Introduction to Machine Learning: statistical analysis of networks (and texts)

P. Latouche

PR Université Paris Descartes & École Polytechnique
pierre.latouche@math.cnrs.fr

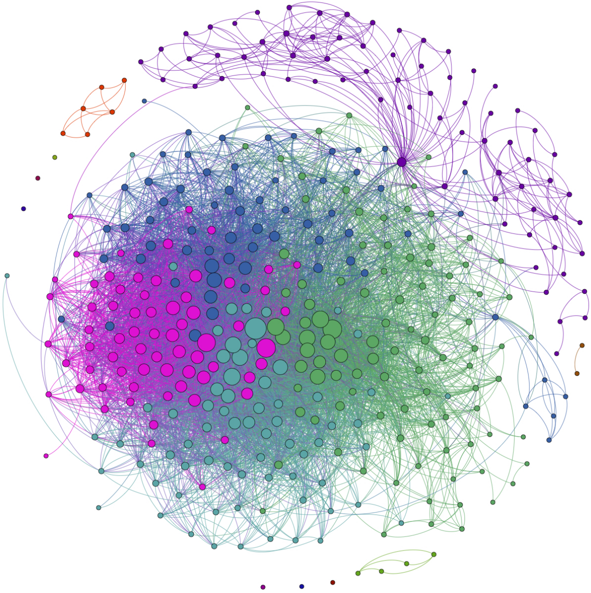
École Polytechnique

Part I

Introduction to graph analysis

Outline

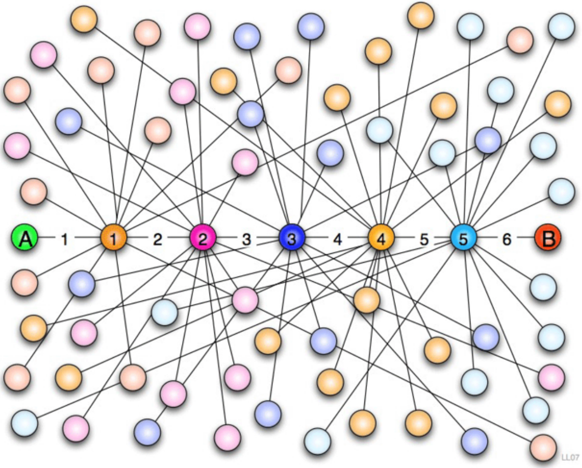
Outline



Example of a social network (Facebook) (PJ Lamberson, UCLA)

“Each user has drawn what specialists called a social graph, the map of all its relationships, which is about to become the ultimate footprint[...]”

F. Filloux, Facebook tisse sa toile, Le Monde Magazine



6 degrees of separation (D. Walker, Wikipedia)

"I read somewhere that everybody on this planet is separated by only six other people. Six degrees of separation between us and everyone else on this planet. The President of the United States, a gondolier in Venice, just fill in the names [...]"

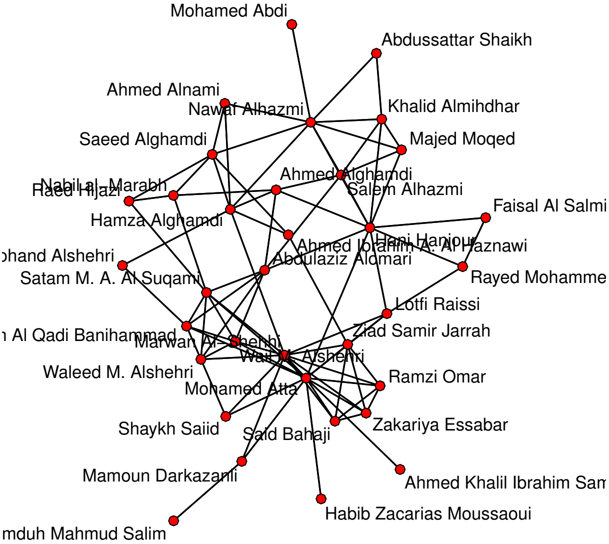
J. Guare, Six degrees of separation (1990)

6 degrees of separation

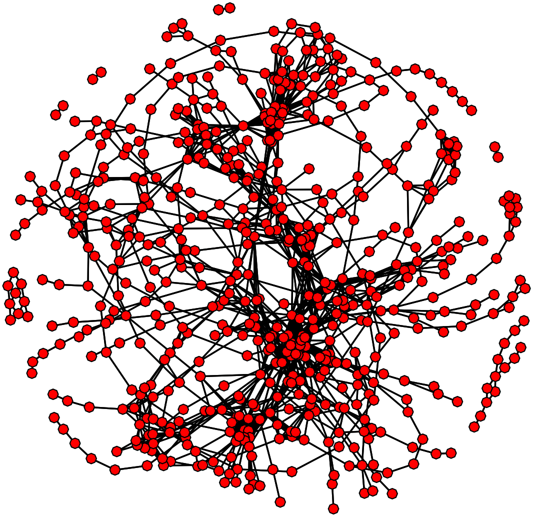
- Theory proposed by Karinthy (1929)
- A play written by John Guare (1990) : “Six degrees of separation”
- Film (1993)

