Outliers detection in networks with missing links

Solenne Gaucher[1], Olga Klopp[2], Geneviève Robin[3]

[1] Université Paris-Sud
[2] ESSEC Business School, ENSAE
[3] CNRS, Université d'Évry Val d'Essonne

22 October, 2021

Example: Les Misérables characters network

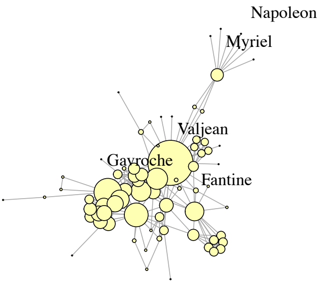


Figure: Network of Victor Hugo's novel Les Misérables characters

Network data:

- $\mathcal{V}$: Set of **nodes** (characters: Jean Valjean, Fantine, etc.).
- $\mathcal{E}$: Set of **edges** between nodes (whenever two characters appear in the same chapter).

General framework for network data analysis

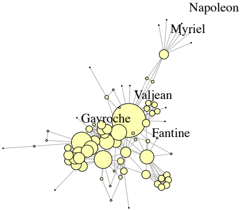


Figure: Network $\mathcal{G}$

$\Leftrightarrow$

	Napoleon	Myriel	Mlle Baptistine	...
Napoleon	0	1	0	
Myriel	1	0	1	
Mlle Baptistine	0	1	0	
$\vdots$				

Table: Adjacency matrix A

$$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}, \quad |\mathcal{V}| = n$$

$$A \in \{0,1\}^{n \times n}, \quad A_{ij} = 1 \Leftrightarrow (i,j) \in \mathcal{E}$$

General framework for network data analysis

- **Imperfect data setting**:
  1. Missing values (possibly many: machine failure, individual non response, etc.)
  2. Outliers (hubs, adversary agents, etc.): "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism"

- **Objectives**:
  1. Predict missing values: estimation of connection probabilities
  2. Detect outliers: support recovery
  3. Scalable method ($\sim$10,000 nodes)

The Stochastic Block Model (SBM) [Holland et al., 1983]

- Classical network model in probabilistic framework.
- **Main ideas**:
  1. The nodes (characters) are partitioned into unknown communities (narrative units).
  2. The probability that two nodes are connected (appear in the same chapter) depends on their respective communities.
  3. Communities assignment and connection probabilities are learned from data.

The Stochastic Block Model (SBM)

- Nodes distributed across $K > 0$ communities w.p. $\pi_1, \dots, \pi_K$.
- Denote $z_{ik} = \begin{cases} 1 \text{ if node } i \text{ is in community } k \\ 0 \text{ otherwise} \end{cases}$:

$$(z_{i1}, \dots, z_{iK}) \sim \mathrm{Multi}(1, (\pi_1, \dots, \pi_K)).$$

- The probability of connections between nodes is given by $\mathrm{Q} \in [0,1]^{K \times K}$ symmetric matrix of connection probabilities between communities:

$$\mathbb{P}(A_{ij} = 1 | z_{ik} = 1, z_{jl} = 1) = \mathrm{Q}_{k,l}$$

The Stochastic Block Model (SBM)

$$\begin{bmatrix} \cdot & 1 & 0 & 0 & 1 \\ 1 & \cdot & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 1 \\ 0 & 0 & 0 & \cdot & 1 \\ 1 & 0 & 1 & 1 & \cdot \end{bmatrix}$$

A

$$\begin{array}{ccc} \alpha_1 & \begin{bmatrix} \cdot & Q_{1,1} & Q_{1,2} & Q_{1,2} & Q_{1,2} \\ Q_{1,1} & \cdot & Q_{1,2} & Q_{1,2} & Q_{1,2} \\ Q_{1,2} & Q_{1,2} & \cdot & Q_{2,2} & Q_{2,2} \\ Q_{1,2} & Q_{1,2} & Q_{2,2} & \cdot & Q_{2,2} \\ Q_{1,2} & Q_{1,2} & Q_{2,2} & Q_{2,2} & \cdot \end{bmatrix} & \alpha_2 \end{array}$$

$\mathbb{E}[A]$

- Up to reordering of the nodes, the expected adjacency matrix is *block-wise constant*. Its *rank* is at most K.

