Stochastic mixed membership models and link prediction: a study of homophily and preferential attachment in social networks

Adrien Dulac[a,b], **Eric Gaussier**[a], **and Christine Largeron**[b,1]

[a]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG - F-38000 Grenoble; [b]Université Jean Monnet, Laboratoire Hubert Curien - F-42000 Saint-Etienne

This manuscript was compiled on April 6, 2017

We assess here whether standard stochastic mixed membership models are adapted to link prediction in social networks by studying how they handle homophily and preferential attachment. To do so, 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 attachement, 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.

Stochastic mixed membership models | Social networks | Homophily | Preferential attachment

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 (1, 2). 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 (3) and its non-parametric extension (4), and the mixed-membership stochastic block model (5), and its non parametric extension (6, 7).

Although drawn from a wide range of domains, real world social networks exhibit general properties. In this work, we focus on two important such properties, namely *homophily* and *preferential attachment* (8, 9), and assess to which extent link prediction models, as the ones mentioned above, comply with them.

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*. 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 $\mathbf{\Phi}$ 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 $\mathbf{\Phi}$ are generated. As mentioned before, and to be as general as possible, we consider here the non-parametric versions of the latent feature model (4), referred to as `ILFM`, and of the mixed-membership stochastic block model (6, 7), referred to as `IMMSB`. All our results are nevertheless also valid for the finite versions of these models.

In `ILFM`, the factor matrix $\mathbf{F}$ is generated by an Indian Buffet Process, and each element of the matrix $\mathbf{\Phi}$ is generated according to a normal distribution with 0 mean and fixed, common variance. The Indian Buffet Process yields binary vectors; the k^{th} coordinate of the vector $\mathbf{f}_i$, denoted f_{ik}, is thus either 1 or 0, meaning that node i belongs or not to the k^{th} latent factor. In `IMMSB`, the factor matrix $\mathbf{F}$ is obtained with a Hierarchical Dirichlet Process, and each element of the matrix $\mathbf{\Phi}$ is generated according to a Beta distribution with fixed, common parameters. The Hierarchical Dirichlet Process yields this time membership probabilities; f_{ik} directly encodes the probability that node i belongs to the k^{th} latent factor. We refer the interested reader to (4) and (6) for more details on these processes.

Once the factor and weight matrices have been generated, the probability to generate a link between any two nodes i

Significance Statement

We introduce formal definitions of the compliance of probabilistic link prediction models to homophily and preferential attachment in social networks, and show 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 attachement, 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.

All authors contributed equally to the theoretical development and experimental design. A. Dulac furthermore implemented all the models and ran the experiments.

*As mentioned in (10), 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.

[1]To whom correspondence should be addressed. E-mail: Christine.Largeron@univ-st-etienne.fr

and j for ILFM is given by:

$$p(y_{ij} = 1|\mathbf{F}, \mathbf{\Phi}) = \sigma(\mathbf{f}_i \mathbf{\Phi} \mathbf{f}_j^\top) \quad [1]$$

where $\sigma()$ is the sigmoid function and $\top$ denotes the transpose. In IMMSB, one has first to select two latent factors from $\mathbf{f}_i$ and $\mathbf{f}_j$, and then generate or not a link between i and j. This amounts to: (a) choose a class from the class membership distributions according to a categorical distribution:

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

and (b) generate a link with probability:

$$p(y_{ij} = 1|, \mathbf{\Phi}) = \phi_{z_{i\to j} z_{i\leftarrow j}} \quad [2]$$

where $\phi_{kk'}$ denotes the element of $\mathbf{\Phi}$ at row k, column k'.

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 $\mathbf{\Phi}$, from which new links are predicted. In the remainder, we denote by $\hat{\mathbf{F}}$ and $\hat{\mathbf{\Phi}}$ the estimates of $\mathbf{F}$ and $\mathbf{\Phi}$, that are usually obtained through standard (collapsed) Gibbs sampling and Metropolis-Hastings algorithms (4, 7, 11, 12). We furthermore denote by $\mathcal{M}_e$ the version of both ILFM and IMMSB models in which $\mathbf{F}$ and $\mathbf{\Phi}$ are assumed known and fixed to $\hat{\mathbf{F}}$ and $\hat{\mathbf{\Phi}}$. We now investigate whether, from the learned parameters $\hat{\mathbf{F}}$ and $\hat{\mathbf{\Phi}}$, the new links generated produce a network that comply with homophily and preferential attachement.

