PSEUDO-LIKELIHOOD METHODS FOR COMMUNITY DETECTION IN LARGE SPARSE NETWORKS

Authors: Amini et al. 2013
Presented by: Namdar Homayounfar
STA 4513

October 16, 2014

Introduction

Introduction

- **Network Data**: Observed edges between nodes, possibly accompanied by additional information on the notes/edges.
- **One of the Fundamental Analysis Tasks:** Detecting and modelling community structure within the network.
- **Literature:**
  1. Algorithmic approaches in physics: Greedy methods such as hierarchical clustering and algorithms based on optimizing a global criterion over all possible partitions, such as normalized cuts and modularity.
  2. Model based methods in statistics: Postulate and fit a probabilistic model for a network with communities - stochastic block models and its extensions.

Introduction — Stochastic Block Models (SBM)

Most commonly used and best studied model for community detection.

- Network with n nodes defined by an $n \times n$ adjacency matrix A — Symmetric and no self loops
- Model Assumption: True node labels $c = (c_1, \ldots, c_n) \in \{1, \ldots, K\}^n$ drawn independently from the multinomial distribution with parameter $\pi = (\pi_1, \ldots, \pi_K)$
- $\pi_i > 0$ for all i and K is the known number of communities
- Conditional on the labels, A_{ij} for $i < j$ independent Bernoulli:

$$\mathrm{E}\left[A_{ij} \middle| c\right] = P_{c_i c_j}$$

- $P = [P_{ab}]$ is a $K \times K$ symmetric matrix

Introduction — SBM

- **Inference Task:** Infer node labels c from A which involves estimating π and P
- SBM implies the same expected degree for all nodes within a community
- Excludes networks with "hub" nodes commonly encountered in practice
- Many extensions such as the *mixed membership models*
- This paper uses the extension to *degree-corrected block models*

Introduction — Degree Corrected SBM

- Remove the "hub" exclusion constraint by augmenting $\mathrm{E}\left[A_{ij}\middle|c\right] = P_{c_i c_j}$ with

$$\mathrm{E}\left[A_{ij}\middle|c\right] = \theta_i \theta_j P_{c_i c_j}$$

- θ_i's are node degree parameters satisfying an identifiability constraint.
- In [22], Bernoulli distribution for A_{ij} was replaced with Poisson — Ease of technical derivations, good approximation for a range of networks

Introduction — Fitting SBMs

- Bayesian framework:
  1. **MCMC:** work only for networks with a few hundred nodes
  2. **Variational Inference:** faster than Gibbs sampling in MCMC but do not scale to million of nodes
  3. **Belief Propagation** recently proposed in [14]: Comparable to the method in this presentation in theoretical complexity but slower in practice
- Non-Bayesian Framework:
  1. **Profile Likelihood:** For a given label assignment parameters can be estimated by plug-in, they can be profiled out and the resulting criterion can be maximized over all label assignments by greedy search — speed depends on the search method and the number of iterations — world for thousands but not millions of nodes
  2. **A method of Moments** approach: involves counting all occurrences of specific patterns in a graph which is computationally challenging.

Introduction — Consistency

- **Profile Likelihood:** Give consistent estimates of the labels under both the SBM and the degree-corrected version
  1. Strong Consistency — Probability of the the estimated label vector being equal to the truth converging to 1 — the average graph degree λ_n has to grow faster than $\log n$
  2. Weak Convergence — The fraction of misclassified nodes converging to 0 — only need $\lambda_n \to \infty$
- **Variational Methods and Belief Propagation :** Asymptotic behaviour analyzed for both the sparse $[\lambda_n = O(1)]$ and the dense $[\lambda_n \to \infty]$ regimes. Consistency is impossible to achieve unless in the dense regime. In sparse case, can only claim that the estimated labels are correlated with the truth better than random guessing, but not that they are consistent.

Contributions of this paper

1. Propose a fast pseudo-likelihood algorithm for fitting the block model, as well as its variation conditional on node degrees that allows for fitting networks with highly variable node degrees within communities.
    - Ignore the symmetry assumption in the adjacency matrix. Apply block compression — divide the nodes into blocks and only look at the likelihood of the row sums within blocks.
    - Accurate and fast approximation for fitting a block model to networks with tens of millions of nodes
2. Proof of consistency of one step of the algorithm
3. Propose spectral clustering with perturbations, a new clustering method, used to initialize pseudo-likelihood in practice.

