Sparse Network Estimation

Olga Klopp

Joint works with

Alexandre Tsybakov

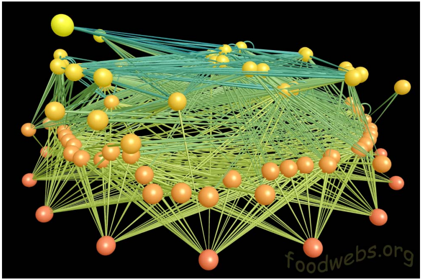
Nicolas Verzelen

Solenne Gaucher

Geneviéve Robin

Network model

Network analysis has become an important research field driven by applications in social sciences, computer sciences, statistical physics, biology,. . .



East-river trophic network [Yoon et al.(04)]

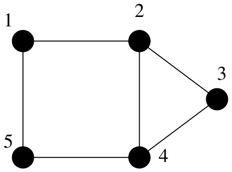
Approach

- The modeling of real networks as random graphs.
- Model-based statistical analysis.

Graph Notations

A (simple, undirected graph) $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ consists of

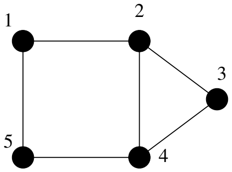
- a set of vertices $V = \{1, \dots n\}$
- a set of edges $E \subset \{\{i, j\} : i, j \in V$ and $i \neq j\}$



Graph Notations

A (simple, undirected graph) $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ consists of

- a set of vertices $V = \{1, \dots n\}$
- a set of edges $E \subset \{\{i, j\} : i, j \in V$ and $i \neq j\}$



$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

The corresponding adjacency matrix is denoted $\mathbf{A} = (\mathbf{A}_{i,j}) \in \{0, 1\}^{n \times n}$, where $\mathbf{A}_{i,j} = 1 \Leftrightarrow (i, j) \in E$

Sparsity

Main integral characteristics

- number of vertices n
- number of edges $|E|$

Sparsity

Main integral characteristics

- number of vertices n
- number of edges $|E|$

Maximal number of edges

$$\frac{n(n-1)}{2}$$

Sparsity

Main integral characteristics

- number of vertices n
- number of edges $|E|$

Maximal number of edges

$$\frac{n(n-1)}{2}$$

- Dense graph $|E| \asymp n^2$
- Real world networks are sparse : $|E| = o(n^2)$
  - more difficult to handle

Stochastic Block-Model (SBM) Holland et al. (1980)

- Fit observed networks to parametric or non-parametric models of random graphs.
- **SBM** popular in applications: it allows to generate graphs with a community structure
  - Parameters:
    - Partition of n nodes into k disjoint groups $\{C_1, \dots, C_k\}$
    - Symmetric $k \times k$ matrix Q of inter-community edge probabilities.
  - Any two vertices $u \in C_i$ and $v \in C_j$ are connected with probability Q_{ij}.
  - **Regularity Lemma:** basic approximation units for more complex models.

Non-parametric Model

- SBM does not allow to analyze the fine structure of extremely large networks, in particular when the number of groups is growing.
- Non-parametric models of random graphs: **Graphon Model**
  - Graphons are symmetric measurable functions

$$W : [0,1]^2 \to [0,1].$$

  - Play a central role in the recent theory of graphs limits: every graph limit can be represented by a graphon.
  - Graphons give a natural way of generating random graphs.

Graphon Model

- **Graphon Model:**
  - $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ are latent i.i.d. uniformly distributed on $[0, 1]$.

$$\boldsymbol{\Theta}_{ij} = W_0(\xi_i, \xi_j).$$

  - The diagonal entries $\boldsymbol{\Theta}_{ii}$ are zero and $\boldsymbol{\Theta}_0 = (\boldsymbol{\Theta}_{ij})$
  - Given $\boldsymbol{\Theta}_0$ the graph is sampled according to the **inhomogeneous random graph model**:
    - ⋆ vertices i and j are connected by an edge with probability $\boldsymbol{\Theta}_{ij}$ independently from any other edge.
  - If W_0 is a step-function with k steps, the graph is distributed as a SBM with k groups.

