Detect Community Better
Variational EM for Blockmodels with Covariates

Fan Chen [1] and Arash A. Amini [2]

[1] School of Mathematical Sciences, Zhejiang University, China

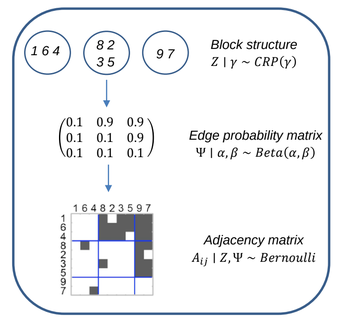
[2] Department of Statistics, University of California, Los Angeles, CA

1. Summary

Community detection is one of essential destinations of statistical learning on network datasets. Modern industries construct vast amount of data consisting of both graphical edge but also node feature. User information in Facebook ego networks serves as a example with that regard. Adding node covariates into Stochastic Blockmodels (SBMC), we approximately estimate SBMC by Variational EM (VEM) method, with some theoretical guarantees. Though simulation and robustness test, we explore the performance of new proposed methods, after which we implement them on Facebook ego networks and gain the similar results.

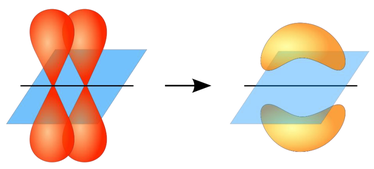
2. Stochastic Blockmodel for community detection

- SBM is a model of community-structure networks.
- Network ⟺ adjacency matrix $A \in \{0,1\}^{n\times n}$ $(A_{ij} = 1 \text{ iff. } i \leftrightarrow j)$.
- Undirected network ⟺ symmetric A.
- K-community structure ⟺ *membership matrix* $Z \in \{0,1\}^{n\times K}$.
$$Z^\top = [z_1, \dots, z_n]$$
where $z_i \in \{0,1\}^K$: membership vector of node i.
($z_{ik} = 1$ iff i belongs to community k)
- SBM in addition has an *edge probability matrix* $\Psi \in [0,1]^{K\times K}$.
- SBM: $\mathbb{P}(A_{ij} = 1 \mid z_i, z_j, \{z_\ell\}_{\ell\neq i,j}) = z_i^\top \Psi z_j, i < j$.
- Compactly: $\mathbb{E}[A \mid Z] = Z\Psi Z^\top$.



3. Adding node covariates

- R-Node features (covariates) ⟺ sample matrix $X \in \mathbb{R}^{n\times R}$, $X^\top = [X_1, \dots, X_n]$
- Given Z, suppose covariates matrix $(X|Z)$ and adjacency matrix $(A|Z)$ are statistically independent.
- Graphical model gives
$p_\theta(A, X, Z) = p_\theta(A \mid Z)p_\theta(X \mid Z)p_\theta(Z)$.



4. Variational inference for blockmodel with covariates

- Maximize log-likelihood $\ell(\theta; A) \coloneqq \log \sum_z p_\theta(z, A)$.
- Extended confidential lower bound on $\ell(\theta; A)$
$\mathcal{J}(\theta, q; A, X) \coloneqq \mathbb{E}_q[\log p_\theta(A, X, Z) - \log q(Z)]$ for all density q and θ are parameters of A, X. (e.g. for Gauss PDF, $\theta = (\alpha, \pi, \mu, \Sigma)$)
- Mean-field approximation: a restriction of the form $q(z) = \prod_j q_j(z_j)$
- Variational EM: alternate between τ-ascend and θ-ascent.

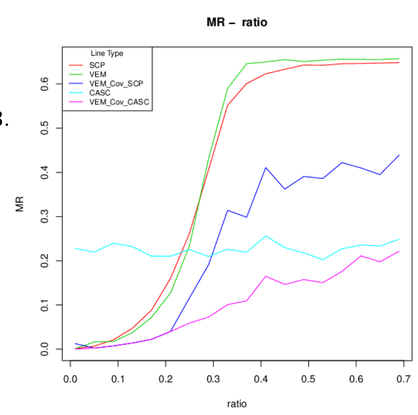
Evidence lower bound → Mean-field approach → Variational inference → EM approach

5. Theoretical guarantees

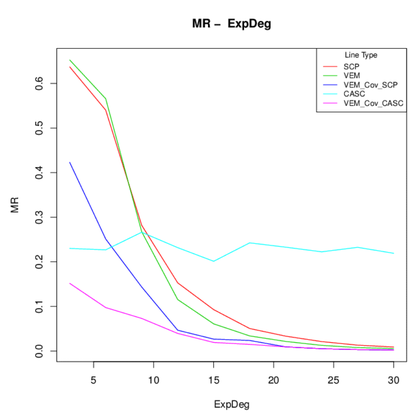
- Misclassified rate bound:
$$\frac{|\mathcal{M}|}{N} \leq \frac{c_0 KP\delta \log(8N/\epsilon)}{N\lambda_K^2}.$$
(Celisse & al. (2011).)
- SBMC is a 'Degenerated' model:
assume true label are z^*, then
$\Pr\left\{\frac{\sum_{z\neq z^*} g(z;A,X)}{g(z^*;A,X)} > t \,\middle|\, Z = z^*\right\} = \mathcal{O}\left(ne^{-\kappa(t)n}\right)$. (N. Binkiewicz, 2015)
- Ongoing work about the convergence
$\Pr(\cdot \mid A, X) \to \delta\{z^0\}$

6. Numerical results: simulation

- $N = 1500, K = 3, R = 3$.
- Link probability matrix
$$\begin{bmatrix} p & q & q \\ q & p & q \\ q & q & p \end{bmatrix}$$
- Feature data obey multivariate normal distribution $\mathrm{N}_R(\mu, \Sigma)$.