Algorithms

Algorithms — Pseudo-likelihood

In principle, the joint likelihood of A and c can be maximized via the EM algorithm. However, the E-step involves optimizing over all possible label assignments — NP-hard. Instead, the authors come up with a Pseudo-likelihood (PL) function and then estimate the parameters.

Algorithms — Pseudo-likelihood

- Introduce an initial labelling vector $e = (e_1, \ldots, e_n)$, $e_i \in \{1, \ldots, K\}$ — partitions the nodes into K groups
- The main quantity that we work with are the block sums along the columns defined as:

$$b_{ik} = \sum_j A_{ij} 1(e_j = k)$$

  for $i = 1, \ldots, n$ and $k = 1, \ldots, K$
- Let $b_i = (b_{i1}, \ldots, b_{iK})$
- b_{ik} = the number of neighbours of the ith node that belong to the kth group

Algorithms — Pseudo-likelihood

- Let R be the $K \times K$ matrix with

$$R_{ka} = \frac{1}{n} \sum_{i=1}^{n} 1(e_i = k, c_i = a)$$

- Let $R_{k\bullet}$ = the kth row of R
- Let $P_{\bullet l}$ = the l-th column of P — ($\mathrm{E}\left[A_{ij} \middle| c\right] = P_{c_i c_j}$ in SBM)
- Let $\lambda_{lk} = nR_{k\bullet}P_{\bullet l}$ and $\Lambda = \{\lambda_{lk}\}$
- For each node i, conditional on labels $c = (c_1, \ldots, c_n)$ with $c_i = l$:
  - (A) $\{b_{i1}, \ldots, b_{iK}\}$ are mutually independent.
  - (B) b_{ik}, is approximately poisson with mean λ_{lk}
- With true labels $\{c_i\}$ unknown, each b_i can be viewed as a mixture of Poisson vectors, identifiable if Λ has no identical rows.

Algorithms — Pseudo-likelihood

- Let $\lambda_l = \sum_k \lambda_{lk}$
- Ignore the dependence among $\{b_i, i = 1, \ldots, n\}$
- Use the Poisson assumption and treat $\{c_i\}$ as latent variable and write the pseudo log-likelihood as follows:

$$\ell_{PL}(\pi, \Lambda; \{b_i\}) = \sum_{i=1}^{n} \log \left(\sum_{l=1}^{K} \pi_l e^{-\lambda_l} \prod_{k=1}^{K} \lambda_{lk}^{b_{ik}} \right)$$

- Can be maximized via the EM algorithm for mixture models

Algorithms — EM Algorithm for Pseudo-likelihood

- Initialize the parameters. (How to initialize $e = (e_1, \ldots, e_n)$?)
- Repeat T times:
  1. Compute the block sums $\{b_{il}\}$
  2. Using current parameter estimates $\{pi\}$ and $\{\Lambda\}$, estimate probabilities for node labels by

$$\hat{\pi}_{il} = P_{PL}(c_i = l | b_i) = \frac{\hat{\pi}_l \prod_{m=1}^{K} \exp(b_{im} \log \hat{\lambda}_{lm} - \hat{\lambda}_{lm})}{\sum_{k=1}^{K} \hat{\pi}_k \prod_{m=1}^{K} \exp(b_{im} \log \hat{\lambda}_{km} - \hat{\lambda}_{km})}$$

  3. Given label probabilities, update parameter values as follows:

$$\hat{\pi}_l = \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_{il}, \qquad \hat{\lambda}_{lk} = \frac{\sum_i \hat{\pi}_{il} b_{ik}}{\sum_i \hat{\pi}_{il}}$$

  4. Return to 2 unless the parameter estimates have converged.
  5. Update labels by $e_i = \arg\max_l \hat{\pi}_{il}$ and return to 1
  6. Update $\hat{P}$ by $\hat{P}_{lk} = \frac{\sum_{i,j} A_{ij} \hat{\pi}_{il} \hat{\pi}_{jk}}{\sum_i 1(e_i = k) \sum_j 1(e_j = k)}$

