A study of stochastic mixed membership models for link prediction in social networks

—

Dulac Adrien

Christine Largeron – Eric Gaussier

LIG / AMA

Juillet 2017

COMMUNAUTÉS DE RECHERCHE ACADÉMIQUE Auvergne-Rhône-Alpes

T.I.C. ET USAGES INFORMATIQUES INNOVANTS

UNIVERSITÉ Grenoble Alpes

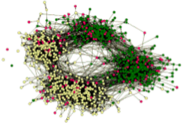
La Région Auvergne-Rhône-Alpes

Table of contents

Summary

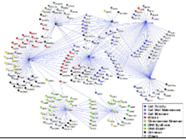
Complex Networks

Friendship Network

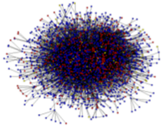


Business ties in US biotech-industry

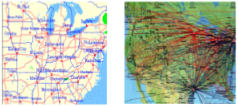
Genetic interaction network



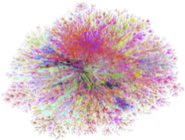
Protein-Protein Interaction Networks



Transportation Networks



Internet



Social Networks Properties

Important properties observed in social networks

Property	Measure
Small World	Diameter (Millgram 1978, Leskovec 2008)
Community	Modularity, Clustering Coefficient (Newman 2006)
Homophily	Homophily Coefficient (McPherson 2001)
Preferential Attachment / Burstiness	Degree Distribution (Barabási 1999)
Sparsity	Density (Haldous, Hoover 1981)

Statement

We focus on two models in the class of Mixed-Membership Models:

- IMMSB (infinite Mixed-Membership Stochastic Blockmodel)
- ILFM (infinite Latent Feature Model)

Do those models comply with preferential attachment? homophily?

Statement

We focus on two models in the class of Mixed-Membership Models:

- IMMSB (infinite Mixed-Membership Stochastic Blockmodel)
- ILFM (infinite Latent Feature Model)

Do those models comply with preferential attachment? homophily?

Claims

	Preferential Attachment		
	global	local	homophily
ILFM	No	No	depends on similarity
IMMSB	No	Yes	depends on similarity

Summary

Probabilistic Model

The Joint Distribution

$P(Y, \Theta \mid \sigma)$: How observations of a natural phenomena arise

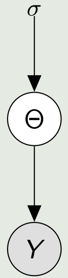
Likelihood

$P(Y \mid \Theta)$: How observations depend on the latent variables

Prior

$P(\Theta \mid \sigma)$: Describe the latent variables of the model

Graphical Model



Probabilistic Model

The Joint Distribution

$P(Y, \Theta \mid \sigma)$: How observations of a natural phenomena arise

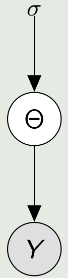
Likelihood

$P(Y \mid \Theta)$: How observations depend on the latent variables

Prior

$P(\Theta \mid \sigma)$: Describe the latent variables of the model

Graphical Model



The generative process

$$\Theta \sim P(\Theta \mid \sigma)$$
$$y \sim P(y \mid \Theta)$$

Inference / Optimisation

Posterior Distribution: $P(\Theta \mid Y) = \frac{P(Y|\Theta)P(\Theta)}{P(Y)}$

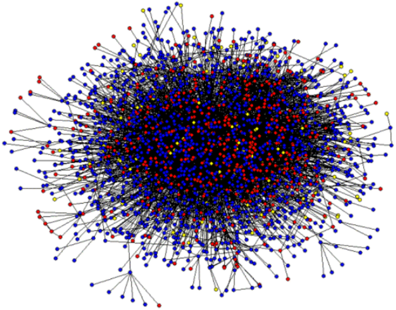
Social Network to Graph

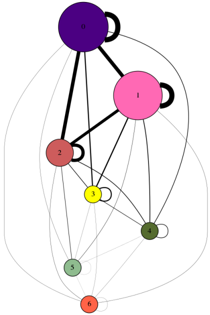
A graph : $G = (V, E)$

- V: set of vertices, drawn as nodes,
- E: set of edges, drawn as lines connecting vertices,

containing N nodes : $|V| = N$

Define a adjacency matrix : $Y = (y_{ij}) \quad i, j \in V^2, \left| \begin{array}{ll} y_{ij}=1 & \text{if} \quad (i,j) \in E \\ y_{ij}=0 & \text{otherwise} \end{array} \right.$