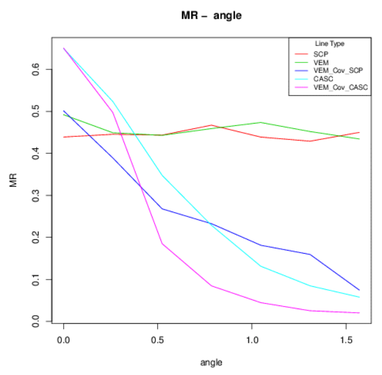


In-out ratio β: link probability without blocks over within blocks, i.e. $\beta = q/p$. The higher β, the harder to detect community

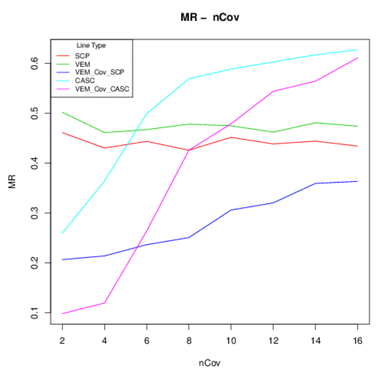


Expected degree λ: number of *expected* total connection, i.e. $\lambda = p[(1-\beta)S + \beta(N-S)]$. The higher expected degree, the less sparse adjacency matrix, thus the easier to detect community

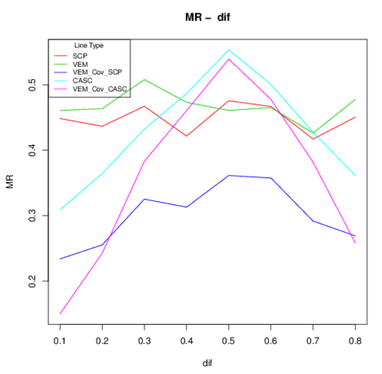
7. Numerical results: robustness and sensitivity



Intersection angle of every two mean vectors: the divergence of covariates. The larger intersection angle, the lower inner product, thus the easier to detect community.

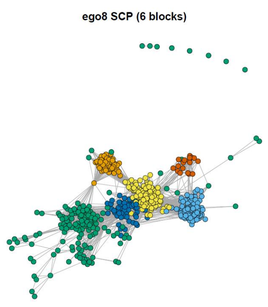


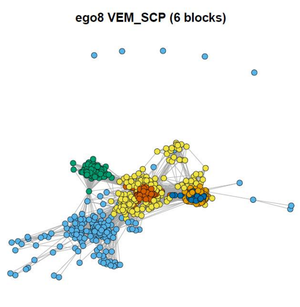
Number of covariates: dimension of feature data space. The higher dimension, the more difficult to utilize them, thus leading to larger MR.

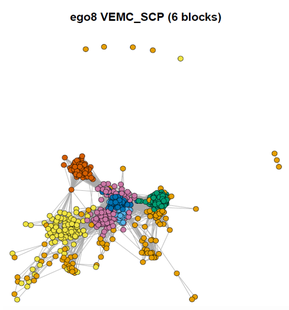


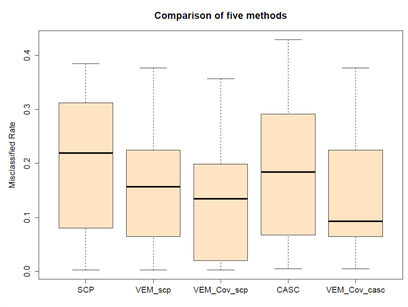
Difference covariate rate: proportion of deliberately assigned covariate data of other (neighboring) blocks. 50% is a peak, because the 'dif' more than that would make the covariate data shift a step, thus $\mathrm{dif}' = 1 - \mathrm{dif}$.

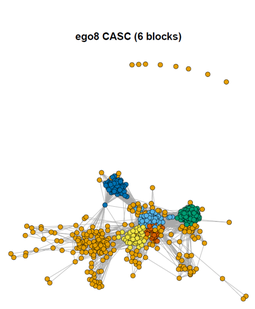
8. Illustration on Facebook ego networks

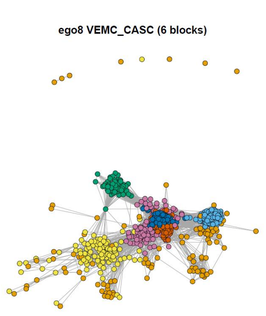


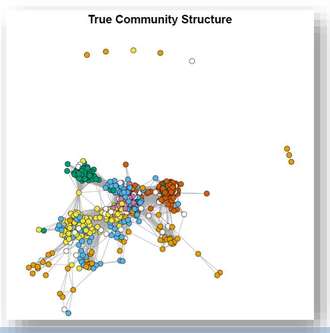












As a illustration, we implement our methods on Facebook ego networks (J. McAuley, 2007) together with existing SCP (A. Amini, 2013) and CASC (N. Binkiewicz, 2015), comparing with true community structure, to which one can see that VEMCs are greatly more similar. The box-plot also shows the median and quantiles of each algorithm.

9. Further work

- More scalable datasets
(Amazon (300k nodes, J. Leskovec, 2007), in process)
- More detailed theoretical guarantees
- Generalized linear model on (binary) covariates
(in process)

10. Acknowledge

I would like to show the great gratitude to Prof. Arash Amini for accepting me as CSST student and providing me with a strong support. Thanks for Prof. Amini for his enormous contribution, descriptive advice and pleasant personality. I would also like to appreciate Mr. Kang Ying's generous help with preprocessing work on Amazon dataset.