Algorithms — PL Conditional on Node Degrees

- For networks with hub nodes or those with substantial degree variability within communities, the block model can provide a poor fit.
- Divides the nodes into low-degree and high-degree groups
- Supported empirically and by theory.
- What to do:
  (a) Extension to the degree-corrected SBM — has an extra degree parameter for every node which has to be estimated — difficult to write down a PL and use EM
  (b) In this paper, the authors consider the PL conditional on the observed node degrees — Whether these degrees are similar or not will not then matter, and the fitted parameters will reflect the underlying block structure rather than the similarities in degrees.

Algorithms — EM for CPL

The conditional pseudo-likelihood can be obtained using the following observation:

- If random variables X_k are independent Poisson with means μ_k, their distribution conditional on $\sum_k X$ is multinomial.

Thus

- The distribution of $(b_{i1}, \ldots, b_{iK})$ conditional on labels c with $c_i = l$ and the node degree $d_i = \sum_k b_{ik}$, is multinomial with parameters $(d_i; \theta_{l1}, \ldots, \theta_{lK})$, where $\theta_{lK} = \frac{\lambda_{lk}}{\lambda_l}$
- The conditional pseudo log-likelihood is then given by:

$$\ell_{CPL}(\pi, \Theta; \{b_i\}) = \sum_{i=1}^{n} \log \left(\sum_{l=1}^{K} \pi_l \prod_{k=1}^{K} \theta_{lk}^{b_{ik}} \right)$$

Algorithms — EM for CPL

Now modify steps 2 and 3 of the unconditional EM algorithm to

1. based on current parameter estimates $\{pi\}$ and $\left\{\hat{\theta_{lk}}\right\}$, estimate probabilities for node labels by

$$\hat{\pi}_{il} = P_{CPL}(c_i = l|b_i) = \frac{\hat{\pi}_l \prod_{m=1}^{K} \hat{\theta}_{lm}^{b_{im}}}{\sum_{k=1}^{K} \hat{\pi}_k \prod_{m=1}^{K} \theta_{km}^{b_{im}}}$$

2. Given label probabilities, update parameter values as follows:

$$\hat{\pi}_l = \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_{il}, \qquad \hat{\lambda}_{lk} = \frac{\sum_i \hat{\pi}_{il} b_{ik}}{\sum_i \hat{\pi}_{il} d_i}$$

Algorithms — Initialization of e

Clustering based on 1- and 2-degrees

- Separate the nodes by degree using a clustering algorithm such as the one-dimensional k-means
- Works for only a certain type of block models, identifiable from their degree distribution
- In this paper, the authors consider a two-dimensional K-means clustering on the degree and the number of paths of length 2 from node i

Algorithms — Initialization of e

Spectral clustering

- Based on the spectral properties of the adjacency matrix A or its graph Laplacian.
- Let $D = diag(d_1, \ldots, d_n)$ — diagonal matrix collecting node degrees.
- Look at the eigenvalues of the normalized graph Laplacian $L = D^{-1/2} A D^{-1/2}$
- Choose a small number, say $r = K - 1$ of the eigenvectors corresponding to the r largest eigenvalues with the largest omitted.
- These vectors provide an r-dimensional representation of the nodes.
- Apply K-means to find clusters.

Algorithms — Initialization of e

Spectral clustering with perturbations

- The authors found that spectral clustering performs poorly for community detection for sparse graphs with expected degree $\lambda < 5$
- They provide a modification: Connect all disconnected components which belong to the same community by adding artificial "weak" links.
- Regularize the adjacency matrix A by adding $\alpha/p \times \lambda/n$ multiplied by the adjacency matrix of an ErdosRenyi graph on n nodes with edge probability p, where α is a constant.
- Found $\alpha/p = 0.25$ and $p = 1$ works well for their range of simulations.
- Now do the usual steps of the spectral clustering.