Unsupervised Learning

Learn a "good" representation of the data in order to do:

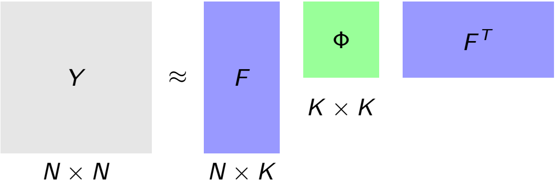
- Links Prediction
- Clustering and Visualization

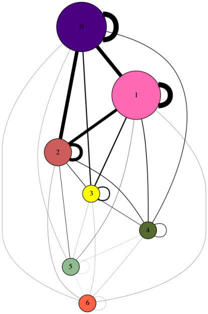
Unsupervised Learning

Learn a "good" representation of the data in order to do:

- Links Prediction
- Clustering and Visualization

Approximate the adjacency matrix : $Y \approx F \Phi F^T$





A latent feature matrix : $F \in \mathcal{F}^{N \times K}$ – Local variable for each node
A latent weights matrix : $\Phi \in \mathcal{P}^{K \times K}$ – Global variable shared by all nodes
Assume K the dimension of the model (dimension of latent variables)

Random Graph model

Given a Graph $G = (V, E)$, find F and Φ such that:

$$Y \sim P(Y|\ F, \Phi) = F\Phi F^T$$

F : a latent feature matrix (describe the membership of nodes)

Φ : a latent weight matrix (describe the correlation between class)

Random Graph model

Given a Graph $G = (V, E)$, find F and Φ such that:

$$Y \sim P(Y|\ F, \Phi) = F\Phi F^T$$

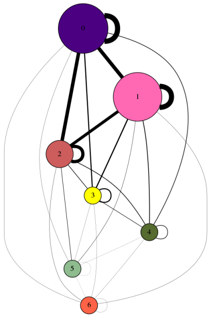
F : a latent feature matrix (describe the membership of nodes)

Φ : a latent weight matrix (describe the correlation between class)

Likelihood

$$P(y_{ij} = 1|F, \Phi) = f_i \Phi f_j^T$$

What prior over F and Φ ?



Random Graph Model (cont.)

The Stochastic Blockmodel[1]

- We have K blocks and,
- Any node i belongs to one block c:
- Each weights ϕ_{kl} , $k, l \in (0, .., K-1)^2$ encode the strength between two blocks.

Arbitrary Example:

Block membership for node 4 and 7 (K=3 and N=9)

$f_4 = \boxed{0 \,|\, 1 \,|\, 0}$ $\quad f_7 = \boxed{0 \,|\, 0 \,|\, 1}$ $\quad \phi_{12} = 0.3$

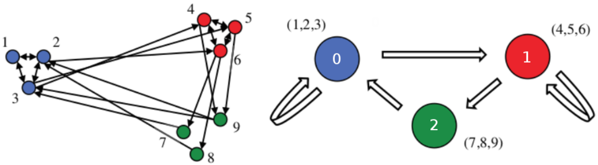


Figure: Stochastic Blockmodel

[1] goldenberg2010survey

Mixed Membership Models

Relax the single membership assumption (soft clustering).

Binary features

$f_i =$

1	0	1	1	0

Also relax the fixed number of features K with an Indian Buffet Process (IBP).

or

Proportion features

$f_i =$

.5	.2	.1	.1	.1

$\sum f_i = 1$

Also relax the fixed number of features K with a (Hierarchical) Dirichlet Process (DP).

Random Graph Model (cont.)