Sparse Graphon Model

- The expected number of edges $\asymp n^2 \;\Rightarrow\;$ **dense** case.
- In real life networks often **sparse**
- **Sparse Graphon Model:**
  - Take $\rho_n > 0$ such that $\rho_n \to 0$ as $n \to \infty$.
  - The adjacency matrix $\mathbf{A}$ is sampled according to graphon W_0 with scaling parameter ρ_n:

$$\mathbf{\Theta}_{ij} = \rho_n W_0(\xi_i, \xi_j),\; i < j.$$

  - $\rho_n =$ "expected proportion of non-zero edges",
  - the number of edges is of the order $O(\rho_n n^2)$,
    - ★ $\rho_n = 1$ dense case
    - ★ $\rho_n = 1/n$ very sparse

Network Model

From a single observation of a graph

Problem 1:

Estimate the matrix of connection probabilities $\mathbf{\Theta}_0$

and

Problem 2:

Estimate the sparse graphon function $f_0(x, y) = \rho_n W_0(x, y)$

- We observe the $n \times n$ adjacency matrix $\mathbf{A} = (\mathbf{A}_{ij})$ of a graph
- $\mathbf{A}$ has been sampled according to the inhomogeneous random graph model with a fixed matrix $\mathbf{\Theta}_0$ or to the graphon model with graphon W_0
- Given a single observation $\mathbf{A}$, we want to estimate $\mathbf{\Theta}_0$ or f_0.

Graphon: invariance with respect to the change of labeling

- Graphon estimation is **more challenging** than probability matrix estimation
- Multiple graphons can lead to the same distribution on the space of graphs of size n.
- The topology of a network is **invariant with respect to any change of labeling** of its nodes
- We consider **equivalence classes** of graphons defining the same probability distribution on random graphs.

Loss function for graphon estimation

- Consider a sparse graphon $f(x,y) = \rho_n W(x,y)$
- $\tilde{f}(x,y)$ estimator of $f(x,y)$
- The squared error is defined by

$$\delta^2(f,\tilde{f}) := \inf_{\tau \in \mathcal{M}} \int \int_{(0,1)^2} |f(\tau(x),\tau(y)) - \tilde{f}(x,y)|^2 \mathrm{d}x\mathrm{d}y$$

$\mathcal{M}$ is the set of all measure-preserving bijections $\tau : [0,1] \to [0,1]$

Property (Lovász 2012)

$\delta(\cdot,\cdot)$ defines a metric on the quotient space $\mathcal{W}$ of graphons.

Minimax rate for sparse SBM in Frobenius norm

K., Tsybakov & Verzelen (2017)

$$\inf_{\widehat{\mathbf{\Theta}}} \sup_{\mathbf{\Theta}_0 \in \mathcal{T}[k,\rho_n]} \mathbb{E}_{\mathbf{\Theta}_0}\left[\frac{1}{n^2}\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\right\|_2^2\right] \asymp \min\left\{\rho_n\left(\frac{\log k}{n} + \frac{k^2}{n^2}\right), \rho_n^2\right\}$$

- $\rho_n = 1$: **Gao et al.(2014)**, the minimax rate over $\mathcal{T}[k,1]$

$$\frac{k^2}{n^2} + \frac{\log k}{n}$$

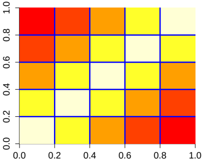
  - $k > \sqrt{n\log(k)}$: nonparametric rate $\dfrac{k^2}{n^2}$
  - $k < \sqrt{n\log(k)}$: clustering rate $\dfrac{\log k}{n}$

From probability matrix estimation to graphon estimation

- To any $n \times n$ probability matrix $\mathbf{\Theta}$ we can associate a graphon.
- Given a $n \times n$ matrix $\mathbf{\Theta}$ with entries in $[0, 1]$, define the **empirical graphon** $\widetilde{f}_{\mathbf{\Theta}}$ as the following piecewise constant function:

$$\widetilde{f}_{\mathbf{\Theta}}(x, y) = \mathbf{\Theta}_{\lceil nx \rceil, \lceil ny \rceil}$$

for all x and y in $(0, 1]$.