Numerical Results

Numerical Results

- Simulate two scenarios: SBM and degree-corrected SBM
- Fix $K = 3$ and $\pi = (1/3, 1/3, 1/3)$
- Conditional on the labels the edges are generated as independent Bernoulli r.v.s with $\mathrm{E}\left[A_{ij} \middle| c\right] = \theta_i \theta_j P_{c_i c_j}$
- The parameters θ_j are drawn independently from the distribution of Θ with $P(\Theta = 0.2) = \rho$ and $P(\Theta = 1) = 1 - \rho$
- $\rho = 0$ corresponds to the regular block model and $\rho = 0.9$ corresponds to a network where 10% of the nodes can be viewed as hubs.

Numerical Results

- Matrix P controlled by two parameters:
  1. $\beta =$ out-in-ratio — will be varied from 0 to 0.2
  2. the weight vector w which determines the relative degrees within communities — $w = (1, 1, 1)$ contains no information about communities in node degrees. $w = (1, 5, 10)$, degrees provide relevant information for clustering
- Also we will vary the overall expected network degree λ from 1 to 15
- P is constructed in such a way to have have λ as expected degree:

$$P = \frac{\lambda}{(n-1)(\pi^T P^{(0)} \pi)(\mathrm{E}(\Theta))^2} P^{(0)}$$

with $P^{(0)} = 1(\beta \neq 0) diag(w) \beta^{-1} + 1(\beta = 0) diag(w)$

Numerical Results

- To compare the results with the true labels use the normalized mutual information (NMI):

$$NMI(c, e) = -\sum_{i,j} R_{ij} \log \frac{R_{ij}}{R_{i+}R_{+j}} \Big(\sum_{i,j} R_{i,j} \log R_{i,j} \Big)^{-1}$$

- Always between 0 and 1 (perfect match)
- For n large, matching $50\%, 70\%$ and 90% corresponds to NMI of approximately 0.12, 0.26, and 0.58