6 degrees of separation

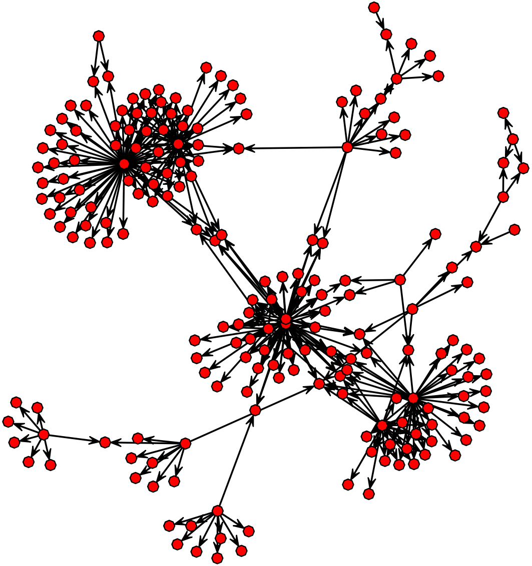
- On this question :
  - Gurevich (1961, PhD, MIT)
  - Kochen (Austrian mathematician)
  - Milgram (Psychologist, Harvard)
    - obedience to authority experiment
    - small world experiment
  - Watts (Columbia university) :
    - experiment : 48000 senders, 19 targets. Email = a package to be transmitted
    - On average : 6 degrees of separation
  - Leskovec and Horvitz : msn. Similar results
  - Twitter + Facebook : 3.4, ...



Terrorist network [Kre01]



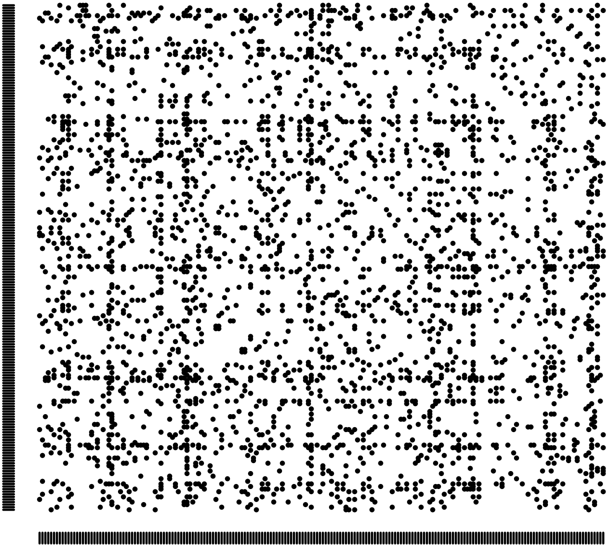
Metabolic network of (bacteria) Escherichia coli [LFS06]



Subset of the regulation network of yeast [MSOI+02]

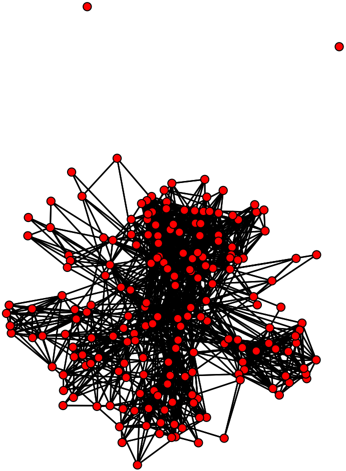
Outline

Dot plot representation



Network of blogs [ZAM08]

Fruchterman et Reingold [FR91]



Network of blogs [ZAM08]

Graph visualization

- As circle (circle)
- From the spectrum (eigen)
- Fruchterman et Reingold
- Hall
- Kamada et Kawai
- Multi dimensional scaling (mds)
- Use of eigen values
- Random

Kamada et Kawai [KK89] I

- Goal :
  - to avoid edge overlaps
  - to position nodes and edges uniformly in $\mathbb{R}^2$
- Dynamical system (related to Hooke's law)
  - nodes are particles
  - edges are springs
- Goal : to minimize the total elastic energy
- $E = \sum_{i<j} \frac{1}{2} k_{ij} (\|p_i - p_j\| - l_{ij})^2$
- l_{ij} proportional to the shortest distance between i and j
- Newton-Raphson

Fruchterman et Reingold [FR91] I

- Related to Kamada et Kawai (and Hooke's law)
- Simulation of the dynamical system
- Interaction between particles
- Forces induce movements
- Neighbors attract each others
- Repulsive forces applied on all particles
- $f_a(d) = d^2/k$ where d is the distance between two particles (attraction)
- $f_r(d) = -k^2/d$

Outline

- Oriented graph (directed) : (V, E) where $V = \{V_1, \ldots, V_n\}$ is the set of nodes and E is the set of edges. E is made of ordered pairs from $V \times V$
- Non oriented graph (undirected) : (V, E). The pairs in E are non ordered
- Valued graph: (V, E, f) where (V, E) is a graph and $f : E \rightarrow F$ is an application
- Degree of a node: number of edges of the node
- Path : a finite or infinite sequence of edges which connect a sequence of vertices all distinct from one another
- $G = (V, E)$ a graph. The subgraph associated to the subset A of V is the graph G_A defined by $G_A = (A, E \cap A \times A)$

Data structures

- Adjacency matrix $*$
- Incidence matrix
- Edge list $*$
- ...