- This provides a way of deriving an estimator of the graphon function $f(\cdot, \cdot) = \rho_n W(\cdot, \cdot)$ from **any** estimator of the probability matrix $\mathbf{\Theta}_0$.

From probability matrix estimation to graphon estimation

- **Empirical graphon** $\widetilde{f}_{\mathbf{\Theta}}(x,y) = \mathbf{\Theta}_{\lceil nx\rceil,\lceil ny\rceil}$.
- For any estimator $\widehat{\mathbf{T}}$ of $\mathbf{\Theta}_0$:

$$E\left[\delta^2(\widetilde{f}_{\widehat{\mathbf{T}}}, f)\right] \le 2E\left[\frac{1}{n^2}\|\widehat{\mathbf{T}} - \mathbf{\Theta}_0\|_F^2\right] + 2\underbrace{E\left[\delta^2\left(\widetilde{f}_{\mathbf{\Theta}_0}, f\right)\right]}_{\textbf{agnostic error}}$$

(from the triangle inequality). Here, $\widetilde{f}_{\widehat{\mathbf{T}}}$ and $\widetilde{f}_{\mathbf{\Theta}_0}$ are empirical graphons.

Bound for the δ-risk of step-function graphon

Step function graphons: For some $k \times k$ symmetric matrix $\mathbf{Q}$ and some $\phi : [0,1] \to [k]$,

$$W(x,y) = \mathbf{Q}_{\phi(x),\phi(y)} \quad \text{for all } x, y \in [0,1] \; .$$

Theorem (K., Tsybakov and Verzelen, 2017)

Consider the ρ_n-sparse step-function graphon model W in $\mathcal{W}[k]$. The restricted LS empirical graphon estimator $\widehat{f}$ satisfies

$$E\left[\delta^2\left(\widehat{f}, f\right)\right] \leq C\left[\rho_n\left(\frac{k^2}{n^2} + \frac{\log(k)}{n}\right) + \rho_{\mathbf{n}}^{\mathbf{2}}\sqrt{\frac{\mathbf{k}}{\mathbf{n}}}\right].$$

Sparse network estimation problem

- The optimal rates can be achieved by the Least Squares Estimator
- But: it is not realizable in polynomial time
- Possible gap between the minimax optimal rate and the best rate achievable by computationally feasible methods?
- **Hard thresholding estimator**

Hard thresholding estimator

- **Achieves the best known rate in Frobenius distance in the class of polynomial-time estimators**
- Singular value decomposition of $\mathbf{A}$:

$$\mathbf{A} = \sum_{j=1}^{\mathrm{rank}(\mathbf{A})} \sigma_j(\mathbf{A}) u_j(\mathbf{A}) v_j(\mathbf{A})^T$$

- Tuning parameter $\lambda > 0$:

$$\widetilde{\mathbf{\Theta}}_\lambda = \sum_{j:\sigma_j(\mathbf{A}) \geq \lambda} \sigma_j(\mathbf{A}) u_j(\mathbf{A}) v_j(\mathbf{A})^T$$

Singular value hard thresholding estimator of Θ_0.

Hard thresholding estimator for sparse SBM

Theorem (K. & Verzelen, 2018)

With high probability

$$\frac{1}{n}\|\widetilde{\mathbf{\Theta}}_\lambda - \mathbf{\Theta}_0\|_2 \;\; \leq \;\; C\sqrt{\frac{\rho_n \mathbf{k}}{n}} \; ,$$

where C is a numerical constant.

