Clustering of nodes in social networks

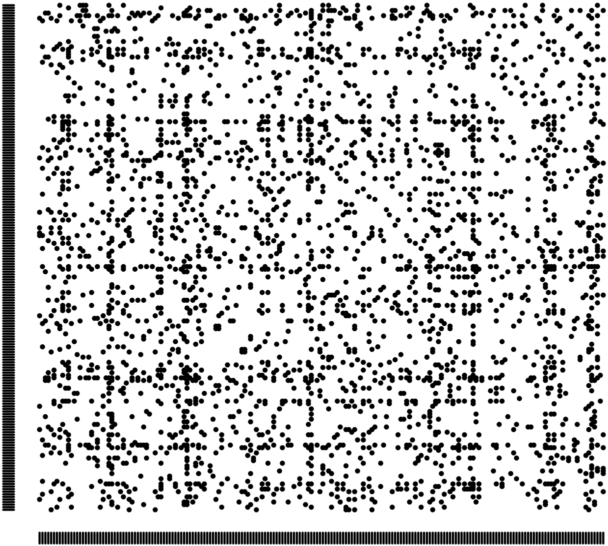
Pierre Latouche

Université Paris Descartes
Laboratoire MAP5
http://www.math-info.univ-paris5.fr/~platouch

Outline

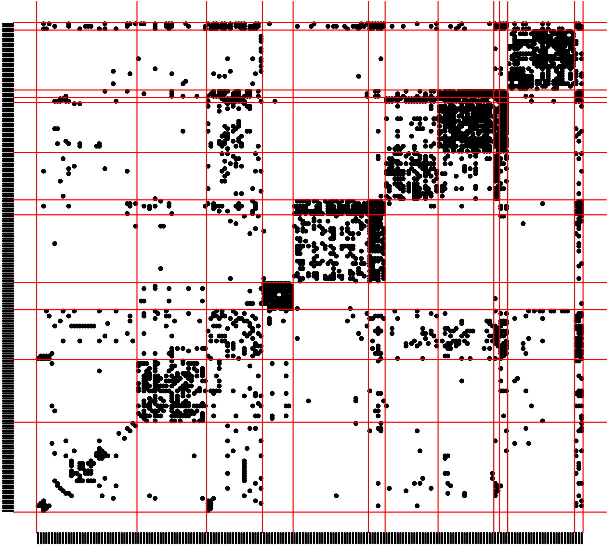
Outline

Dot plot



Network of blogs (Zanghi et al., 2008)

Clustering



Network of blogs (Zanghi et al., 2008)

Clustering

- Goal : to build a partition of the nodes
- A partition of K clusters is a family of sets P_k such that
  - $P_k \cap P_l = \emptyset, \forall k \neq l$
  - $\cup_{k=1}^{K} P_k = V$
- How ?
  - Nodes in a cluster should be *similar*
  - Nodes in different clusters should *disimilar*
- How to measure the similarity between nodes ?

- *NP* hard problem
- Looking for communities = looking for very specific clusters
- More edges between nodes of the same cluster
- Based on a criterion called *modularity*

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - P_{ij}) \delta(C_i, C_j)$$

- P_{ij} : expectation of the number of edges between i and j, under a null model
- $\delta(C_i, C_j) = 1$ if i and j are in the same community ($C_i = C_j$)

Introduction

Types of networks : (→ development of statistical approaches)

- Binary + static edges
- Discrete / continuous / categorical / ...
- Covariates on vertices / edges
- Dynamic edges :
  - Continous time → point processes
  - Discrete time → Markov,...