Numerical Results

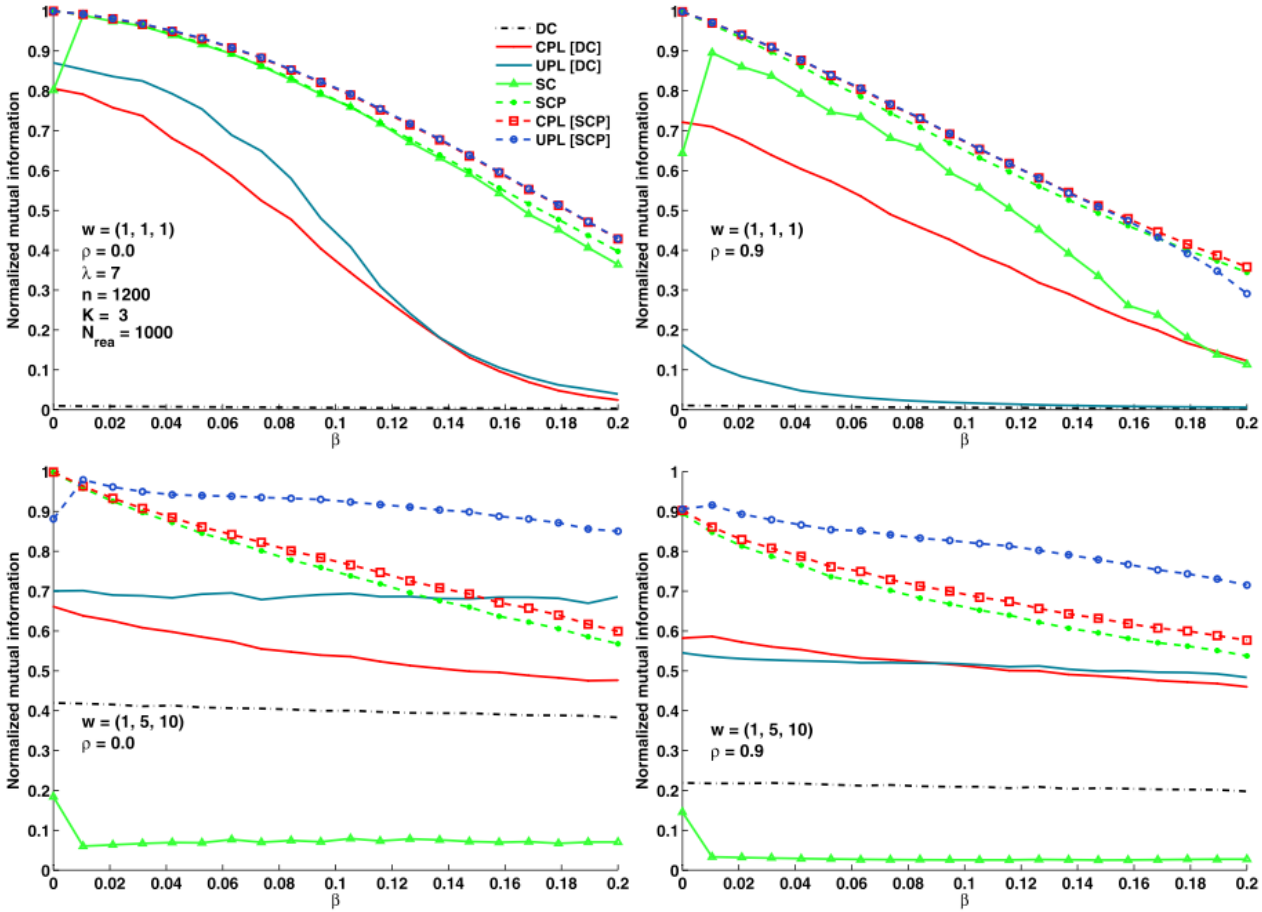

Figure : The NMI between true and estimated labels as a function of "out-in-ratio" β

