Higher-order fluctuations in dense random graph models: Statistical aspects

Adrian Röllin

National University of Singapore

UK Easter Probability Meeting,
March 2023, Manchester

joint work with Gursharn Kaur and Li Shang

Dense random graphs

Subgraph densities

$G_1, G_2, \ldots$: dense graph sequence.

F: finite simple graph on k vertices.

Subgraph density of F in G_n:

$$t_F(G_n) := \frac{\#\text{ injective homomorphisms of } F \text{ into } G_n}{n(n-1)\cdots(n-k+1)}$$

Inhomogeneous Erdős-Rényi random graph

Fix graphon $\kappa \colon [0,1]^2 \to [0,1]$.

$U = (U_1, \ldots, U_n)$: i.i.d. uniform on $[0,1]$.

Connect vertices i and j with probability $\kappa(U_i, U_j)$.

Denote this graph by $G(n, \kappa)$.

Law of Large Numbers

Lovász and Szegedy (2006)

Theorem. *Let $G_n \sim G(n, \kappa)$ for all n. Then, almost surely, G_n is a dense graph sequence and, almost surely, G_n converges to κ in the metric space of graphons; that is,*

$$t_F(G_n) \xrightarrow[a.s.]{} \mathbb{E} \prod_{i \overset{F}{\sim} j} \kappa(U_i, U_j)$$

Fluctuations of subgraph densities

or “What is the CLT of dense graph limit theory?”

A discouraging observation

The key quantities $t_F(G_n)$ tell us virtually nothing about the fluctuations.

Let $G_n \sim G(n,p)$, for p fixed.

Then, for any F,

$$\lim_{n\to\infty} \mathrm{Cor}(t_F(G_n), t_{\bullet\!-\!\bullet}(G_n)) = 1$$

The dominant fluctuations of subgraph densities are determined by $t_{\bullet\!-\!\bullet}(G_n)$.

And for general graphons...

In $G(n, \kappa)$, the $t_F(G_n)$ are in general dominated by sums of the form $\sum_{i=1}^{n} g_{F,\kappa}(U_i)$.

Fluctuations of subgraph densities are dominated by vertex labels (in general), and all information about randomness of edges is lost in the limit.

Finer fluctuations

Janson and Nowicki (1991), Janson (1997).

Generalised U-statistics:

$$\sum_{1\leqslant a_1<\cdots<a_k\leqslant n} g\big(U_{a_1},\ldots,U_{a_k};Y_{a_1a_2},\ldots,Y_{a_{k-1}a_k}\big)$$

where U_i are i.i.d. and Y_{ij} are i.i.d.

The key result: such statistic allow a Hoeffding-type decomposition, but it's more complicated than for regular U-statistics.

Mind you: The classical CLT does not exhibit “finer fluctuations”.

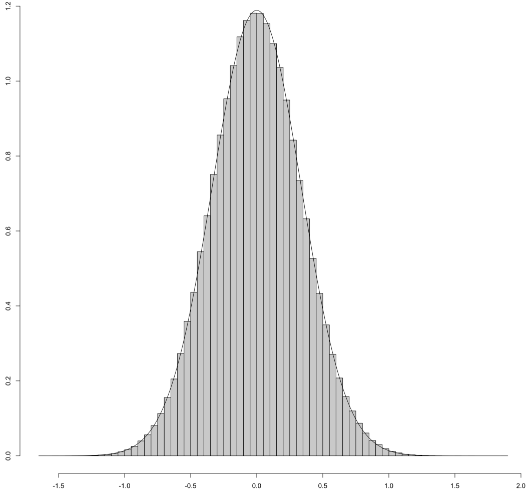
Consider $G(n,\kappa)$ with $\kappa = \begin{array}{|c|c|}\hline \delta & \beta \\ \hline \alpha & \delta \\ \hline \end{array}$