Limitations of the SBM

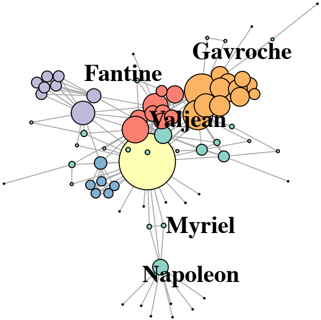


Figure: SBM applied to Les Misérables characters network (nodes colored by community)

Some nodes are not well modeled:

- Hubs (Jean Valjean, Myriel)
- Mixed memberships (Gavroche)
- Neutral nodes (Napoléon)
- Adversarial nodes

Unobserved edges:

- Nonresponse, dropout
- Expensive exploration of interactions

Our contributions

1. New algorithm to estimate connection probabilities in a network, robust to outliers and missing values
2. Achieves exact detection of the outliers (new result in low-rank plus sparse matrix decomposition)
3. Estimation guarantees for connection probabilities (best known error for polynomial time algorithms)
4. Scalable to moderate networks ($\sim$1e4 nodes, 1e6 edges)
5. Numerical illustration & applications to epidemiology and social network analysis

General model

- Decompose the set of edges into $\mathcal{V} = \mathcal{I} \cup \mathcal{O}$, where $\mathcal{I}$ is a set of $n - s$ inliers, and $\mathcal{O}$ is a set of s outliers.

General model

- Decompose the set of edges into $\mathcal{V} = \mathcal{I} \cup \mathcal{O}$, where $\mathcal{I}$ is a set of $n - s$ inliers, and $\mathcal{O}$ is a set of s outliers.
- For $(i, j) \in \mathcal{I}^2$, $A_{ij} \sim \text{Bernoulli}(\mathsf{L}^\star_{ij})$, where $\mathsf{L}^\star$ is a *low-rank matrix* in $[0, 1]$.

General model

- Decompose the set of edges into $\mathcal{V} = \mathcal{I} \cup \mathcal{O}$, where $\mathcal{I}$ is a set of $n - s$ inliers, and $\mathcal{O}$ is a set of s outliers.
- For $(i, j) \in \mathcal{I}^2$, $A_{ij} \sim \text{Bernoulli}(\mathsf{L}^\star_{ij})$, where $\mathsf{L}^\star$ is a *low-rank matrix* in $[0, 1]$.
- For $(i, j) \in \mathcal{O} \times \mathcal{I}$, $A_{ij} \sim \text{Bernoulli}(\mathsf{S}^\star_{ij})$, where $\mathsf{S}^\star$ is an *column-wise sparse matrix* in $[0, 1]$.

General model

- Decompose the set of edges into $\mathcal{V} = \mathcal{I} \cup \mathcal{O}$, where $\mathcal{I}$ is a set of $n - s$ inliers, and $\mathcal{O}$ is a set of s outliers.

- For $(i, j) \in \mathcal{I}^2$, $A_{ij} \sim \text{Bernoulli}(\mathsf{L}^\star_{ij})$, where $\mathsf{L}^\star$ is a *low-rank matrix* in $[0, 1]$.

- For $(i, j) \in \mathcal{O} \times \mathcal{I}$, $A_{ij} \sim \text{Bernoulli}(\mathsf{S}^\star_{ij})$, where $\mathsf{S}^\star$ is an *column-wise sparse matrix* in $[0, 1]$.

- For $(i, j) \in \mathcal{O} \times \mathcal{O}$, $A_{ij} \sim \text{Bernoulli}(\mathsf{S}^\star_{ij} + \mathsf{S}^\star_{ji})$, where $\mathsf{S}^\star$ is an *column-wise sparse matrix* in $[0, 1]$.