Infinite Mixed Membership Stochastic Model (IMMSB)[2]

$\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$

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

$\quad \boldsymbol{f}_i \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta})$

For each $(m, n) \in \{1, .., K\}^2$

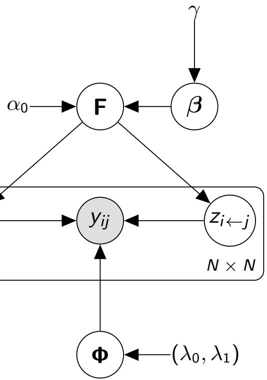
$\quad \phi_{mn} \sim \mathrm{Beta}(\lambda_0, \lambda_1)$

For each $(i, j) \in V^2$

$\quad z_{i\rightarrow j} \sim \mathrm{Mult}(\boldsymbol{f}_i)$

$\quad z_{i\leftarrow j} \sim \mathrm{Mult}(\boldsymbol{f}_j)$

$\quad y_{ij} \sim \mathrm{Bern}(z_{i\rightarrow j} \Phi z_{i\leftarrow j}^T)$



[2]**IMMSB**.

Infinite Latent Feature Model (ILFM)[3]

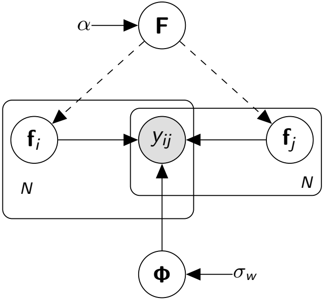
$F \sim \mathrm{IBP}(\alpha)$

For each $(m, n) \in \{1, .., K\}^2$

$\phi_{mn} \sim \mathrm{Gaussian}(0, \sigma_w)$

For each $(i, j) \in V^2$

$y_{ij} \sim \mathrm{Bern}(\boldsymbol{f}_i \Phi \boldsymbol{f}_j^T)$



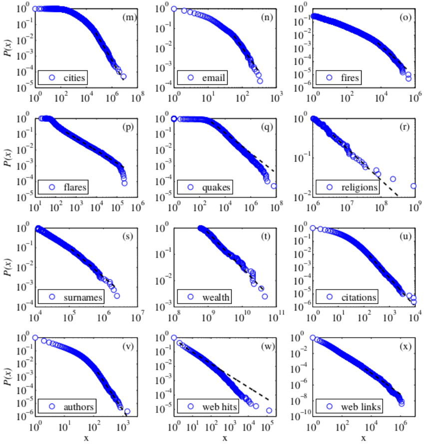
[3] **ILFRM**.

Summary

Preferential Attachment

$$\Delta_n P(d_i \geq n+1 | d_i \geq n) > 0$$

Figure: power law distribution in empirical data



4

Characterisation of Burstiness

Three Levels.

Definition

Global Preferential attachment

$$\Delta_n P(d_i \geq n+1 | d_i \geq n) > 0$$

Definition

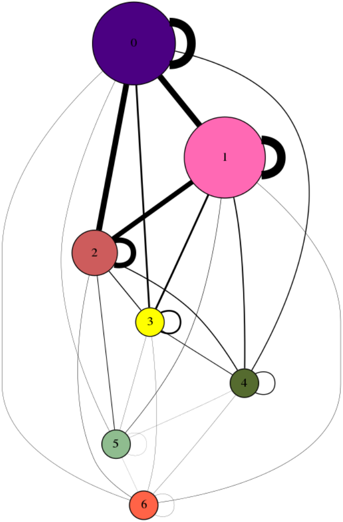
Local Preferential attachment

$$\Delta_n P(d_{ic} \geq n+1 | d_{ic} \geq n) > 0$$

Definition

Feature Burstiness

$$\Delta_n P(\boldsymbol{f}_k \geq n+1 | \boldsymbol{f}_k \geq n) > 0$$



Summary

Theoretical Results

Summary of our theoretical results.[5]

Theoretical results

	Preferential Attachment		
	global	local	homophily
ILFM	No	No	depends on similarity
IMMSB	No	Yes	depends on similarity

[5] **A study of stochastic mixed membership models for link prediction in social networks: Under Revisio**

Empirical Results (I)

Figure: Synthetic Datasets.