- Also minimax optimal in the **cut distance**

- **Cut distance:**
  - Two random graphs with the same edge density are close
  - Reflects global and local structural similarities
  - Cornerstone in the limit graphs theory (**Lovász and Szegedy (2004), Borgs et al (2008), (2012)**):
    - ★ Every graph limit can be represented by a **graphon**
    - ★ A sequence $(\mathcal{G}_n)$ of simple graphs is convergent if and only if it is a Cauchy sequence in the **cut metric**.
- Estimating well the graphon W_0 in the cut distance allows to estimate well the number of small patterns induced by W_0

Matrix cut norm

Matrix cut norm (Frieze and Kannan (1999)):

Matrix $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{n \times n}$

$$\|\mathbf{A}\|_{\square} = \frac{1}{n^2} \max_{S,T \subset [n]} \left| \sum_{i \in S, j \in T} A_{ij} \right|$$

- $S = T$, $S \cap T = \emptyset$ or $T = \bar{S}$
- $$\|\mathbf{A}\|_{\square} \leq \frac{1}{n^2} \|\mathbf{A}\|_1 \leq \frac{1}{n} \|\mathbf{A}\|_2$$

where $\|\mathbf{A}\|_1 = \sum_{i,j} |A_{ij}|$ and $\|\mathbf{A}\|_2 = \sqrt{\sum_{i,j} A_{ij}^2}$

Cut norm of graphons

Cut norm of graphons

$$\|W\|_\square = \sup_{S,T\subset[0,1]} \left| \int_{S\times T} W(x,y)\mathrm{d}x\mathrm{d}y \right|$$

- S and T measurable subsets
- $S = T$, $S \cap T = \emptyset$ or $T = \bar{S}$
- $\|W\|_\square \leq \|W\|_1 \leq \|W\|_2 \leq \|W\|_\infty \leq 1$

Minimax rate for sparse SBM in cut norm K. & Verzelen, 2018

$$\inf_{\widehat{\mathbf{\Theta}}} \sup_{\mathbf{\Theta}_0 \in \mathcal{T}[k,\rho_n]} \mathbb{E}_{\mathbf{\Theta}_0}\left[\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\right\|_{\square}\right] \asymp \min\left(\sqrt{\frac{\rho_n}{n}}, \rho_n\right)$$

- Faster than the minimax rate of convergence in Frobenius norm:

$$\inf_{\widehat{\mathbf{\Theta}}} \sup_{\mathbf{\Theta}_0 \in \mathcal{T}[k,\rho_n]} \mathbb{E}_{\mathbf{\Theta}_0}\left[\frac{1}{n}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_2\right] \asymp \min\left\{\left(\sqrt{\frac{\rho_n \log k}{n}} + \frac{\sqrt{\rho_n} k}{n}\right), \rho_n\right\}$$

  - **Few blocks** $k \lesssim \sqrt{n}$: gain of $\log(k)$ factor
  - **Large** $k \gtrsim \sqrt{n}$: gain of $k/\sqrt{n}$ factor

Graphon estimation problem: step-function graphon

Thresholding empirical graphon estimator

$$\mathbb{E}_W \left[\delta_{\square} \left(\widetilde{f}_{\widetilde{\mathbf{\Theta}}_\lambda}, f_0 \right) \right] \leq C \left(\rho_n \sqrt{\frac{k}{n \log(k)}} + \sqrt{\frac{\rho_n}{n}} \right)$$

- Empirical graphon associated to the hard thresholding estimator is minimax optimal in the cut-distance.
- Achieves best known convergence rates with respect to δ_1 and δ_2-distance among polynomial time algorithms.

Link Prediction

Link prediction

- Networks are often **incomplete**: detecting interactions can require significant experimental effort
- Replace exhaustive testing for every connection by deducing the pairs of nodes which are most likely to interact
- Predict the probabilities of connections from partial observation of the graph

Maximum Likelihood Estimator

- **Wolfe and Olhede (2013)**, **Bickel et al (2013)**, **Amini et al (2013)**, **Celisse et al (2012)** , **Tabouy et al (2017)** ...
- Also NP hard ...
- Computationally efficient approximations:
  - Pseudo-likelihood methods
  - Variational approximation
- Quite successful in practice

Is MLE minimax optimal?