Types of clusters : (→ development of statistical approaches)

- Communities (transitivity)
- Heterogeneous clusters
- Partitions, overlapping clusters, hierarchy

Introduction

Essentially, two starting points :

- The latent position model [HRH02]
- The stochastic block model [WW87, NS01]

Outline

Erdös-Rényi model

- Two nodes connect with probability μ : $X_{ij} \sim \mathcal{B}(\mu)$
- So $D_i = \sum_{j=1}^{n} X_{ij}$ is (approximately) drawn from a Poisson distribution
  - $D_i \sim \mathcal{B}(n-1, \mu) \approx \mathcal{P}(n\mu)$
  - $\forall k, \mathbb{P}(D_i = k) \approx e^{-np}(n\mu)^k / k! \not\propto k^{-a}$
  - Not a power law !
- AND : homogenous model !
- A lot of developments on theoretical aspects
- Not adapted to real networks

Outline

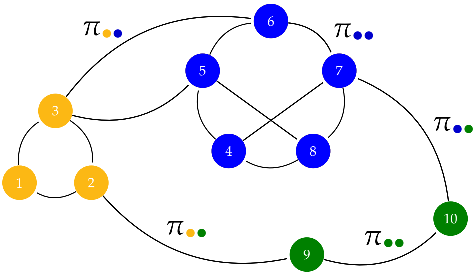
Stochastic Block Model (SBM) [WW87, NS01]

- Z_i independent hidden variables :
  - $Z_i \sim \mathcal{M}\big(1, \alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)\big)$
  - $Z_{ik} = 1$: vertex i belongs to class k
- $X|Z$ edges drawn independently :

$$X_{ij}|\{Z_{ik}Z_{jl} = 1\} \sim \mathcal{B}(\pi_{kl})$$

- A mixture model for graphs :

$$X_{ij} \sim \sum_{k=1}^{K} \sum_{l=1}^{K} \alpha_k \alpha_l \mathcal{B}(\pi_{kl})$$



π
6
π
5
7
3
π
4
8
1
2
10
π
9
π

Maximum likelihood estimation

- **Log-likelihoods of the model** :
  - Observed-data : $\log p(X|\alpha,\pi) = \log\left\{\sum_Z p(X,Z|\alpha,\pi)\right\}$
    $\hookrightarrow K^N$ terms
- Expectation Maximization (EM) algorithm requires the knowledge of $p(Z|X,\alpha,\pi)$

Problem
$p(Z|X,\alpha,\pi)$ is not tractable (no conditional independence)

Variational EM
Daudin et al. [DPR08]

Maximum likelihood estimation

- **Log-likelihoods of the model** :
  - Observed-data : $\log p(X|\alpha,\pi) = \log\left\{\sum_Z p(X,Z|\alpha,\pi)\right\}$
    $\hookrightarrow$ K^N terms
- Expectation Maximization (EM) algorithm requires the knowledge of $p(Z|X,\alpha,\pi)$

Problem

$p(Z|X,\alpha,\pi)$ is not tractable (no conditional independence)

Variational EM

Daudin et al. [DPR08]

Maximum likelihood estimation

- **Log-likelihoods of the model** :
  - Observed-data : $\log p(X|\alpha, \pi) = \log\left\{\sum_Z p(X, Z|\alpha, \pi)\right\}$
    $\hookrightarrow K^N$ terms
- Expectation Maximization (EM) algorithm requires the knowledge of $p(Z|X, \alpha, \pi)$

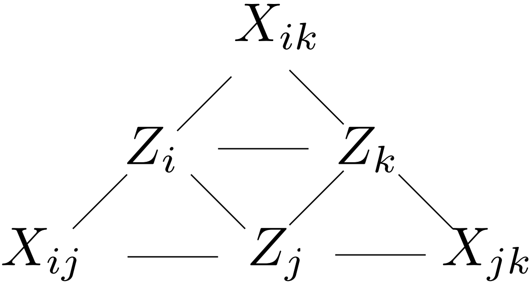
Problem

$p(Z|X, \alpha, \pi)$ is not tractable (no conditional independence)

Variational EM

Daudin et al. [DPR08]

Graphical model and moral graph



Moral graph of SBM

Model selection

Criteria

Since $\log p(X|\alpha, \pi)$ is not tractable, we *cannot* rely on :

