathbf{p}(d_i \mid F^{-i\cdot}, F_i) \mathbf{p}(F_i \mid F^{-i\cdot})}{\mathbf{p}(d_i \mid F^{-i\cdot})} \tag{41}$$

$$\propto \sum_{F_i} \prod_{j' \in \mathcal{V}(i) \cup j} \sigma(F_i W F_{j'}^T) \prod_{j' \notin \mathcal{V}(i)} 1 - \sigma(F_i W F_{j'}^T) \prod_k \frac{m_{ik}}{N} \tag{42}$$

The term $\prod_k \frac{m_{ik}}{N}$ latter equation comes from the conditional probability of a feature f_{ik} for an IBP prior and applying a chain of product rule. The product is the only term that depend on d_i and we refer to it as $f(d_i)$. Under the assumption that all observed links have higher probability than the observed non-links to bind, we can choose an index dictionary g to reorder the terms such as:

$$\underbrace{\sigma(F_i W F_{g(1)}^T) \geq .. \geq \sigma(F_i W F_{g(p)}^T)}_{g(.) \in \mathcal{V}(i) \cup j} \geq \underbrace{.. \geq \sigma(F_i W F_{g(N)}^T)}_{g(.) \notin \mathcal{V}(i)} \tag{43}$$

We then have with some regularity:

$$f(d_i) \geq \sigma(F_i W F_{g(p)}^T)^{d_i + 1} (1 - \sigma(F_i W F_{g(p+1)}^T))^{N - d_i} \tag{44}$$

$$log(f(d_i)) \geq d_i log(\frac{\sigma(F_i W F_{g(p)}^T)}{1 - \sigma(F_i W F_{g(p+1)}^T)}) + cst \tag{45}$$

Reste a valider 2 point:

- Le passage la proportion
- Borner $log(f(d_i))$ entre deux fonction croissantes et montrer qu'on oscille pas l'intrieur ?!

Then a sufficient condition for the burstiness is to have: $\sigma(F_i W F_{g(p)}^T) > 1 - \sigma(F_i W F_{g(p+1)}^T)$. If σ is the sigmoid this is equivalent ot have $F_i W F_{g(p)}^T > -F_i W F_{g(p+1)}^T$.

In other term to enable burstiness in the ILFM, the model need to ensure some deterministic behavior when regarding the realization of outcomes and there actual distribution.

B. Burstiness for MMSB

In the latent class models each dyads has two underlying class assignments for each node of the couple. We note $Z \in N \times N \times 2$ the matrix that represents those class assignments. We seek for the following form of the likelihood, that we marginalize over all the possible couples classes $c = (k, k')$:

$$\mathbf{p}(y_{ij} = 1 \mid Y^{-i\cdot}, Z^{-ij}, d_i) = \sum_{c=(k,k')} \mathbf{p}(y_{ij} = 1 \mid Y^{-i\cdot}, d_i, c)\mathbf{p}(c \mid Z^{-ij}) \tag{46}$$

Here note that within the sum, the left hand term is conditionally independent of Z^{-ij}. And the right hand term is independent of the adjacency terms $Y^{-i\cdot}$ since it do not belongs to the Markov blanket of c random variable.

The first term is the likelihood for the links between (i, j) given the class of each node (k, k'). Due to the Beta-Bernoulli conjugacy of the model, ϕ and θ can be marginalized out, and it simplify to:

$$\mathbf{p}(y_{ij} = 1 \mid Y^{-i\cdot}, d_i, c) = \frac{C_{c1}^{-i\cdot} + d_{ic} + \lambda_1}{C_{c\cdot}^{-ij} + \lambda_0 + \lambda_1} \tag{47}$$

Where C_{c1} denotes the count matrix for all interactions having value 1 (link present) with the classes couple being $c = (k, k')$. Thus $C_{c1} = \sum_{i,j} \mathbf{1}(z_{i\to j} = k, z_{i\leftarrow j} = k', y_{ij} = 1)$ and $C_{c\cdot} = \sum_{i,j} \mathbf{1}(z_{i\to j} = k, z_{i\leftarrow j} = k')$

We recognize the likelihood form of the Gibbs update [**?**], except that we isolate the term depending of the degree on i, d_i. Hence the term d_{ic} is the element of the degree with a classes couple $c = (k, k')$ and $d_{ic} = \sum_{j' \neq j} \mathbf{1}(z_{i\to j'} = k, z_{i\leftarrow j'} = k', y_{ij'} = 1)$.