γ

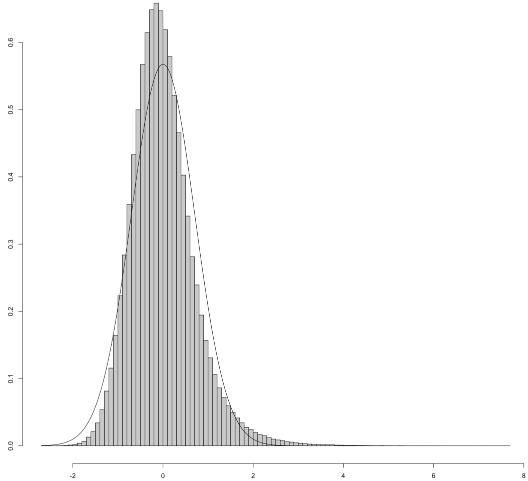
$$t_{\bullet\!-\!\bullet}(G_n) = \mathbb{E}\kappa(U_1,U_2) + \frac{2\rho_1 n^{1/2} V_\bullet}{(n-1)} + \frac{\rho_2\left(V_\bullet^2 - \gamma(1-\gamma)\right)}{n-1} + \frac{2^{1/2} V_{\bullet\!-\!\bullet}}{n^{1/2}(n-1)^{1/2}} + \frac{(\beta-\alpha)V_\bullet}{n^{1/2}(n-1)},$$

where $\rho_1 = \alpha\gamma - \beta(1-\gamma) + (1-2\gamma)\delta$ and $\rho_2 = \alpha + \beta - 2\delta$.

$$V_\bullet = n^{-1/2}\sum_i \left(\mathrm{I}[U_i \leqslant \gamma] - \gamma\right), \quad V_{\bullet\!-\!\bullet} = \binom{n}{2}^{-1/2}\sum_{i_1<i_2}\left(Y_{i_1 i_2} - \kappa(U_{i_1},U_{i_2})\right).$$



$n = 10{,}000$, $\alpha = \beta = 0.8$, $\delta = 0.1$, $\gamma = 0.2$, $\rho_1 = -0.42$.



$n = 10{,}000$, $\alpha = \beta = 0.8$, $\delta = 0.1$, $\gamma = 0.5$, $\rho_1 = 0$.

$$t_{\triangle}(G_n) = R_{0.0} + R_{0.5} + R_{1.0} + R_{1.5} + R_{2.0} + R_{2.5}$$

$$R_{0.0} = \mathbb{E}\, t_{\triangle}(G_n)$$

$$R_{0.5} = \frac{c_2 V_{\bullet}}{n^{1/2}}, \quad R_{1.0} = \frac{c_3 (V_{\bullet}^{\,2} - \gamma(1-\gamma))}{n} + \cdots ,$$

$$R_{1.5} = \cdots + \frac{c_4 V_{\triangle} + c_5 V_{\vee,1} + c_6 V_{\diagup,1} V_{\bullet} + c_7 V_{\diagup,2} V_{\bullet}}{n^{3/2}},$$

$$R_{2.0} = \cdots , \quad R_{2.5} = \cdots .$$

$$V_{\bullet} = n^{-1/2} \sum_{i} \hat{Z}_i, \qquad V_{\diagup,1} = \binom{n}{2}^{-1/2} \sum_{i<j} \hat{Y}_{ij},$$

$$V_{\vee,1} = \binom{n}{3}^{-1/2} \sum_{i<j<k} \kappa(U_i, U_k) \hat{Y}_{ij} \hat{Y}_{jk}, \quad V_{\triangle} = \binom{n}{3}^{-1/2} \sum_{i<j<k} \hat{Y}_{ij} \hat{Y}_{jk} \hat{Y}_{ik}, \quad \ldots$$

$$\hat{Z}_i = \mathrm{I}[U_i \leqslant \gamma] - \gamma, \qquad \hat{Y}_{ij} = Y_{ij} - \kappa(U_i, U_j)$$

Centred subgraph counts

We propose to use

$$T_F(G_n) = \binom{n}{k}^{-1/2} \sum_{a_1 < \cdots < a_k} \prod_{i \overset{F}{\sim} j} \left(Y_{a_i a_j} - \kappa(U_{a_i}, U_{a_j}) \right),$$

as fundamental local graph statistics.