- $AIC = \log p(X|\hat{\alpha}, \hat{\pi}) - M$
- $BIC = \log p(X|\hat{\alpha}, \hat{\pi}) - \frac{M}{2} \log \frac{N(N-1)}{2}$

ICL

Biernacki et al. [BCG00] $\hookrightarrow$ Daudin et al. [DPR08]

Variational Bayes EM $\hookrightarrow$ *ILvb*

Latouche et al. [LBA12]

Others

McDaid et al. [MDMNH13]

Model selection

Criteria

Since $\log p(X|\alpha, \pi)$ is not tractable, we *cannot* rely on :

- $AIC = \log p(X|\hat{\alpha}, \hat{\pi}) - M$
- $BIC = \log p(X|\hat{\alpha}, \hat{\pi}) - \frac{M}{2} \log \frac{N(N-1)}{2}$

ICL

Biernacki et al. [BCG00] $\hookrightarrow$ Daudin et al. [DPR08]

Variational Bayes EM $\hookrightarrow$ *ILvb*

Latouche et al. [LBA12]

Others

McDaid et al. [MDMNH13]

Bayesian framework

- **Conjugate prior distributions** :
  - $p\left(\alpha | n^0 = \{n^0_1, \ldots, n^0_K\}\right) = \mathrm{Dir}(\alpha;\, n^0)$
  - $p\left(\pi | \eta^0 = (\eta^0_{kl}), \zeta^0 = (\zeta^0_{kl})\right) = \prod_{k \leq l} \mathrm{Beta}(\pi_{kl};\, \eta^0_{kl}, \zeta^0_{kl})$
- **Non informative Jeffreys prior** :
  - $n^0_k = 1/2$
  - $\eta^0_{kl} = \zeta^0_{kl} = 1/2$

Variational Bayes EM [LBA09]

- $p(Z, \alpha, \pi | X)$ not tractable

Decomposition

$$\log p(X) = \mathcal{L}(q) + \mathrm{KL}\left(q(\cdot) \,\|\, p(\cdot|X)\right)$$

where

$$\mathcal{L}(q) = \sum_Z \int \int q(Z, \alpha, \pi) \log \left\{ \frac{p(X, Z, \alpha, \pi)}{q(Z, \alpha, \pi)} \right\} d\alpha d\pi$$

Factorization

$$q(Z, \alpha, \pi) = q(\alpha) q(\pi) q(Z) = q(\alpha) q(\pi) \prod_{i=1}^{N} q(Z_i)$$

Variational Bayes EM [LBA09]

E-step

- $q(Z_i) = \mathcal{M}(Z_i;\, 1, \boldsymbol{\tau_i} = \{\tau_{i1}, \ldots, \tau_{iK}\})$

M-step

- $q(\alpha) = \mathrm{Dir}(\alpha;\, n)$
- $q(\pi) = \prod_{k \leq l}^{K} \mathrm{Beta}(\pi_{kl};\, \eta_{kl}, \zeta_{kl})$

A new model selection criterion : ILvb [LBA12]

- $\log p(X|K) = \mathcal{L}(q) + \mathrm{KL}(...)$
- After convergence, use $\mathcal{L}(q)$ as an approximation of $\log p(X|K)$

ILvb

$$IL_{vb} = \log\left\{\frac{\Gamma(\sum_{k=1}^{K} n_k^0)\prod_{k=1}^{K}\Gamma(n_k)}{\Gamma(\sum_{k=1}^{K} n_k)\prod_{k=1}^{K}\Gamma(n_k^0)}\right\}
+ \sum_{k\leq l}^{K}\log\left\{\frac{\Gamma(\eta_{kl}^0+\zeta_{kl}^0)\Gamma(\eta_{kl})\Gamma(\zeta_{kl})}{\Gamma(\eta_{kl}+\zeta_{kl})\Gamma(\eta_{kl}^0)\Gamma(\zeta_{kl}^0)}\right\} - \sum_{i=1}^{N}\sum_{k=1}^{K}\tau_{ik}\log\tau_{ik}$$