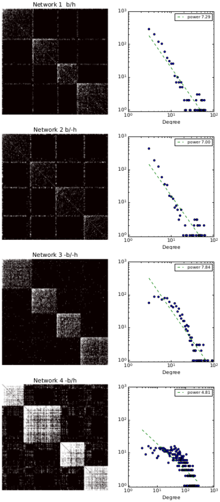
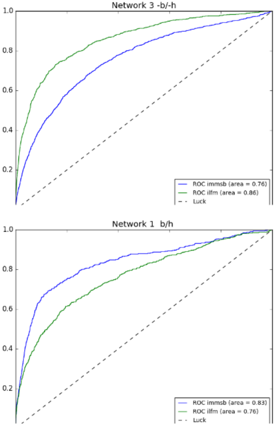
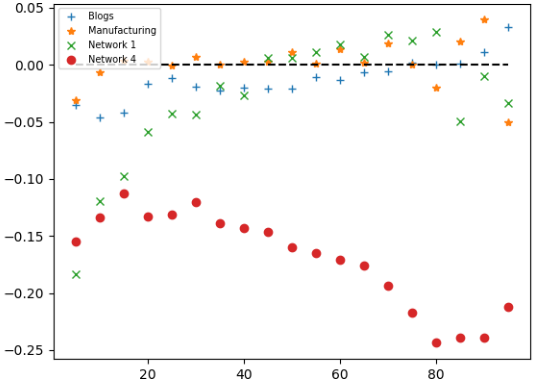


Figure: Prediction Performance, AUC curves.



Empirical Results (II)

Figure: Comparison performance of IMMSB and ILFM based on the size of the training set.



Summary

Limitation and future work

Representation Theorem for exchangeable graph[6].

Aldous-Hoover Theorem

An array $(Y_{ij})_{ij \in \mathbb{N}}$ is jointly exchangeable if and only if :

$$Y_{ij} \sim F(U_i, U_j, U_{ij}) \tag{1}$$

- With $F : [0,1]^3 \to [0,1]$ a random measure,
- $(U_i)_{i \in \mathbb{N}}$ and $(U_{ij \in \mathbb{N}})$ i.i.d Uniform[0,1] random variables.

A corollary of this theorem is that random exchangeable graph are either **empty or dense**.

➔ What network structure and invariance to comply with the Global preferential attachment ?

[6] **orbanz2015bayesian**.

Conclusion

Wrap up

- Formal Definition of social networks properties [Homophily and Preferential Attachments].
- Study of compliance of Mixed Membership Models with the properties.
- Empirical implications on prediction performances.

Future work

- Extend to weighted networks (and scalable inference).
- Extend to temporal networks.

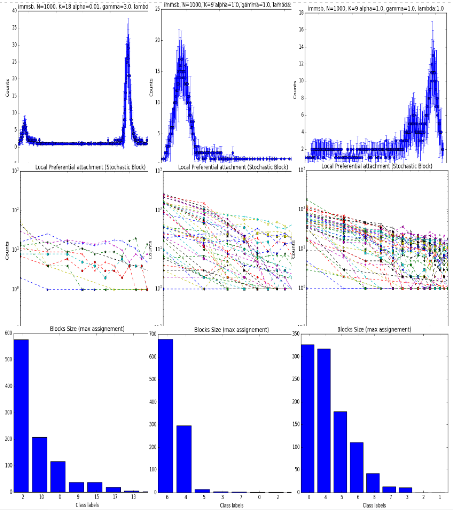
Bottom

Thank You

Characterisation of Burstiness (cont .)

Generative Process for IMMSB.

Figure: IMMSB Simulation.



Characterisation of Burstiness (cont .)

Is there a proof ?

$$P(d_i = n|\ \alpha) \quad ?$$

Problem

- d_i random variable ?
- $P(\boldsymbol{Y}_i \mid \alpha, \lambda) = \int_F \int_\Phi p(f_i|\alpha) p(\Phi|\lambda) \prod_{j \neq i} p(f_j|\alpha) p(y_{ij} \mid f_i, f_j, \Phi)\ dFd\Phi$

Intractable...