Janson & Nowicki/Kaur & R.: For any collection of graphs $F_1, \ldots, F_d$, the statistics $T_{F_1}, \ldots, T_{F_d}$ are jointly close to a multivariate Gaussian law.

Statistical Applications

Test statistics

Family of uncorrelated test statistics:

$$Z_F(G_n) = \frac{\sum_{a_1 < \cdots < a_k} \prod_{i \overset{F}{\sim} j} \left(Y_{a_i a_j} - p_{a_i a_j} \right)}{\left(\sum_{a_1 < \cdots < a_k} \prod_{i \overset{F}{\sim} j} p_{a_i a_j} \left(1 - p_{a_i a_j} \right) \right)^{1/2}},$$

where p_{ij} are the hypothesised edge probabilities.

Choices of F determines what is being tested.

Test statistics

In practice, p_{ij} will be replaced by some estimates $\hat{p}_{ij} = \hat{p}_{ij}(G_n)$, which come from fitting a particular random graph model.

Hence, we consider instead

$$\hat{Z}_F(G_n) = \frac{\displaystyle\sum_{a_1 < \cdots < a_k} \prod_{i \overset{F}{\sim} j} \left(Y_{a_i a_j} - \hat{p}_{a_i a_j} \right)}{\left(\displaystyle\sum_{a_1 < \cdots < a_k} \prod_{i \overset{F}{\sim} j} \hat{p}_{a_i a_j} \left(1 - \hat{p}_{a_i a_j} \right) \right)^{1/2}},$$

Interpretation

$\hat{Z}_{\diagup}(G_n)$: Test total number of edges against expected number of edges.

$\hat{Z}_{\vee}(G_n)$: Test pairwise dependence; large pos. value → increased simultaneous presence or absence of adjacent edges.

$\hat{Z}_{\triangle}(G_n)$: Large pos. values → increased simultaneous presence triangles or “one on, two off” configurations; this means, presence of one edge suppresses or encourages presence of other two edges simultaneously.

Simulation study

$\kappa =$ stochastic block model with 4 groups.

Connection probabilities given by

$$K = \begin{pmatrix} 0.45 & 0.34 & 0.82 & 0.60 \\ 0.34 & 0.70 & 0.98 & 0.57 \\ 0.82 & 0.98 & 0.03 & 0.82 \\ 0.60 & 0.57 & 0.82 & 0.25 \end{pmatrix}$$

$n = 200$ vertices.

Reconstruction of community labels and estimation of connection probabilities done via off the shelf Variational EM algorithm, R package `blockmodels`.

Simulation study

Result based on one realisation of the network.

	Number of groups							
	1	2	3	4	5	6	7	8
	Data simulated from 4 × 4 stochastic block model							
$\hat{z}$	0.00	0.01	0.00	**0.17**	0.08	−0.07	0.01	−0.02
$\hat{z}$	4.65	2.47	2.27	**−0.57**	−0.51	−0.73	0.16	−1.30
$\hat{z}$	−18.57	−0.38	1.04	**0.03**	0.22	0.15	−0.23	−0.11
$\hat{z}$	57.36	2.43	0.77	**−0.24**	−0.07	−0.04	−0.33	−0.33
$\hat{z}$	−5.39	2.31	2.62	**−0.93**	−0.56	−0.39	−0.65	−0.36
$\hat{z}$	1.90	2.42	1.55	**1.29**	0.27	0.83	1.61	0.14

Hospital encounter network ‘rfid’

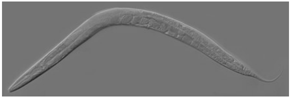
Contacts among patients and health care workers in a hospital unit in Lyon over the course of four days in 2010.

75 participants consented to wear RFID sensors, which recorded when any two of them were in face-to-face contact with each other during a 20-second interval of time.