Extensions and results

- Many extensions have been proposed for SBM
  - Overlapping clusters : MMSBM [ABFX08], OSBM [LBA11]
  - Covariates [ZVA10, MRV10]
  - Continuous, discrete, categorial edges [MRV10, JLB^{+}14, MR14]
  - ...
- Identifiability of SBM [AMR11]
- Consistency of variational approaches in SBM [CDP12, BCCZ13, MM15]

Outline

Latent position cluster model

- Z_i independent hidden variables :

$$Z_i \sim \sum_{k=1}^{K} \alpha_k \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2\, \mathbf{I}),$$

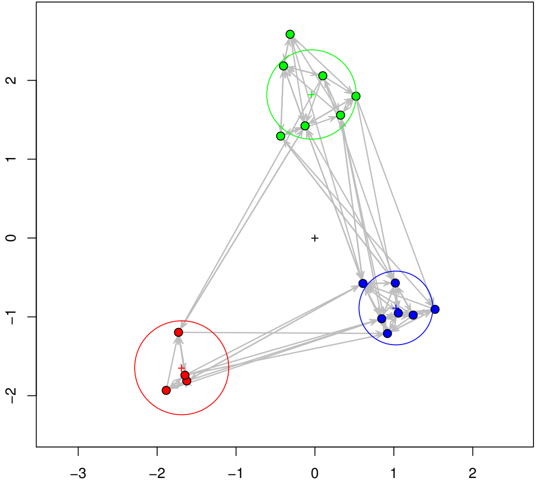
- $X|Z$ edges drawn independently :

$$X_{ij}|Z_i, Z_j, \mathbf{Y}_{ij} \sim \mathcal{B}\left(g(a_{Z_i,Z_j,\mathbf{Y}_{ij}})\right).$$

The function $g(x) = 1/(1 + e^{-x})$ is the logistic sigmoid function. Moreover $a_{Z_i,Z_j,\mathbf{Y}_{ij}}$ is given by :

$$a_{Z_i,Z_j,\mathbf{Y}_{ij}} = \mathbf{Y}_{ij}^{\intercal}\, \boldsymbol{\beta}_0 - \beta_1 |Z_i - Z_j|, \tag{1}$$

where $\boldsymbol{\beta}_0$ as the same dimensionality as $\mathbf{Y}_{ij}$ and β_1 is a scalar



An analysis with LPCM

Outline

What about ERGM ?

- Used in many applications
- Less and less by statisticians

Outline

STBM : Context and notations

We are interesting in clustering the M nodes of a network into Q groups :

- the network is represented by its $M \times M$ adjacency matrix A :

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between i and j} \\ 0 & \text{otherwise} \end{cases}$$

- if $A_{ij} = 1$, the textual edge is characterized by a set of D_{ij} documents, where each document W_{ij}^d is made of N_{ij}^d words :

$$\mathbf{W}_{ij} = (W_{ij}^1, ..., W_{ij}^d, ..., W_{ij}^{D_{ij}}),$$

$$\text{where} \quad W_{ij}^d = (W_{ij}^{d1}, ..., W_{ij}^{dn}, ..., W_{ij}^{dN_{ij}^d}),$$

- in practice, the user has to provide a list of textual edges : Alexis ; Arthur ; "I am very happy to try Linkage" Louis ; Nathan ; "Do you know that Romain tried Linkage ?"