Convergence rate for the MLE

The conditional log-likelihood:

$$\mathcal{L}(\mathbf{A};\mathbf{\Theta}) = \sum_{i<j} \mathbf{A}_{ij}\log(\mathbf{\Theta}_{ij}) + (1-\mathbf{A}_{ij})\log(1-\mathbf{\Theta}_{ij})$$

Theorem (Gaucher & K., 2019)

With high probability

$$\|\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}}_{ML}\|_2^2 \leq C\rho_n\left(\mathcal{K}(\mathbf{\Theta}_0,\widetilde{\mathbf{\Theta}}) + \frac{\rho_n^2}{(1-\rho_n)^2\wedge\gamma_n^2}\left(k^2+n\log(k)\right)\right).$$

- $0<\gamma_n\leq(\mathbf{\Theta}_0)_{ij}\leq\rho_n<1$
- $\widetilde{\mathbf{\Theta}}$ the best approximation among SBM to $\mathbf{\Theta}_0$ in the sense of the Kullback Leibler divergence
- **Minimax optimal** if $\gamma_n \asymp \rho_n$

Partial observations of the network

- $\mathbf{X} \in \{0,1\}^{n\times n}_{sym}$ **the sampling matrix**:

  $\mathbf{X}_{ij} = 1$ if we observe $\mathbf{A}_{ij}$ and $\mathbf{X}_{ij} = 0$ otherwise
- Conditionally on $\mathbf{\Theta}_0$, $\mathbf{X}$ is independent from the adjacency matrix $\mathbf{A}$
- $\mathbf{X}_{ij}$ are mutually independent
- $\mathbf{\Pi} \in [0,1]^{n\times n}_{sym}$ the **matrix of sampling probabilities**:

  $\mathbf{X}_{ij} \overset{ind.}{\sim}$ Bernoulli$(\mathbf{\Pi}_{ij})$

Partial observations of the network

Particular cases:

- node-based sampling schemes (e.g. the exo-centered design)
- random dyad sampling schemes
- ...

MLE with missing observations

The conditional log-likelihood:

$$\mathcal{L}_{\mathbf{X}}(\mathbf{A};\boldsymbol{\Theta}) = \sum_{i<j} \mathbf{X}_{ij}\left(\mathbf{A}_{ij}\log(\boldsymbol{\Theta}_{ij}) + (1-\mathbf{A}_{ij})\log(1-\boldsymbol{\Theta}_{ij})\right).$$

Theorem (Gaucher & K., 2019)

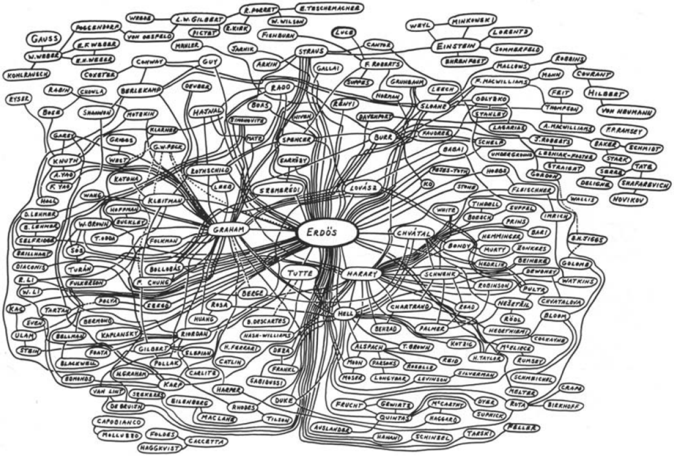
With high probability

$$\|\boldsymbol{\Theta}_0 - \widehat{\boldsymbol{\Theta}}\|^2_{2,\boldsymbol{\Pi}} \leq C'\rho_n\left(\mathcal{K}_{\boldsymbol{\Pi}}(\boldsymbol{\Theta}_0, \widetilde{\boldsymbol{\Theta}}) + \frac{\rho_n^2}{(1-\rho_n)^2 \wedge \gamma_n^2}\left(k^2 + n\log(k)\right)\right).$$

- $\|\boldsymbol{\Theta}_0 - \widehat{\boldsymbol{\Theta}}\|^2_{2,\boldsymbol{\Pi}} = \sum_{ij} \boldsymbol{\Pi}_{ij}(\boldsymbol{\Theta}_0 - \widehat{\boldsymbol{\Theta}})^2_{ij}$
- **Minimax optimal** if $c_1 p \leq \boldsymbol{\Pi}_{ij} \leq c_2 p$ [**Gao et al, 2016**]