The second term of equation (46), can be rewrited by noting that the classes of the couple c are independent and that the term $Y^{-j\cdot}$ can be dropped because it is not present in the Markov blanket of the class assignment:

$$\mathbf{p}(c \mid Z^{-ij}) = \mathbf{p}(z_{i\to j} = k \mid Z^{-ij})\mathbf{p}(z_{i\leftarrow j} = k' \mid Z^{-ij}) \tag{48}$$

Again, the two members of the right hand equation (48) are the Gibbs updates for the topic assignments of nodes for the interaction (i, j). Both members reduce to simple form due to the conjugacy between the Dirichlet and Multinomial [**?**] or concurrently from the Chinese Restaurant Franchise [**?**]:

$$\mathbf{p}(z_{i\to j} = k \mid Z^{-ij}) = \frac{N_{ik}^{-ij} + \alpha_k}{N_{i\cdot}^{-ij} + \alpha_\cdot} \tag{49}$$

$$\mathbf{p}(z_{i\leftarrow j} = k' \mid Z^{-ij}) = \frac{N_{jk'}^{-ij} + \alpha_{k'}}{N_{j\cdot}^{-ij} + \alpha_\cdot} \tag{50}$$

Finally, one can see that the only term depending on the degree d_i is isolated, and we can rewrite equation (46), with term depending only on d_i, k, i and j:

$$\mathbf{p}(y_{ij} = 1 \mid Y^{-i\cdot}, Z^{-ij}, d_i) = \sum_{c=(k,k')} A_c(B_c + d_{ic}) \tag{51}$$

Where A_c and B_c are two positive function of c.

$$A_c = \frac{N_{ik}^{-ij} + \alpha_k}{N_{i\cdot}^{-ij} + \alpha_\cdot} \frac{N_{jk'}^{-ij} + \alpha_{k'}}{N_{j\cdot}^{-ij} + \alpha_\cdot} \frac{1}{C_{c\cdot}^{-ij} + \lambda_0 + \lambda_1} \tag{52}$$

$$B_c = C_{c1}^{-i\cdot} + \lambda_1 \tag{53}$$

As we sum over all possible couple classes, the probability to have a link will augment with the degree with the classes couple corresponding to the element of the degree with the same couple. Hence the probability to observe a link for node i is strictly crescent with his degree d_i.

a) Preferential Attachment:

The model is bursty hence it can handle the preferential attachment at the network level.

b) Local Preferential Attachment:

The Local preferential attachment is similar to the notion of burstiness but inside a community/class of the network. Assuming that we know the class of i $z_{i\to j}$ to be k, the probability to have a link becomes:

$$\mathbf{p}(y_{ij} = 1 \mid Y^{-i\cdot}, d_i, z_{i\to j}) = \sum_{k'} \mathbf{p}(y_{ij} = 1 \mid Y^{-i\cdot}, Z^{-ij}, d_i, c = (k, k'))\mathbf{p}$$

(54)

$$= \sum_{k'} \frac{C_{c1}^{-i\cdot} + d_{ic} + \lambda_1}{C_{c\cdot}^{-ij} + \lambda_0 + \lambda_1} \frac{N_{jk'}^{-ij} + \alpha_{k'}}{N_{j\cdot}^{-ij} + \alpha_\cdot} \tag{55}$$

$$= \sum_{k'} A'_{k'}(B'_{k'} + d_{i(k,k')}) \tag{56}$$

Here the probability increases with the degree independently of the interactions classes. This means that burstiness is possible inside but also between communities.

c) Communities Distribution:

....Need to count the table for each classes in Chinese Restaurant Franchise (CRF), to evaluate the distribution according to the hyperprior of HDP...

VIII. Empirical results

To validate our theoretical results we fitted our models on synthetic networks and track how well we can reproduce the properties of interest on a generated network.

The synthetic network has 1000 nodes and 4 communities and a density of 0.05. [See the ref of the generator for the ground true on the preferential attachment effect...]

IX. Homophily Indicator

We consider a social network defined as an attributed graph G = (V,E), where V is a set of N nodes representing entities, $E \in V \times V$ is a set of m edges representing relationships between pairs of entities. Each node $i \in V$ is described by K features and s is a similarity function which allows to compare two vertices according to their features. We consider that two vertices are similar, denoted s(x,y), if s(x,y) is lower than a threshold.

Given a contingency table defined as follows:
$a = Card\{(x,y) \in V \times (V-1)/(x,y) \in E \wedge s(x,y)\}$
$b = Card\{(x,y) \in V \times (V-1)/(x,y) \in E \wedge \neg s(x,y)\}$
$c = Card\{(x,y) \in V \times (V-1)/(x,y) \notin E \wedge s(x,y)\}$
$d = Card\{(x,y) \in V \times (V-1)/(x,y) \notin E \wedge \neg s(x,y)\}$
$\frac{N*(N-1)}{2}$ is the total count of the cells in the contingency table.