1. 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 (13, 14). The characteristics often considered are inherent to the individuals and may represent their social status, their preferences or their interest. 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 (15). However, this definition, which relies on a single characteristic (as 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 1.1 (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 $\hat{\mathbf{F}}$ 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:

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

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. 1). Furthermore, marginalizing over the z variables in IMMSB leads to:

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

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

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

as $\hat{\phi}_{kk} = 0$. Let us now consider $\hat{\mathbf{f}}_i = \hat{\mathbf{f}}_j = (0, 1, 0)$ and $\hat{\mathbf{f}}_{i'} = (0.5, 0, 0.5)$ and $\hat{\mathbf{f}}_{j'} = (0, 1, 0)$. 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 $\hat{\mathbf{f}}_{i'} = (0.5, 0, 0.5)$ by $\hat{\mathbf{f}}_{i'} = (1, 0, 1)$, 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 connected. It is furthermore possible to make these models homophilic under s_l by imposing constraints on the weight matrix $\mathbf{\Phi}$ (and hence its estimate $\hat{\mathbf{\Phi}}$); 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 (16) to find overlapping communities through assortativity constraints in the mixed membership stochastic block model.

2. 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[†]. 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 (17).

Preferential attachment relates to a general phenomenon known as *burstiness*[‡] which describes the fact that some events

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

[‡] A.L. Barabási, for example, uses the term *preferential attachement* in (18), and *burstiness* in (19).

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 (20). In these domains, simple definitions of burstiness, that 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, have been proposed (21, 22), for both discrete and continuous probability distributions. We adapt here these definitions for preferential attachment in social networks.

In the context of social networks, the notion of preferential attachement 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 2.1 (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 attachement iff, for any node i, $p(d_i \geq n+1 \mid d_i \geq n, \mathcal{M}_e)$ increases with $n \in \mathbb{N}$;*

(2) Local Preferential Attachment*: we say that a probabilistic link prediction model $\mathcal{M}_e$ satisfies the local preferential attachement 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 $x \in [0, N-1]$. Furthermore, $d_{i,k}$ is defined as the expectation, 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 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 $\hat{\mathbf{F}}$ and $\hat{\mathbf{\Phi}}$, one has:

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

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:

$$0 \leq p(d_i \geq n+1 \mid d_i \geq n, \mathcal{M}_e) \leq (1 - c^{N-n}) \xrightarrow[n \to N]{} 0$$

which shows that $p(d_i \geq n+1 \mid d_i \geq n, \mathcal{M}_e)$ does not increase with n. We thus have the following property:

Proposition 2.1. *Both `ILFM` and `IMMSB` do not satisfy global preferential attachment.*

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

Proposition 2.2. *`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 (12) 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 to 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 model well 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.

3. 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 (23) for assessing whether empirical data behaves according to a power law (as mentioned before, power laws are the standard bursty distributions in social networks (18)). 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.

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.

Table 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 (24). 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 [§], 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 [¶], 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 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 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 (25) 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-vlaue equals to 1 for Network1 and Blogs, and 0 and 0.4 for Network2 and Manufacturing.

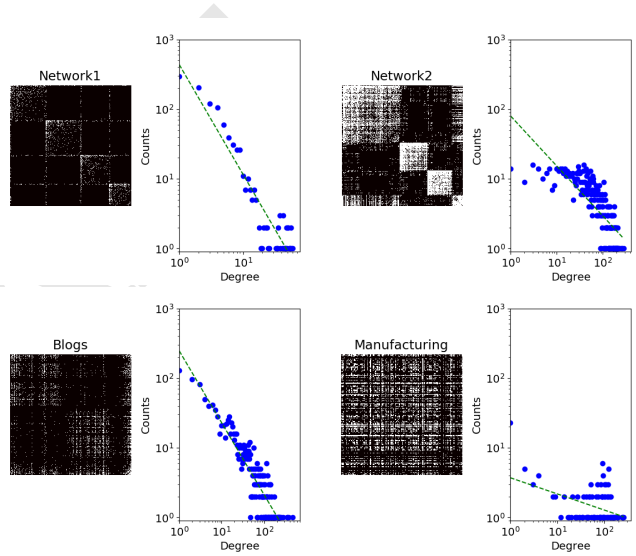

Fig. 1. Adjacency matrices (left) and global degree distributions (right) for the four training datasets. In the adjacency matrices, a white dot corresponds to a 1 and a block dot to a 0.

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 (12). 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.