Conclusion

- **Least Squares Estimator**:
  - attains the optimal rates in a minimax sense,
  - not realizable in polynomial time
- better choice: **Thresholding estimator** (slower rates of convergence)
- **MLE:**
  - minimax optimal
  - has computationally efficient approximations
- **Link Prediction:**
  - MLE: enables rank unobserved pairs of nodes
  - Minimax optimality of this approach
  - Works for quite general sampling schemes


ERDÖS
GRAHAM
HARARY
TUTTE
CHVÁTAL
GAUSS
EINSTEIN
MINKOWSKI
LORENTZ
SOMMERFELD
WEYL
HILBERT
COURANT
VON NEUMANN
F.P.RAMSEY
STRAUS
CONWAY
GUY
BERLEKAMP
KNUTH
KLEITMAN
RADO
SPENCER
SZEMERÉDI
LOVÁSZ
BURR
SLOANE
RÉNYI
HAJNAL
KATONA
ROTHSCHILD
F. CHUNG
BOLLOBÁS
BERGE
HELL
CHARTRAND
SCHWENK
BONDY
ROBINSON
ALSPACH
KARP
DUKE
ULAM
STEIN
KAC
TURÁN
POLYA
SÓS
W.WEBER
KOHLRAUSCH
EYSER
SHANNON
CANTOR
FISHBURN
GALLAI
TARSKI
FELLER
BIRKHOFF
ROTA
QUINTAS
FRUCHT
GEWIRTZ
HAGGARD
MACLANE
RHODES
TILSON
HARPER
CAPOBIANCO
MOLLUZZO
FOLDES
CACCETTA
HAGGKVIST
NOVIKOV
SHAFAREVICH
DELIGNE
TATE
SCHMIDT
BAKER
BLOOM
GOLOMB
WATKINS
IMRICH
PRINS
BARI
HEMMINGER
SCHELP
BABAI
HOBBS
FLEISCHNER
GORDON
STANLEY
ODLYZKO
LEECH
NORMAN
DAVENPORT
SUPPES
F. ROBERTS
GRUNBAUM
MALLOWS
F. MACWILLIAMS
THOMPSON
FEIT
MANN
ROBBINS
A.MACWILLIAMS
J. ROBERTS
STARK
SERRE
WALLIS
COXETER
GARDNER
CHOWLA
RABIN
BOES
WANG
HALL
D.LEHMER
E.LEHMER
DIACONIS
W. LI
R. LI
EVEN
BELLMAN
KAPLANSKY
FOATA
GILBERT
POLLAK
M.GRAHAM
VAN LINT
DE BRUIJN
EILENBERG
HUANG
ROSA
RIORDAN
CATLIN
CARLITZ
SABIDUSSI
FRANKL
DEZA
B.DESCARTES
PALMER
BEHZAD
T.BROWN
KOTZIG
REID
H.TAYLOR
SILVERMAN
MOSER
MOON
PARSONS
LONGYEAR
LEVINSON
SCHMEICHEL
MELTER
CRAPO
SUPNICK
DYER
MCCARTHY
AUSLANDER
HANANI
SCHINZEL
NEŠETŘIL
RÖDL
READ
PULTR
CHVATALOVA
DEWDNEY
HEDRLIN
BEINEKE
HARTY
BONDY
STONE
WHITE
TINDELL
SUFFEL
KO
BOESCH
SIMONOVITS
NIVEN
SARKÖZY
HAJNAL
GRIGGS
WEST
LEEB
HOFFMAN
BUCKLEY
W.BROWN
T.ODDA
FOLKMAN
FULKERSON
TARTAN
KLARNER
MOTZKIN
JARNIK
ARKIN
MAHLER
PICTET
E.KIRK
W.WILSON
R.PORRET
E.TESCHEMACHER
L.W.GILBERT
WEBER
POGGENDORF
VON OESFELD
E.F.WEBER
E.H.WEBER
LUCE
EHRENFEST
A.YAO
F.YAO
GAREY
SELFRIDGE
BRILLHART
DOWLING
COLEMAN
HEDETNIEMI
M^C.ELIECE
RUMSEY
COCKAYNE

Thank You !