General model

- Low-rank matrix $L^\star$ generalizes the block-wise constant SBM model (of rank K).
- $S^\star$ contains the arbitrary connection probabilities of outliers:

$$S^\star = \begin{bmatrix} \cdot & 0 & 0 & S^\star_{1,4} & S^\star_{1,5} \\ 0 & \cdot & 0 & S^\star_{2,4} & S^\star_{2,5} \\ 0 & 0 & \cdot & S^\star_{3,4} & S^\star_{3,5} \\ 0 & 0 & 0 & \cdot & S^\star_{4,5} \\ 0 & 0 & 0 & S^\star_{5,4} & \cdot \end{bmatrix}$$

- Column-wise sparse $\Rightarrow$ small number of ouliers s compared to the total number of nodes n.
- $\mathbb{E}[A_{ij}] = L^\star_{ij} + S^\star_{ij} + S^\star_{ji}$

Estimation procedure

- Objective function:

$$\Phi_\epsilon(S, L) \triangleq \underbrace{\frac{1}{2}\|\Omega \odot (A - L - S - (S)^\top)\|_F^2}_{\text{data fitting term}} + \underbrace{\lambda_1 \|L\|_\star}_{\text{low-rank penalty}} + \underbrace{\lambda_2 \|S\|_{2,1}}_{\text{column-wise sparse penalty}}$$

Estimation procedure

- Objective function:

$$\Phi_\epsilon(S, L) \triangleq \underbrace{\frac{1}{2}\|\Omega \odot (A - L - S - (S)^\top)\|_F^2}_{\text{data fitting term}} + \underbrace{\lambda_1 \|L\|_\star}_{\text{low-rank penalty}} + \underbrace{\lambda_2 \|S\|_{2,1}}_{\text{column-wise sparse penalty}}$$

- Estimation problem:

$$(\hat{S}, \hat{L}) \in \operatorname{argmin}_{S \in [0,1]^{n \times n}, L \in [0,1]^{n \times n}} \Phi_\epsilon(S, L)$$

Estimation procedure

- Objective function:

$$\Phi_\epsilon(S, L) \triangleq \underbrace{\frac{1}{2}\|\Omega \odot (A - L - S - (S)^\top)\|_F^2}_{\text{data fitting term}} + \underbrace{\lambda_1 \|L\|_\star}_{\text{low-rank penalty}} + \underbrace{\lambda_2 \|S\|_{2,1}}_{\text{column-wise sparse penalty}}$$

- Estimation problem:

$$(\hat{S}, \hat{L}) \in \operatorname{argmin}_{S \in [0,1]^{n \times n}, L \in [0,1]^{n \times n}} \Phi_\epsilon(S, L)$$

- In practice:

$$(S^{\text{opt}}, L^{\text{opt}}) \in \operatorname{argmin}_{S,L} \Phi_\epsilon(S, L)$$

Assumptions on missing values

- Let Π_{ij} denote the probability to observe the entry A_{ij}: assume $\Pi_{ij} \geq \mu_n > 0$.

- Denote by ν_n and $\tilde{\nu}_n$ two sequences such that
  - For all $i \in \mathcal{I}$, $\sum_{j \in \mathcal{I}} \Pi_{ij} \leq \nu_n n$
  - For all $i \in \mathcal{V}$, $\sum_{j \in \mathcal{O}} \Pi_{ij} \leq \tilde{\nu}_n n$

Assumptions on connections

- Bounded parameters: $\|\mathbf{L}\|_\infty \leq \rho_n$, $\|\mathbf{S}\|_\infty \leq \gamma_n$
  - $\rho_n n$ average degree of inliers
  - $\gamma_n n$ average degree of outliers

- Number of observed edges:

$$\nu_n \rho_n n \geq \log(n), \quad \tilde{\nu}_n \gamma_n n \geq \log(n)$$

Signal to noise ratio

- For estimation of connection probabilities: $\nu_n \rho_n n \geq \tilde{\nu}_n \gamma_n s$
  - $\nu_n \rho_n n$: average observed degree of inliers
  - $\tilde{\nu}_n \gamma_n n$: average observed degree of outliers

- For detection of outliers: $\sum_{i \in \mathcal{I}} \Pi_{ij} S^{\star}_{ij} \geq C \nu_n \rho_n n$
  - $\sum_{i \in \mathcal{I}} \Pi_{ij} S^{\star}_{ij}$: average number of observed edges between outliers and inliers
  - $\nu_n \rho_n n$: average observed degree of inliers

Estimation of connection probabilities

Theorem

Choose $\lambda_1 = 84\sqrt{\nu_n \rho_n n}$ *and* $\lambda_2 = 19\sqrt{\nu_n \rho_n n}$*. Then, there exists absolute constants* $C > 0$ *and* $c > 0$ *such that with probability at least* $1 - \frac{c}{n}$*,*

$$\left\| \left(\widehat{\mathrm{L}} - \mathrm{L}^*\right)_{|I} \right\|_F^2 \leq \frac{C}{\mu_n}\left(\frac{\nu_n}{\mu_n}\rho_n kn + (\nu_n\rho_n \vee \tilde{\nu}_n\gamma_n)\rho_n sn\right).$$

- Denote by $\hat{\mathcal{O}}$ the set of outliers detected by the MCGD algorithm (the nonzero columns in $\mathsf{S}^{(T)}$).