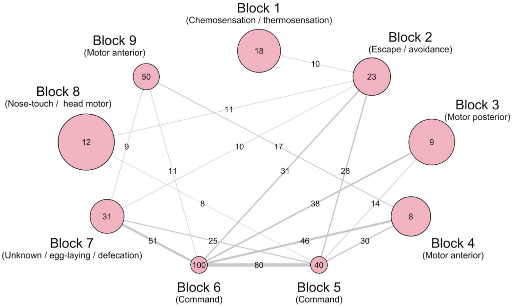
	Number of groups							
	1	2	3	4	5	6	7	8
	Data simulated from 4 × 4 stochastic block model							
$\hat{z}_{/}$	0.00	−0.31	0.01	−0.33	**0.05**	−0.01	0.13	−0.07
$\hat{z}_{\vee}$	71.48	19.49	8.72	9.32	**9.87**	6.25	1.02	−0.15
$\hat{z}_{\triangle}$	11.32	1.49	1.53	4.40	**7.96**	8.24	8.20	8.25
$\hat{z}_{\square}$	136.27	30.38	22.25	17.06	**13.43**	13.15	12.31	12.44
$\hat{z}_{\sqcup}$	44.32	1.11	−1.74	−1.85	**−0.41**	0.50	−1.24	−1.32
$\hat{z}_{\curlywedge}$	44.23	−3.36	4.55	5.87	**7.85**	2.97	−0.52	0.28

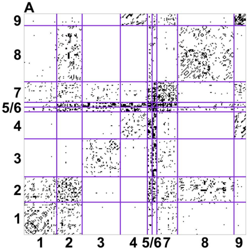
Connectome of *C. elegans*

Adult C. elegans hermaphrodite has 959 somatic cells, out of which 302 are neurons. 279 used for analysis.



Pavlovic, D. M., et al. “Stochastic blockmodeling of the modules and core of the Caenorhabditis elegans connectome” PloS one 9.7 (2014):
V-EM algorithm, `mixer` (defunct), `blockmodels`





Gr'ps	$\hat{z}$	$\hat{z}$	$\hat{z}$	$\hat{z}$	$\hat{z}$	$\hat{z}$	$\hat{z}$	$\hat{z}$	$\hat{z}$	$\hat{z}$
1	107.90	87.45	298.94	107.67	614.02	577.94	797.81	676.12	1401.06	4822.94
2	39.07	56.51	142.80	-13.41	97.23	82.91	180.43	278.91	167.30	425.86
3	54.91	45.28	113.43	39.59	129.04	97.85	149.08	163.93	363.85	658.33
4	29.21	40.19	83.29	14.98	24.24	41.76	99.89	108.12	89.12	55.71
5	29.24	35.25	72.93	14.77	25.67	37.08	77.93	87.20	90.57	58.21
6	19.29	31.22	58.30	-2.32	11.88	17.15	61.57	59.07	26.20	34.28
7	18.16	28.40	53.84	-3.05	11.70	14.32	48.51	51.34	21.14	33.75
8	16.68	26.79	50.70	-2.17	15.41	13.70	42.04	43.71	17.76	36.60
9	**16.36**	**26.50**	**44.90**	**-1.04**	**16.18**	**13.10**	**40.60**	**40.46**	**20.68**	**40.03**
10	13.89	26.07	45.05	-6.91	16.55	6.96	37.62	47.24	13.03	39.56
11	13.82	25.88	43.85	-4.86	11.51	5.91	34.82	44.99	16.83	34.01
12	14.44	23.73	38.94	-4.11	9.72	3.88	29.26	33.22	16.70	33.15
13	14.18	24.17	38.84	-1.35	10.73	4.44	27.87	34.28	16.41	35.60
14	14.05	23.49	38.20	-0.19	10.89	3.92	26.43	33.79	16.83	35.67
15	7.26	23.98	38.22	-0.30	1.62	5.96	24.96	35.25	5.02	2.18
16	6.05	22.84	36.47	-1.04	0.67	5.29	24.01	32.76	2.55	-0.24
17	6.27	22.80	36.00	-1.19	0.71	6.05	24.32	33.16	3.15	-0.22
18	6.37	22.58	33.82	-1.24	-1.20	6.94	25.44	33.23	1.95	1.08
19	6.42	21.90	32.43	-1.72	-1.00	6.99	23.59	29.95	1.83	1.22
20	6.23	21.20	31.41	-1.37	-0.97	6.61	22.35	28.15	1.93	1.06

Some Theoretical Properties

MLE of connection probabilities

Consider $k \times k$ stochastic block model and assume community labels are known.