Adjacency matrix

- $n \times n$ matrix : $X = (X_{ij})$ such that
- $X_{ij} = 1$ if i and j are linked by an edge, 0 otherwise
- Coding : $O(n^2)$

Edge list

- Each raw : an edge (i, j)
- Caution : directed and undirected case
- Coding : $O(m)$

Indicators

- Degree of a node $d_i = \sum_{j \neq i}^{n} X_{ij}$
  - Directed case: $d_i^{out} = \sum_{j \neq i}^{n} X_{ij}$, $d_i^{in} = \sum_{j \neq i}^{n} X_{ji}$
- Mean degree $\bar{d} = (1/n) \sum_{i=1}^{n} d_i$
- Graph density $\text{den}(G) = m/(n(n-1)/2)$
- Clustering coefficient $2e_i/d_i(d_i - 1)$ if $d_i \geq 2$

Part II

Random graph models

Outline

Erdös-Rényi model

Stochastic block model

Outline

Erdös-Rényi model

Stochastic block model

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

Erdös-Rényi model

Stochastic block model

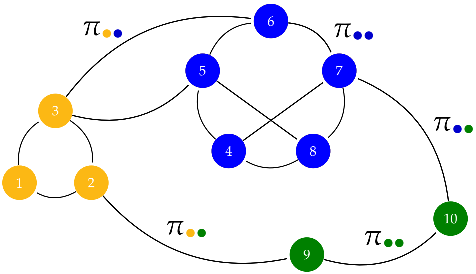
Stochastic Block Model (SBM) [WW87, NS01]

- Z_i independent hidden variables :
  - $Z_i \sim \mathcal{M}\Big(1, \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)\Big)$
  - $Z_{ik} = 1$: 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 :

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


6
5
7
3
4
8
1
2
9
10
π
π
π
π
π

Maximum likelihood estimation

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

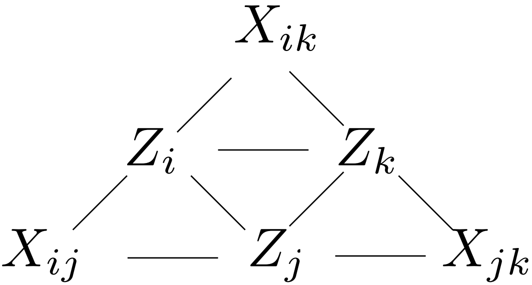
Problem

$p(Z|X, \alpha, \pi)$ 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** :
  - $p\left(\alpha | n^0 = \{n_1^0, \ldots, n_K^0\}\right) = \mathrm{Dir}(\alpha;\, n^0)$
  - $p\left(\pi | \eta^0 = (\eta_{kl}^0), \zeta^0 = (\zeta_{kl}^0)\right) = \prod_{k \leq l} \mathrm{Beta}(\pi_{kl};\, \eta_{kl}^0, \zeta_{kl}^0)$
- **Non informative Jeffreys prior** :
  - $n_k^0 = 1/2$
  - $\eta_{kl}^0 = \zeta_{kl}^0 = 1/2$

Variational Bayes EM [LBA09]

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

Decomposition

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

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

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

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]

Références 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.

Références 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.
- T.M.J Fruchterman and E.M. Reingold, *Graph drawing by force-directed placement*, Sotware Practice and Experience **21** (1991), 1129–1164.
- 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.
- T. Kamada and S. Kawai, *An algorithm for drawing general undirected graphs*, Information Processing Letters **31** (1989), 7–15.

Références III

- V. Krebs, *Unloaking terrotist networks*, Connections **24** (2001), no. 3.
- P. Latouche, E. Birmelé, and C. Ambroise, *Bayesian methods for graph clustering*, pp. 229–239, Springer, 2009.
- 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.
- V. Lacroix, C.G. Fernandes, and M.-F. Sagot, *Motif search in graphs:application to metabolic networks*, Transactions in Computational Biology and Bioinformatics **3** (2006), 360–368.

Références IV

- 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.
- 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.

Références V

- R. Milo, S. Shen-Orr, S. Itzkovitz, D. Kashtan, D. Chklovskii, and U. Alon, *Network motifs: simple building blocks of complex networks*, Science **298** (2002), 824–827.

- 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.

- H. Zanghi, C. Ambroise, and V. Miele, *Fast online graph clustering via erdös-rényi mixture*, Pattern Recognition **41** (2008), no. 12, 3592–3599.

- 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.

Références VI