Homophily. Figure 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

[§] moreno.ss.uci.edu/data.html#blogs

[¶] www.ii.pwr.edu.pl/ michalski/index.php?content=datasets#manufacturing

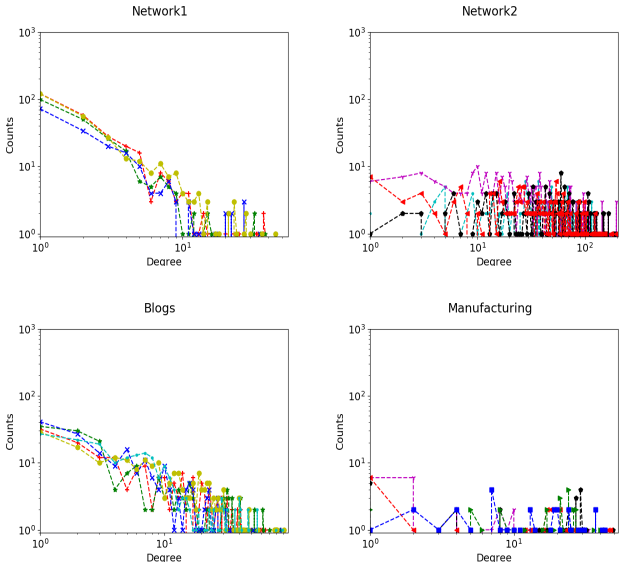

Fig. 2. Local degree distributions for the four training datasets. For Network1 and Network2 the classes come from ground-truth. For Blogs and Manufacturing, classes are obtained by Louvain algorithm.

for IMMSB (left) and ILFM (right). 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 Section 1 that state that both ILFM and IMMSB are homophilic for to the natural similarity but are not homophilic for the latent similarity.

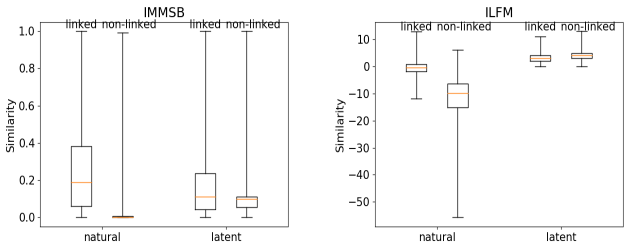

Fig. 3. Natural and latent similarities aggregated over all datasets and computed on linked and non-linked pairs of nodes for IMMSB (left) and ILFM (right).

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 equals to 0.9 for IMMSB and 1 for ILFM, and in Blogs with a p-value equals to 1 for both models; the property is not verified in Network2 and in Manufacturing, where p-values equal 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

Table 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

training datasets. Moreover, one can observe that, in the local case, IMMSB complies with the preferential attachment with p-values equals or close to 1 for the four networks, while ILFM obtained low p-values for the networks that were less locally bursty (respectively equal to 0 and 0.3 for Network2 and 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 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 2.2.

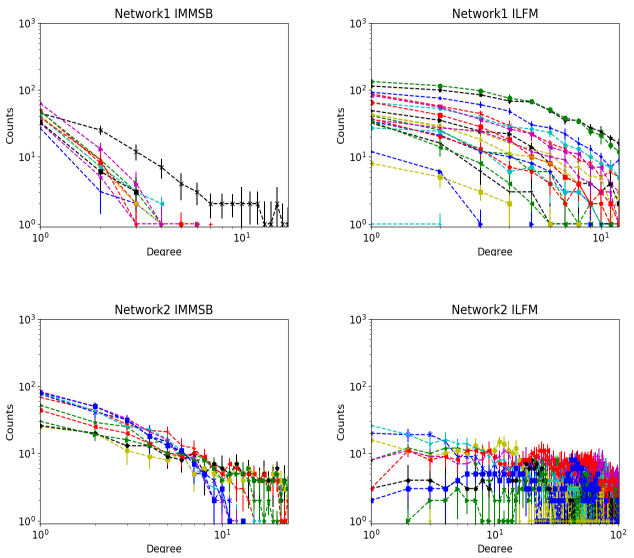

Fig. 4. Local degree distributions for Network1 (top row) and Network2 (bottom row) generated with fitted models IMMSB (first column) and ILFM (second column).