Theorem (Outliers detection)

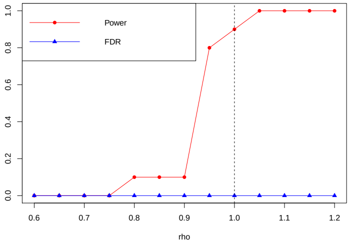
Let $\lambda_2 = 19\sqrt{\rho_n \nu_n n}$. There exists an absolute constant $c > 0$ such that with probability at least $1 - \frac{cs}{n}$:

$$\hat{\mathcal{O}} = \mathcal{O}.$$

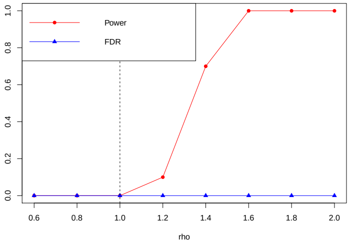
Simulation scheme: SBM with outliers

- $n = 1000$ nodes, $k = 3$ communities
- Connection probabilities: inside community $p = 0.05$, between communities $q = 0.01$
- Add outliers: Hubs connecting to any node with probability π_{hub} and Mixed membership connecting to two communities with probability π_{mix}
- Remove randomly 20% of the links (unobserved links)
- Evaluate the method for outliers detection and link prediction

Outliers detection

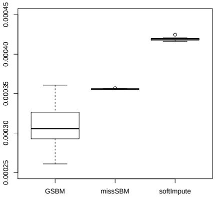


(a) **Hubs** detection: **Power** (red points) and **FDR** (blue triangles) for increasing $\rho_{\text{hub}} \sim \pi_{\text{hub}}/p$, averaged across 10 replications. $\rho_{\text{hub}} = 1$ indicated with dashed black line.

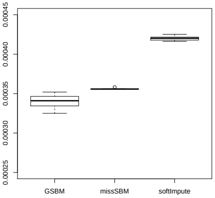


(b) **Mixed membership**: **Power** (red points) and **FDR** (blue triangles) for increasing $\rho_{\text{mix}} \sim \pi_{\text{mix}}/p$, averaged across 10 replications. $\rho_{\text{mix}} = 1$ indicated with dashed black line.

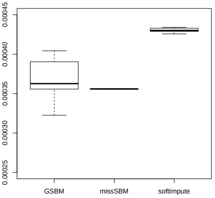
Link prediction



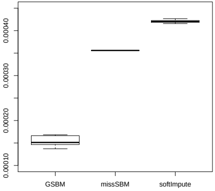
(a) $\tau_{\text{hub}} = 1$



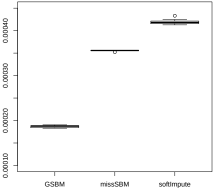
(b) $\tau_{\text{hub}} = 2$



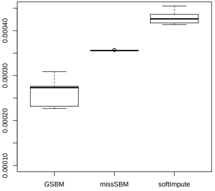
(c) $\tau_{\text{hub}} = 5$



(a) $\tau_{\text{mix}} = 1$



(b) $\tau_{\text{mix}} = 2$



(c) $\tau_{\text{mix}} = 5$

Political Twitter network

- Mentions network between political Twitter accounts, first analyzed in Fraisier et al. [2018], collected during the 2017 French presidential election
- 22,853 (political Twitter accounts)
- 1,896,262 edges edges (mentions in Tweets)
- Each account labeled manually according to political preferences (FI, LR, LREM,PS, RN)
- Apply our method and look at the detected outliers

Political Twitter network: results

- Around 600 detected outliers
- Large hubs: Political figures (candidates: Benoît Hamon, Jean-Luc Mélenchon, etc., journalists and elected officials: Jean-Jacques Bourdin, Alexis Corbière, etc.), main media, unofficial political groups.
- Mixed membership nodes: Accounts affiliated to multiple political parties (smaller hubs: Christine Boutin, La Manif Pour Tous) and individual profiles with no public exposition (@mrericmas: LREM/LR, @erayeye: LR/RN, @Apostillier1: LREM/PS, etc.) that would not be detected using histogram of degrees