Let $\hat{p}_{ij}$ be the MLE estimator of the edge densities between communities:

$$\hat{p}_{ij} = \frac{\#\text{edges from } i\text{'s community to } j\text{'s community}}{\text{size of } i\text{'s community} \times \text{size of } j\text{'s community}}$$

For MLE, we always have $\hat{z}_{\bullet\!-\!\bullet} = 0$

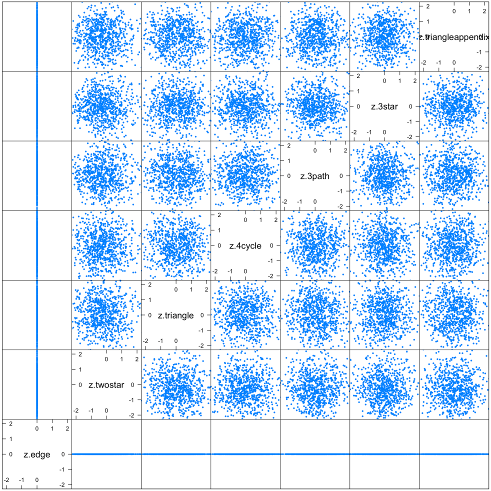
Behaviour of centred subgraph statistics

For fixed F with $v(F)$ vertices and $e(F)$ edges,

$$\mathbb{E}\hat{Z}_F = \mathrm{O}\left(n^{v(F)/2-2\cdot\lceil e(F)/2\rceil}\right) \qquad \text{as } n \to \infty.$$

Worst case is 2-star: $\mathbb{E}\hat{Z}_{\vee} = \mathrm{O}(n^{-1/2})$.

Li & R.: In the dense regime, the $\hat{Z}_F$ are close to a multivariate Gaussian law.



	$\hat{z}_{\diagup}$	$\hat{z}_{\vee}$	$\hat{z}_{\triangle}$	$\hat{z}_{\square}$	$\hat{z}_{\sqcup}$	$\hat{z}_{\curlywedge}$	$\hat{z}_{\triangle-}$
Mean	0.00	-0.25	0.03	-0.02	0.06	-0.03	-0.03

Spectral clustering + MLE

Model

Core-periphery structure, four groups of vertices, (two cores, two peripheries):

$$K = \begin{pmatrix} 0.8 & 0.5 & 0.1 & 0.1 \\ 0.5 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.8 & 0.5 \\ 0.1 & 0.1 & 0.5 & 0.1 \end{pmatrix}$$

We consider 3 regimes:

dense	interm.	sparse
K	$K/n^{0.3}$	$K/n^{0.6}$

Test statistic

Test statistic: $\chi^2 = \hat{z}^2_{\triangle} + \hat{z}^2_{\square} + \hat{z}^2_{\sqcup} + \hat{z}^2_{\curlywedge} + \hat{z}^2_{\triangle\!-}$

Under correct number of groups, perfect classification and MLE estimates, χ^2 is approximately χ^2_5 distributed.

Use this to calculate *p*-values and compare distribution of the *p*-values against uniform distribution on $[0, 1]$.

Correct classifications

Fraction of correctly classified labels by spectral clustering.

	n=200			n=400			n=800		
	dense	interm.	sparse	dense	interm.	sparse	dense	interm.	sparse
avg. cor. classified	0.84	0.51	0.33	0.96	0.52	0.33	0.99	0.52	0.37