STBM : Context and notations

We are interesting in clustering the M nodes of a network into Q groups :

- the network is represented by its $M \times M$ adjacency matrix A :
$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between i and j} \\ 0 & \text{otherwise} \end{cases}$$
- if $A_{ij} = 1$, the textual edge is characterized by a set of D_{ij} documents, where each document W_{ij}^d is made of N_{ij}^d words :
$$\mathbf{W}_{ij} = (W_{ij}^1, ..., W_{ij}^d, ..., W_{ij}^{D_{ij}}),$$
$$\text{where} \quad W_{ij}^d = (W_{ij}^{d1}, ..., W_{ij}^{dn}, ..., W_{ij}^{dN_{ij}^d}),$$
- in practice, the user has to provide a list of textual edges :
Alexis ; Arthur ; "I am very happy to try Linkage"
Louis ; Nathan ; "Do you know that Romain tried Linkage ?"

STBM : A quick reminder on Statistics...

Let us remind that :

- The Bernouilli distribution acts, for a binary random variable $X \in \{0, 1\}$, as follows :

  $$X \sim \mathcal{B}(\pi = 0.7) \rightarrow \{0, 1, 1, 1, 0, 1, 1, 0, 1, 1, ...\},$$

  where π is the probability of success.

- The Multinomial distribution acts, for a categorical random variable $X \in \{1, ..., Q\}$, as follows :

  $$X \sim \mathcal{M}(\rho = (0.2, 0.3, 0.5)) \rightarrow \{2, 1, 3, 3, 2, 3, 2, 1, 3, ...\},$$

  where ρ_q is the probability of getting the value q.

STBM : Modeling of the edges

Let us assume that edges are generated as follows :

- each node i is associated with an (unobserved) group among Q such that :

$$Y_i \sim \mathcal{M}(\rho),$$

where $\rho \in [0,1]^Q$ is the vector of group proportions,

- the presence of an edge A_{ij} between i and j is drawn according to :

$$A_{ij}|Y_{iq}Y_{jr} = 1 \sim \mathcal{B}(\pi_{qr}),$$

where $\pi_{qr} \in [0,1]$ is the connection probability between clusters q and r.

STBM : Modeling of the edges

Let us assume that edges are generated as follows :

- each node i is associated with an (unobserved) group among Q such that :

$$Y_i \sim \mathcal{M}(\rho),$$

  where $\rho \in [0,1]^Q$ is the vector of group proportions,

- the presence of an edge A_{ij} between i and j is drawn according to :

$$A_{ij}|Y_{iq}Y_{jr} = 1 \sim \mathcal{B}(\pi_{qr}),$$

  where $\pi_{qr} \in [0,1]$ is the connection probability between clusters q and r.

STBM : Modeling of the documents

The generative model for the documents is as follows :

- the nth word W_{ij}^{dn} of documents d in W_{ij} is then associated to a latent topic vector Z_{ij}^{dn} according to :

$$Z_{ij}^{dn} | \left\{ A_{ij} Y_{iq} Y_{jr} = 1, \theta \right\} \sim \mathcal{M}\left(\theta_{qr} \right),$$

  where $\theta_{qr} = (\theta_{qrk})_k$ is the vector of topic proportions for the pair (q, r).

- then, given Z_{ij}^{dn}, the word W_{ij}^{dn} is assumed to be drawn from a multinomial distribution :

$$W_{ij}^{dn} | Z_{ij}^{dnk} = 1 \sim \mathcal{M}\left(1, \beta_k = (\beta_{k1}, \ldots, \beta_{kV})\right),$$

  where V is the vocabulary size.

STBM : Modeling of the documents

The generative model for the documents is as follows :

- the nth word W_{ij}^{dn} of documents d in W_{ij} is then associated to a latent topic vector Z_{ij}^{dn} according to :

  $$Z_{ij}^{dn} | \left\{ A_{ij} Y_{iq} Y_{jr} = 1, \theta \right\} \sim \mathcal{M}\left(\theta_{qr} \right),$$

  where $\theta_{qr} = (\theta_{qrk})_k$ is the vector of topic proportions for the pair (q, r).

- then, given Z_{ij}^{dn}, the word W_{ij}^{dn} is assumed to be drawn from a multinomial distribution :

  $$W_{ij}^{dn} | Z_{ij}^{dnk} = 1 \sim \mathcal{M}\left(1, \beta_k = (\beta_{k1}, \ldots, \beta_{kV})\right),$$

  where V is the vocabulary size.