Numerical Results

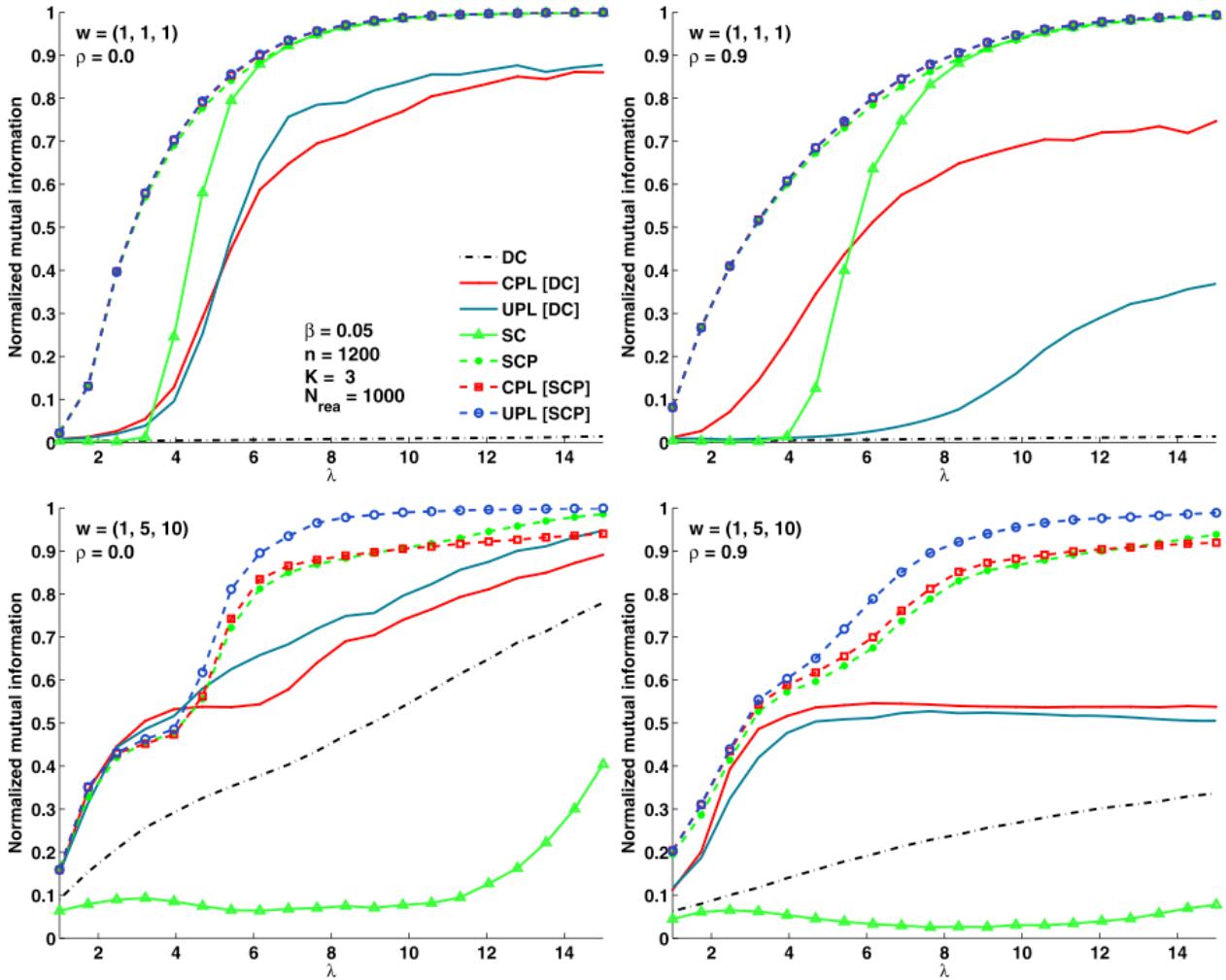

Figure : The NMI between true and estimated labels as a function of average expected degree λ

Numerical Results

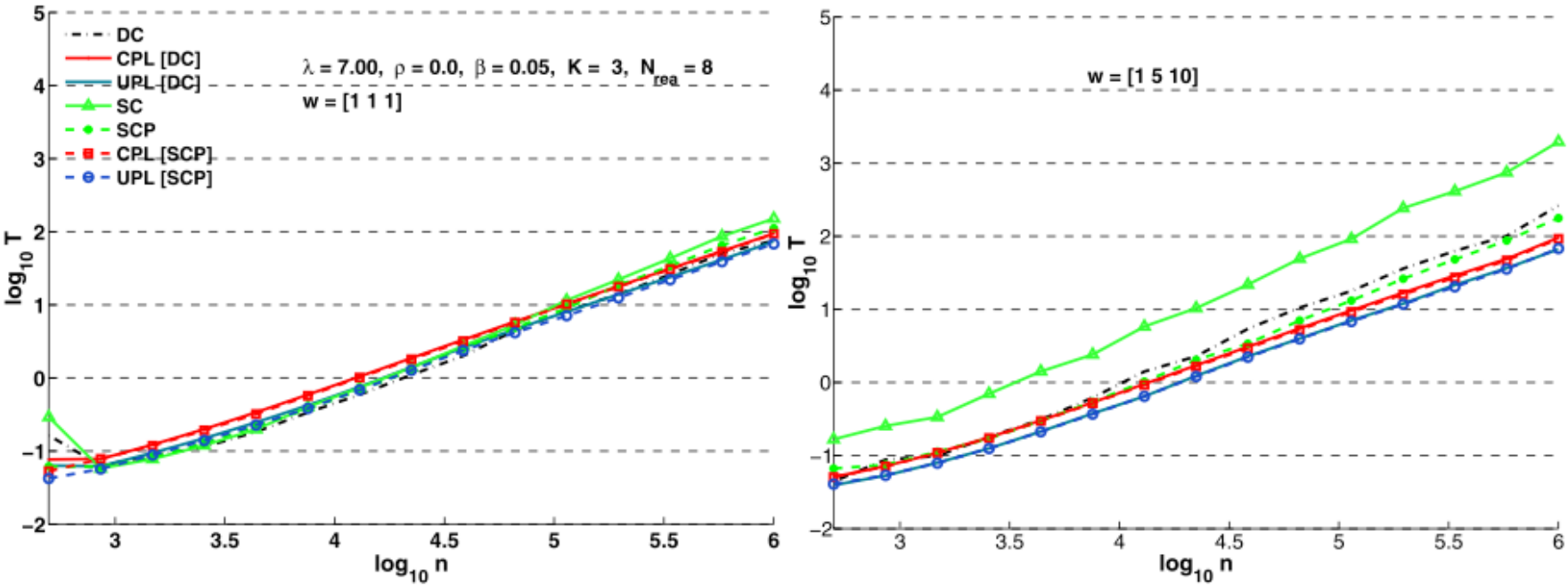

Figure : The runtime in seconds as a function of the number of nodes (loglog scale)