Distribution of p-values

L_1 distance between uniform distribution on $[0,1]$ and empirical CDF of p-values

	n=200			n=400			n=800		
	dense	interm.	sparse	dense	interm.	sparse	dense	interm.	sparse
labels and K known	0.04	0.04	0.10	0.02	0.02	0.03	0.02	0.02	0.04
labels known + MLE	0.06	0.08	0.14	0.01	0.02	0.03	0.02	0.01	0.06
spect. clust. + MLE	1.00	0.98	0.53	1.00	1.00	0.77	0.88	1.00	0.90

dense, $n = 800$, two group core-periphery structure

p	corr. classified	$\hat{z}_{\vee}$	$\hat{z}_{\triangle}$	$\hat{z}_{\square}$	$\hat{z}_{\sqcup}$	$\hat{z}_{\curlywedge}$	$\hat{z}_{\triangle\!-}$	$\hat{z}_{\times}$
0.871	1.0000	0.46	-0.42	-0.77	0.37	0.97	1.01	0.24
0.000	0.9988	11.97	1.22	0.59	1.25	145.22	-0.97	1538.29
0.484	1.0000	0.66	0.47	-1.08	1.12	0.19	-0.05	1.84
0.000	0.9975	23.91	-0.61	-1.33	-1.96	277.36	-0.61	2932.38
0.852	1.0000	-0.38	-0.61	-0.28	-0.34	0.02	0.75	1.44
0.192	1.0000	-1.75	-1.70	0.97	1.26	1.02	-0.62	0.08
0.000	0.9988	11.39	-0.18	-0.66	0.71	133.22	0.58	1396.47
0.197	1.0000	-0.31	-1.52	-0.85	-1.78	-1.70	-0.40	-0.72
0.807	1.0000	1.28	0.28	-0.71	-0.26	-1.15	-0.24	-0.30
0.860	1.0000	0.07	-1.21	0.63	-0.74	-0.61	0.65	-0.24

dense, $n = 200$, two group core-periphery structure

corr. classified	$\hat{z}$	$\hat{z}$	$\hat{z}$	$\hat{z}$	$\hat{z}$	$\hat{z}$	$\hat{z}$
0.970	35.83	-0.00	6.03	-1.04	46.15	12.67	1203.28
0.900	95.38	-0.16	61.52	-6.92	125.62	108.41	2069.49
0.960	40.37	0.40	10.40	-4.73	70.59	15.50	1138.57
0.915	86.63	-1.49	50.49	-17.70	187.85	94.30	2021.23
0.890	100.44	-0.80	68.64	-12.83	205.23	131.99	2016.90
0.935	61.07	0.75	25.43	2.10	116.07	44.88	1508.06
0.960	35.28	0.72	7.40	-1.40	171.52	15.94	783.13
0.945	50.42	-2.16	15.09	-14.07	179.38	25.50	1168.27
0.920	77.78	-2.03	39.29	-16.84	57.96	87.04	1871.28
0.955	52.54	1.34	16.47	9.72	-104.89	30.93	1657.35

Using $\hat{z}_F$ for clustering

k	χ^2 spect. clust.	χ^2 subg. dens.
1	1,079,327	1,079,327
2	61,715	59,757
3	29,734	6,614
4	8,028	16.7
5	25,761	11.2
6	5,645	10.9
7	28,468	2.84
8	52,768	0.82
9	104,240	1.31
10	16,738	0.74

Correctly classified by spect. clustering: 83%

Correctly classified by centr. subg. densities: 96%

Pros and cons

Pros:

- Summands of Z_F are uncorrelated.
- If $F \neq F'$, then Z_F and $Z_{F'}$ are uncorrelated.
- Covariance structure is very simple and can be easily estimated.
- Can be use for actual statistical testing, e.g. goodness-of-fit.
- Can be used for clustering.

Cons:

- Not parameter-free; in practice, need to substitute $\kappa(U_i, U_j)$ by $\hat{p}_{ij}$.
- Calculating Z_F can be computationally more expensive than calculating t_F.
- Interpretation is not as straightforward as for t_F.

Thank You!

G. Kaur and A. Röllin (2021). Higher-order fluctuations in dense random graph models. *Electron. J. Probab.* **26**, article no. 139.

S. Janson (1997). *Gaussian Hilbert Spaces*. Cambridge University Press.

S. Li and A. Röllin (in preparation). Statistical properties of centred subgraph counts in network analysis.