Conclusion

Summary

- New algorithm to analyze network data in presence of outliers and missing links
- Exact detection of outliers
- Estimation error of connection probabilities
- Encouraging empirical results
- R package `gsbm`

Future work

- Classification properties of the algorithm
- Detection of groups of outliers
- Extension to dynamic networks

Generalized SBM [Cai and Li, 2015]

- The set of nodes $\mathcal{V}$ contains $n - s$ "inliers" obeying the SBM and s "outliers" connecting other nodes in an arbitrary way:

$$\mathcal{V} = \underbrace{\mathcal{I}}_{\text{inliers}} \cup \underbrace{\mathcal{O}}_{\text{outliers}}$$

Generalized SBM [Cai and Li, 2015]

- The set of nodes $\mathcal{V}$ contains $n-s$ "inliers" obeying the SBM and s "outliers" connecting other nodes in an arbitrary way:

$$\mathcal{V} = \underbrace{\mathcal{I}}_{\text{inliers}} \cup \underbrace{\mathcal{O}}_{\text{outliers}}$$

- Estimate the matrix of community assignments $Z = (z_{ik})_{1 \leq i \leq n, 1 \leq k \leq K}$ via convex optimization (semi-definite program).

Generalized SBM [Cai and Li, 2015]

- The set of nodes $\mathcal{V}$ contains $n-s$ "inliers" obeying the SBM and s "outliers" connecting other nodes in an arbitrary way:

$$\mathcal{V} = \underbrace{\mathcal{I}}_{\text{inliers}} \cup \underbrace{\mathcal{O}}_{\text{outliers}}$$

- Estimate the matrix of community assignments $Z = (z_{ik})_{1\leq i\leq n, 1\leq k\leq K}$ via convex optimization (semi-definite program).

- Main result: *inliers* are correctly classified into communities with high probability.

SBM with unobserved edges [Tabouy et al., 2017]

- Unobserved dyads (pairs of nodes) in networks
- Estimation of SBM parameters with Variational Expectation-Maximization (VEM)
- Missing Completely At Random (MCAR), Missing At Random (MAR), Not Missing At Random (NMAR) settings
- Unbiased estimation in several NMAR settings
- Does not account for outliers

Optimization algorithm

- R package `gsbm`
- Mixed Coordinate Gradient Descent
- S and L are updated alternatively along descent directions

Algorithm 1 Mixed coordinate gradient descent (MCGD)

1: **Initialization:** $(\mathrm{L}^{(0)}, \mathrm{S}^{(0)}, t) \leftarrow (0, 0, 0)$
2: **for** $t = 1, \ldots, T$ **do**
3: $t \leftarrow t + 1$
4: Compute a proximal update with fixed step size to obtain $\mathrm{S}^{(t)}$.
5: Compute a Conjugate Gradient update to obtain $\mathrm{L}^{(t)}$ (step size given by theory).
6: **end for**
7: **return** $\left(\mathrm{L}^{(T)}, \mathrm{S}^{(T)}\right)$

Convergence of the algorithm

Theorem (Sublinear convergence of MCGD)

For $\delta > 0$, the MCGD algorithm converges to a δ-optimal solution in $\mathcal{O}(1/\delta)$ iterations:

$$|\Phi_\epsilon(\mathsf{S}^{(T_\delta)}, \mathsf{L}^{(T_\delta)}) - \Phi_\epsilon(\mathsf{S}^{opt}, \mathsf{L}^{opt})| \leq \delta,$$

$T_\delta = \mathcal{O}(1/\delta)$.

T. T. Cai and X. Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Annals of Statistics*, 43(3):1027–1059, 2015.

Ophélie Fraisier, Guillaume Cabanac, Yoann Pitarch, Romaric Besançon, and Mohand Boughanem. Élysée2017fr: The 2017 french presidential campaign on twitter. 2018. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17821.

P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

T. Tabouy, P. Barbillon, and J. Chiquet. Variational Inference for Stochastic Block Models from Sampled Data. *ArXiv e-prints*, July 2017.