Example: A Political Blogs Network

A Political Blogs Network

- Dataset on political blogs after 2004 U.S. presidential election.
- Nodes are blogs, edges are hyperlinks between blogs.
- Each blog was manually labeled as liberal or conservative (ground truth)
- Ignore direction of hyperlink
- Analyze the largest connected component which has 1222 nodes and average degree of 27
- The distribution of degrees is highly skewed to the right (median degree is 13 and the maximum is 351)

A Political Blogs Network

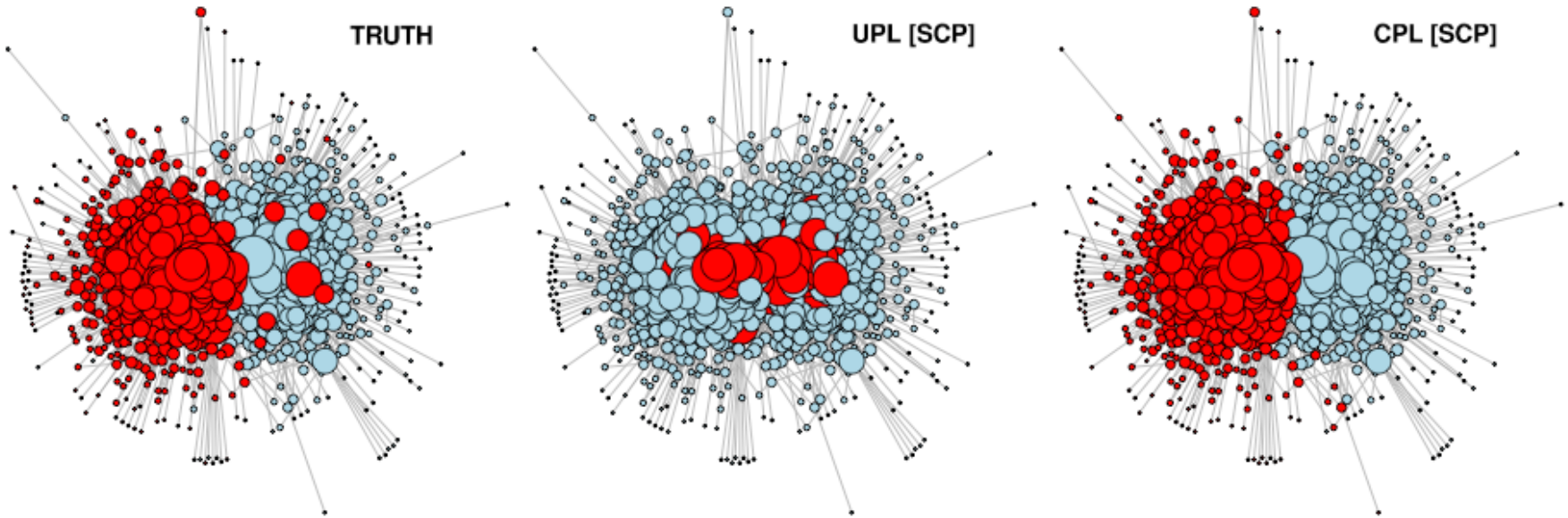

Figure : Political blogs data: true labels and unconditional and conditional pseudo-likelihoods (UPL and CPL) initialized with spectral clustering with perturbations (SCP). Node size is proportional to log degree

Referenece

Amini, Arash A.; Chen, Aiyou; Bickel, Peter J.; Levina, Elizaveta. Pseudo-likelihood methods for community detection in large sparse networks. The Annals of Statistics 41 (2013), no. 4, 2097–2122. doi:10.1214/13-AOS1138. http://projecteuclid.org/euclid.aos/1382547514