The measure that we introduced to evaluate the homophily in the network is given by:

$Hobs(G) = \frac{2[(a+d)-(c+b)]}{N*(N-1)}$
This measure takes is value between -1 and 1. It is equal to 1 when all the pairs of similar vertices are linked and all the pairs of dissimilar verices are not linked. Otherwise, when all pairs of similar vertices are not linked and all pairs of dissimilar verices are linked, it is equal to -1.

This measure of observed homophily in the network can be compared with an expected value computed on a network having the same number of vertices and edges but where the probability of having a link beteween two vertices is independant of their similarity and consequently of their features, which does not respect the homophily property according which two vertices are more likely to be connected if they share common characteristics.

In order to compute the expected homophily indicator we estimate the probability for pairs of vertices of being linked and similar in the following way:
$PR = \frac{2M}{N*(N-1)}$
$PS = \frac{2*Card\{(x,y) \in V \times (V-1)/s(x,y)\}}{N*(N-1)}$
with $PNR = 1 - PR$ and $PNS = 1 - PS$

Then, with the following contingency table:
$a' = \frac{PR*PS*N*(N-1)}{2}$
$b' = \frac{PNR*PS*N*(N-1)}{2}$
$c' = \frac{PR*PNS*N*(N-1)}{2}$
$c' = \frac{PNR*PNS*N*(N-1)}{2}$
we compute the expected homophily as follows:
$Hexpect(G) = \frac{2[(a'+d')-(c'+b')]}{N*(N-1)}$

A social network exhibits homophily if $Hobs(G)$ is higher than $Hexpect(G)$.

X. Supplementary materials

A. Class based derivation

a) Likelihood::

We mention that the ϕ and θ matrix can be reconstructed with Z. From the model, we have the following equalities:

$$\mathbf{p}(y_{ij} \mid \phi_c) = \phi_c^{y_{ij}}(1-\phi_c)^{1-y_{ij}} \tag{57}$$

$$\mathbf{p}(\phi_c \mid \lambda) = \frac{1}{B(\lambda_1,\lambda_0)}\phi_c^{\lambda_1-1}(1-\phi_c)^{\lambda_0-1} \tag{58}$$

Derivation of equation (5.12) of the likelihood:

$$\mathbf{p}(y_{ij} \mid Y^{-ij}, c) \propto \mathbf{p}(y_{ij}, Y^{-ij}, c) \tag{59}$$

$$= \int_{\phi_c} \mathbf{p}(y_{ij} \mid \phi_c)\mathbf{p}(\phi_c \mid \lambda) \prod_{i'j' \neq ij} \mathbf{p}(y_{i'j'} \mid \phi_c) d\phi_c \tag{60}$$

$$\propto \int_{\phi_c} \phi_c^{y_{ij}+C_{c1}^{-ji}+\lambda_1-1}(1-\phi_c)^{1-y_{ij}+C_{c0}^{-ji}+\lambda_0-1} d\phi_c \tag{61}$$

$$\propto \frac{\Gamma(y_{ij}+C_{c1}^{-ji}+\lambda_1)\Gamma(1-y_{ij}+C_{c0}^{-ji}+\lambda_0)}{\Gamma(1+C_{c\cdot}^{-ji}+\lambda_0+\lambda_1)} \tag{62}$$

Considering the case where $y_{ij}=1$, we have :

$$\mathbf{p}(y_{ij}=1 \mid \phi_c) = \frac{C_{c1}^{-ji}+\lambda_1}{C_{c\cdot}^{-ji}+\lambda_0+\lambda_1} \tag{63}$$

b) Class Assignment::

We have from the model the following equalities:

$$\mathbf{p}(\theta_i \mid \alpha) = \frac{\Gamma(\sum_l \alpha)}{\prod_l \Gamma(\alpha)} \prod_l \theta_{il}^{\alpha-1} \tag{64}$$

$$\mathbf{p}(z_{i\to j}=k \mid \theta_i) = \theta_{ik} \tag{65}$$

Derivation of equation (5.14) of class assignment:

$$\mathbf{p}(z_{i\to j}=k \mid Z^{-ij}) = \mathbf{p}(z_{i\to j}=k \mid \{z_{i\to j_0}\}_{j_0\neq j}, \{z_{i\leftarrow j_0}\}_{j_0=1}^{n}) \tag{66}$$

$$\propto \mathbf{p}(z_{i\to j}=k, \{z_{i\to j_0}\}_{j_0\neq j}, \{z_{i\leftarrow j_0}\}_{j_0=1}^{n}) \tag{67}$$