STBM at a glance...

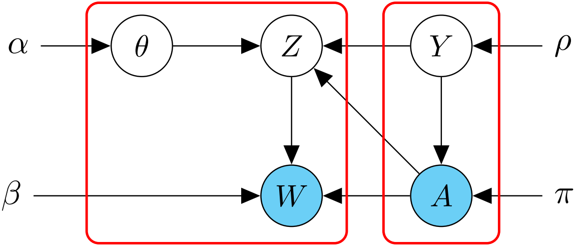


FIGURE – The stochastic topic block model.

Inference and model selection

To estimate the model parameters :

- we proposed an algorithm, that we called a C-VEM algorithm,
- which iteratively estimates the set of model parameters.

Model selection :

- a key point of the methodology is the ability to automatically select the most appropriate values for Q and K,
- to this end, we derived an ICL criterion which identify the best couple (Q^*, K^*) for the data at hand.

Inference and model selection

To estimate the model parameters :

- we proposed an algorithm, that we called a C-VEM algorithm,
- which iteratively estimates the set of model parameters.

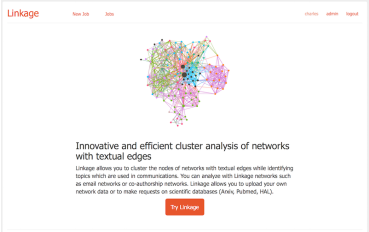
Model selection :

- a key point of the methodology is the ability to automatically select the most appropriate values for Q and K,
- to this end, we derived an ICL criterion which identify the best couple (Q^*, K^*) for the data at hand.

From research to innovation : the linkage project

Linkage is a maturation project :

- Linkage.fr is the result of a maturation project, supported by IDFInnov,
- to handle large networks, it needed a deep re-thinking of :
  - the data structure,
  - the inference algorithm,
  - the visualization tools,
- 6 months of engineering were also necessary to build up the web architecture, with the latest web technologies.

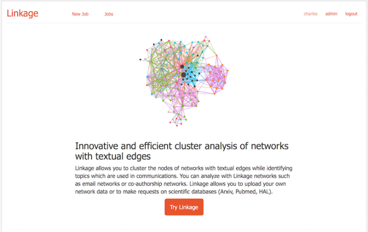


www.linkage.fr

From research to innovation : the linkage project

Linkage is a maturation project :

- Linkage.fr is the result of a maturation project, supported by IDFInnov,
- to handle large networks, it needed a deep re-thinking of :
  - the data structure,
  - the inference algorithm,
  - the visualization tools,
- 6 months of engineering were also necessary to build up the web architecture, with the latest web technologies.



www.linkage.fr

Innovation : the linkage project

Linkage.fr : a SAAS platform to prove the concept

- Linkage.fr aims to demonstrate the abilities of the technology in a few practical situations :
  - analysis of co-authorship networks,
  - analysis of Twitter networks,
  - analysis of Email networks,
- the platform also allows the user to provide their own data as CSV files.

The platform is already well known and used :

- > 800 registered users,
- several operational projects achieved or in progress.

Innovation : the linkage project

Linkage.fr : a SAAS platform to prove the concept

- Linkage.fr aims to demonstrate the abilities of the technology in a few practical situations :
  - analysis of co-authorship networks,
  - analysis of Twitter networks,
  - analysis of Email networks,
- the platform also allows the user to provide their own data as CSV files.

The platform is already well known and used :

- $>$ 800 registered users,
- several operational projects achieved or in progress.

Analysis of the Enron Emails

The Enron data set :

- all emails between 149 Enron employees,
- from 1999 to the bankrupt in late 2001,
- almost 253 000 emails in the whole data base.