Lastly, Figure 5 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

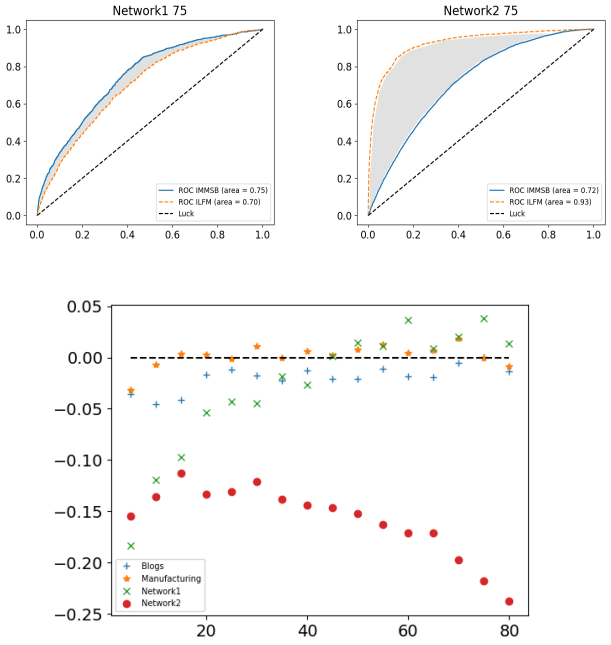

Fig. 5. Top: AUC-ROC curves for Network1 (left) and Network2 (right) with 20 percent of data used for learning. Bottom: Relative performance of IMMSB and ILFM according to the percentage of data used for testing, the rest being used for learning.

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 20 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.

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. Conclusion

We have studied here 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 attachement 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.

1. Liben-Nowell D, Kleinberg JM (2007) The link-prediction problem for social networks. *JASIST* pp. 1019–1031.
2. Hasan MA, Zaki MJ (2011) A survey of link prediction in social networks in *Social Network Data Analytics*, ed. Aggarwal CC. (Springer), pp. 243–275.
3. Meeds E, Ghahramani Z, Neal RM, Roweis ST (2006) Modeling dyadic data with binary latent factors in *Advances in Neural Information Processing Systems*. pp. 977–984.
4. Miller K, Jordan MI, Griffiths TL (2009) Nonparametric latent feature models for link prediction in *Advances in Neural Information Processing Systems*. pp. 1276–1284.
5. Airoldi EM, Blei DM, Fienberg SE, Xing EP (2009) Mixed membership stochastic blockmodels in *Advances in Neural Information Processing Systems*. pp. 33–40.
6. Koutsourelakis PS, Eliassi-Rad T (2008) Finding mixed-memberships in social networks. in *AAAI Spring Symposium: Social Information Processing*. pp. 48–53.
7. Fan X, Cao L, Xu D, Yi R (2015) Dynamic infinite mixed-membership stochastic blockmodel. *Neural Networks and Learning Systems, IEEE Transactions on* 26:2072–2085.
8. Newman M (2010) *Networks: An Introduction*. (Oxford University Press, Inc., New York, NY, USA).
9. Barabási A (2003) *Linked - how everything is connected to everything else and what it means for business, science, and everyday life*. (Plume).
10. Goldenberg A, Zheng AX, Fienberg SE, Airoldi EM (2010) A survey of statistical network models. *Foundations and Trends® in Machine Learning* 2(2):129–233.
11. Griffiths TL, Ghahramani Z (2011) The indian buffet process: An introduction and review. Vol. 12, pp. 1185–1224.
12. Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* 101(476):1566–1581.
13. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual review of sociology* pp. 415–444.
14. Lazarsfeld PF, Merton RK, , et al. (1954) Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society* 18(1):18–66.
15. 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. pp. 601–610.
16. Gopalan PK, Blei DM (2013) Efficient discovery of overlapping communities in massive networks. No. 36, pp. 14534–14539.
17. Leskovec J, Backstrom L, Kumar R, Tomkins A (2008) Microscopic evolution of social networks in *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*. pp. 462–470.
18. Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* (5439):509–512.
19. Barabási AL (2011) *Bursts: The Hidden Patterns Behind Everything We Do, from Your E-mail to Bloody Crusades*. (PLUME, Penguin Book, USA).
20. Church KW, Gale WA (1995) Poisson mixtures. *Natural Language Engineering* 1(02):163–190.
21. Clinchant S, Gaussier E (2008) The BNB distribution for text modeling in *Advances in Information Retrieval*. pp. 150–161.
22. 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.
23. Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. *SIAM review* 51(4):661–703.
24. Largeron C, Mougel PN, Rabbany R, Zaïane OR (2015) Generating attributed networks with communities. *PLoS ONE* 10(4):e0122777.
25. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E (2008) Fast unfolding of community in large networks. *Journal of statistical mechanics: theory and experiment* 10:P10008.