$$= \int_{\theta_i} \mathbf{p}(\theta_i \mid \alpha)\mathbf{p}(z_{i\to j}=k \mid \theta_i) \prod_{j_0\neq j} \mathbf{p}(z_{i\to j_0} \mid \theta_i) \prod_{j_0=1}^{n} \mathbf{p}(z_{i\leftarrow j_0} \mid \theta_i) d\theta_i \tag{68}$$

$$= \int_{\theta_i} \frac{\Gamma(\sum_l \alpha)}{\prod_l \Gamma(\alpha)} \theta_{ik}^{N_{ik}^{-ji}+1} \prod_{l\neq k} \theta_{il}^{N_{il}^{-ij}+\alpha-1} d\theta_i \tag{69}$$

$$\propto \frac{\Gamma(\alpha+N_{ik}^{-ji}+1)\prod_{l\neq k}\Gamma(\alpha+N_{il}^{-ij})}{\Gamma(\sum_l(\alpha+N_{il}^{-ji})+1)} \tag{70}$$

$$\propto \alpha + N_{ik}^{-ji} \tag{71}$$

Finally the equality is maintain by the marginalization constant:

$$\mathbf{p}(z_{i\to j}=k \mid Z^{-ij}) = \frac{N_{ik}^{-ji}+\alpha}{N_{i\cdot}^{-ji}+K\alpha} \tag{72}$$

B. Burstiness for degrees Distribution

From the definition of the burstiness, we have for a random variable d:

$$\mathbf{p}(d \geq n'+1 \mid d \geq n') > \mathbf{p}(d \geq n+1 \mid d \geq n) \quad \{\forall (n,n') \mid n' > n\} \tag{73}$$

We now consider the degree d_i for a node i of a network. We can rewrite the burstiness in the discrete case:

$$\mathbf{p}(d_i = n'+1 \mid d_i = n') > \mathbf{p}(d_i = n+1 \mid d_i = n), \quad \{\forall (n,n') \mid n' > n\} \tag{74}$$

Q. is it equivalent to: $\mathbf{p}(d_i)' > 0$?

Let's suppose that the model $\mathcal{M} = \{\Theta, \Phi\}$ has converged to some local optima. Do new predictions will respect the burstiness property ? To answer this question we need to evaluate the predictive distribution and we assume that the model parameters for all data except for node i is knows. Hence, we denote this knowledge as $\mathcal{M}^{-i}$ whose condition the degree distribution. We omit reference to it in the following. Additionally we write $\mathbf{p}(d_i)$ accounting for $\mathbf{p}(d_i = n)$:

$$\mathbf{p}(d_i = n+1 \mid d_i) = \sum_{\mathcal{M}_i} \mathbf{p}(d_i = n+1 \mid d_i, \mathcal{M}_i)\mathbf{p}(\mathcal{M}_i \mid d_i) \tag{75}$$

The likelihood of the degree can now be written using the conditional independence. Note that we omit reference to the model parameters $\mathcal{M}$:

$$\mathbf{p}(d_i = n+1 \mid d_i) = \sum_{j\notin\mathcal{V}(i)} \mathbf{p}(y_{ij}=1 \mid d_i) \prod_{j'\notin\mathcal{V}(i), j'\neq j} \mathbf{p}(y_{ij'}=0 \mid d_i) \tag{76}$$

$$= \sum_{j\notin\mathcal{V}(i)} (1-\mathbf{p}(y_{ij}=0 \mid d_i)) \prod_{j'\notin\mathcal{V}(i), j'\neq j} \mathbf{p}(y_{ij'}=0 \mid d_i) \tag{77}$$

$$= ... \tag{78}$$

Where $\mathcal{V}(i)$ represent all vertex connected to i and $|\mathcal{V}(i)| = d_i$.

XI. SAMPLING $beta$

According to [**?**], β is distributed as follows:

$$\beta = (\beta_1, ..., \beta_K, \beta_u) \sim Dir(m_{.1}, ..., m_{.K}, \gamma) \tag{79}$$

Where $m_{.k}$ represent the number of tables serving the dish k in all restaurants, in the chinese restaurant franchise. The sampling of the table configuration $\boldsymbol{m}$ can be done using the unsigned Stirling numbers of the first kind $s(n, m)$ [**?**]:

$$\mathbf{p}(m_{ik} = m \mid Z, \boldsymbol{m}^{-jk}, \beta) = \frac{\Gamma(\alpha_0 \beta_k)}{\Gamma(\alpha_0 \beta_k + N_{jk})} s(n_{jk}, m)(\alpha_0 \beta_k)^m \tag{80}$$