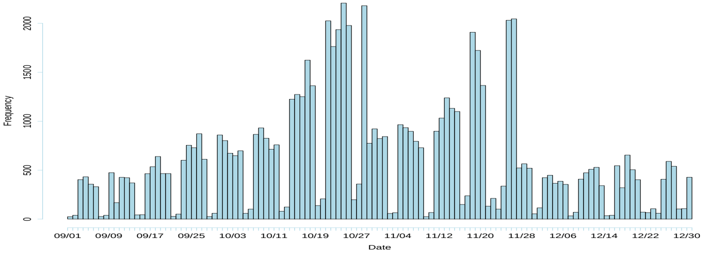


FIGURE – Temporal distribution of Enron emails.

Analysis of the Enron Emails

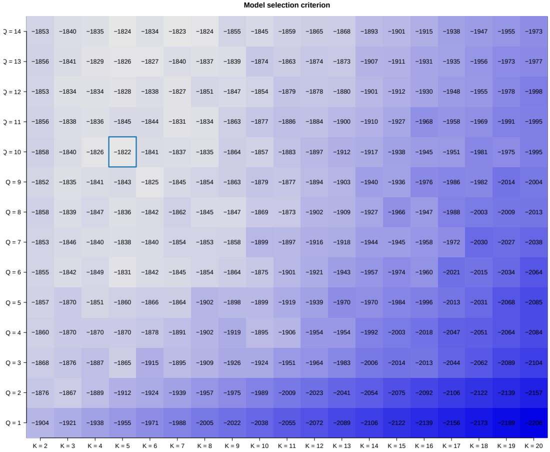


FIGURE – Model selection on the Enron network.

Analysis of the Enron Emails

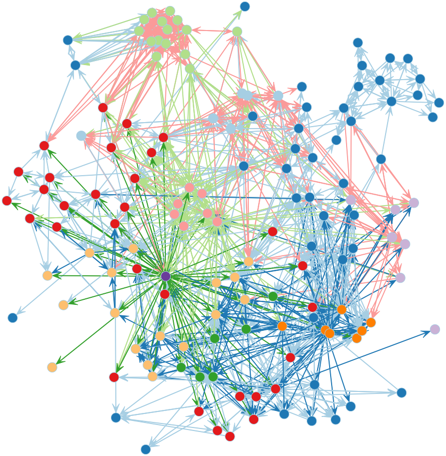


FIGURE – Clustering of the Enron network.

Analysis of the Enron Emails

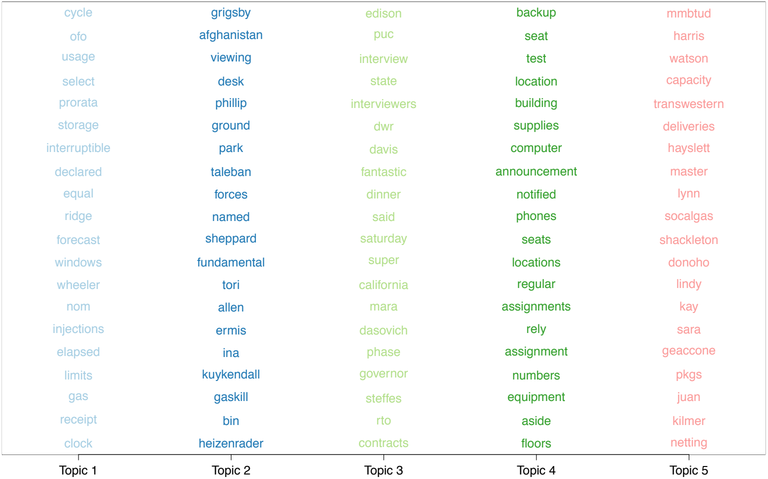


FIGURE – Most specific terms in the found topics for the Enron data.

Analysis of the Enron Emails

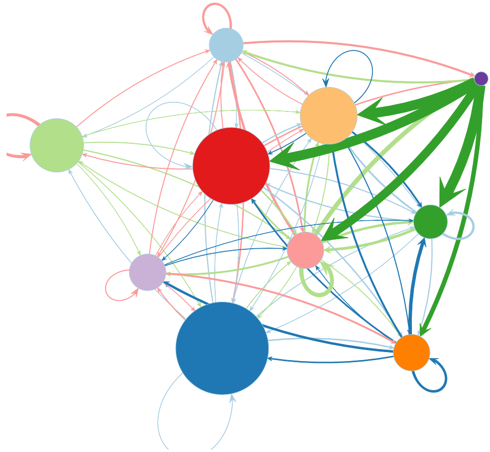


FIGURE – Meta-network for the Enron data set.

References I

- E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing, *Mixed membership stochastic blockmodels*, Journal of Machine Learning Research **9** (2008), 1981–2014.

- Elizabeth S Allman, Catherine Matias, and John A Rhodes, *Parameter identifiability in a class of random graph mixture models*, Journal of Statistical Planning and Inference **141** (2011), no. 5, 1719–1736.

- Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang, *Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels*, Ann. Statist. **41** (2013), no. 4, 1922–1943.

- C. Biernacki, G. Celeux, and G. Govaert, *Assessing a mixture model for clustering with the integrated completed likelihood*, IEEE Trans. Pattern Anal. Machine Intel **7** (2000), 719–725.

References II

- A. Celisse, J.J Daudin, and L. Pierre, *Consistency of maximum likelihood and variational estimators in stochastic block models*, Electronic Journal of Statistics **6** (2012), 1847–1899.
- J. Daudin, F. Picard, and S. Robin, *A mixture model for random graphs*, Statistics and Computing **18** (2008), 1–36.
- Peter D Hoff, Adrian E Raftery, and Mark S Handcock, *Latent space approaches to social network analysis*, Journal of the american Statistical association **97** (2002), no. 460, 1090–1098.
- Y. Jernite, P. Latouche, C. Bouveyron, P. Rivera, L. Jegou, and S. Lamassé, *The random subgraph model for the analysis of an acclesiastical network in merovingian gaul*, Annals of Applied Statistics **8** (2014), no. 1, 377–405.
- P. Latouche, E. Birmelé, and C. Ambroise, *Bayesian methods for graph clustering*, pp. 229–239, Springer, 2009.

References III

- P. Latouche, E Birmelé, and C. Ambroise, *Overlapping stochastic block models with application to the french political blogosphere*, Annals of Applied Statistics **5** (2011), no. 1, 309–336.
- P. Latouche, E. Birmelé, and C. Ambroise, *Variational bayes inference and complexity control for stochastic block models*, Statistical Modelling **12** (2012), no. 1, 93–115.
- A. Mc Daid, T.B. Murphy, Frieln N., and N.J. Hurley, *Improved bayesian inference for the stochastic block model with application to large networks*, Computational Statistics and Data Analysis **60** (2013), 12–31.
- Mahendra Mariadassou and Catherine Matias, *Convergence of the groups posterior distribution in latent or stochastic block models*, Bernoulli **21** (2015), no. 1, 537–573.

References IV

- C. Matias and S. Robin, *Modeling heterogenity in random graphs through latent space models : a selective review*, Esaim Prooceedings and Surveys **47** (2014), 55–74.
- Mahendra Mariadassou, Stéphane Robin, and Corinne Vacher, *Uncovering latent structure in valued graphs : a variational approach*, The Annals of Applied Statistics (2010), 715–742.
- K. Nowicki and T.A.B. Snijders, *Estimation and prediction for stochastic blockstructures*, Journal of the American Statistical Association **96** (2001), 1077–1087.
- Y.J. Wang and G.Y. Wong, *Stochastic blockmodels for directed graphs*, Journal of the American Statistical Association **82** (1987), 8–19.

References V

H. Zanghi, S. Volant, and C. Ambroise, *Clustering based on random graph model embedding vertex features*, Pattern Recognition Letters **31** (2010), no. 9, 830–836.