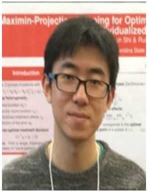
Goodness-of-Fit and the Best Approximation: an Adversarial Approach

Qiwei Yao

London School of Economics, q.yao@lse.ac.uk

Jinyuan Chang
SWUFE, Chengdu



Chengchun Shi
LSE

Mingcong Wu
SWUFE, Chengdu

Xinyang Yu
LSE

Outline

Assessing goodness-of-fit, or selecting a relevant model for networkdata analysis

Li, Levina and Zhu (2020). Network cross-validation by edge sampling. *Biometrika*, pp.257-

Jin, Ke, Tang and Wang (2025). Network goodness-of-fit for block-model family.

Kaji, Manresa and Pouliot (2023). An adversarial approach to structural estimation. *Econometrica*, pp. 2041-

Motivating example: Transitivity Model (Chang et al. 2024)

Let $\mathbf{X}_t = (X_{i,j}^t)$ denote the adjacency matrix at time t:

$$P(X_{i,j}^t = 1 | X_{i,j}^{t-1} = 0, \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \cdots) = \xi_i \xi_j \frac{e^{aU_{i,j}^{t-1}}}{1 + e^{aU_{i,j}^{t-1}} + e^{bV_{i,j}^{t-1}}},$$

$$P(X_{i,j}^t = 0 | X_{i,j}^{t-1} = 1, \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \cdots) = \eta_i \eta_j \frac{e^{bV_{i,j}^{t-1}}}{1 + e^{aU_{i,j}^{t-1}} + e^{bV_{i,j}^{t-1}}},$$

where $U_{i,j}^{t-1} = \sum_k X_{i,k}^{t-1} X_{j,k}^{t-1}$ is no. of common friends of nodes i and j at time $t-1$ — used by Facebook and LinkedIn,

$V_{i,j}^{t-1} = \sum_k \{X_{i,k}^{t-1}(1 - X_{j,k}^{t-1}) + (1 - X_{i,k}^{t-1}) X_{j,k}^{t-1}\}/2$ is a distance measure bwt nodes i and j,

ξ_i, η_i, a and b are non-negative parameters.

Real data example: Email interactions

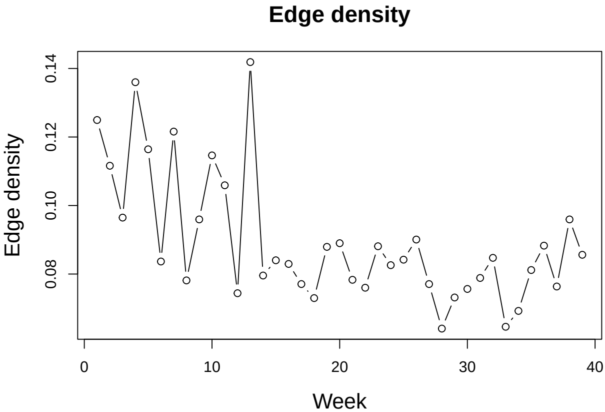
The email interactions in a medium-sized Polish manufacturing company in January – September 2010 (Michalski et al., 2014)

Consider $p = 106$ of the most active participants out of an original 167 employees

$n = 39$ represents 39 weeks, and $X_{i,j}^t = 1$ if participants i and j exchanged at least one email during Week t.

To gain some insight, we first present some preliminary summaries of the data.

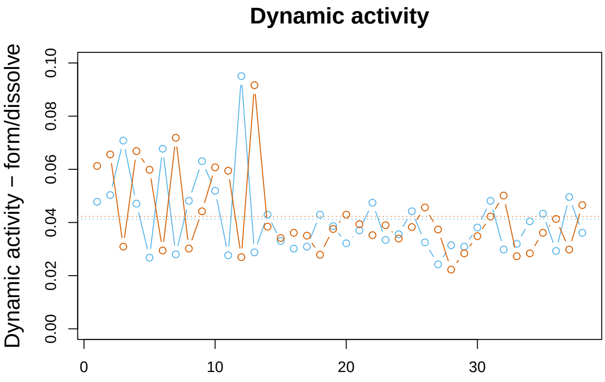
Edge density



Plot of percentage of edges $D_t = \frac{2}{p(p-1)} \sum_{1 \le i < j \le p} X_{i,j}^t$ against t.

A change-point at $t = 14$: Period 1 – first 13 points, Period 2 – last 26 points

Densities of newly formed edges, and newly dissolved edges



Plot of percentage of grown $D_{1,t} = \frac{2}{p(p-1)} \sum_{1 \le i < j \le p} (1 - X_{i,j}^{t-1}) X_{i,j}^{t}$ and dissolved $D_{0,t} = \frac{2}{p(p-1)} \sum_{1 \le i < j \le p} X_{i,j}^{t-1} (1 - X_{i,j}^{t})$ against t.

As $\bar{D}_{1,\cdot} \approx \bar{D}_{2,\cdot} \approx 0.04$, the relative frequency to grow new edge is about 5%, and that to dissolve existing edge is about 45%.

Empirical evidence for transitivity effects

Recall the transitivity model

$$\alpha_{i,j}^t = \xi_i \xi_j \frac{e^{aU_{i,j}^{t-1}}}{1 + e^{aU_{i,j}^{t-1}} + e^{bV_{i,j}^{t-1}}}, \quad \beta_{i,j}^t = \eta_i \eta_j \frac{e^{bV_{i,j}^{t-1}}}{1 + e^{aU_{i,j}^{t-1}} + e^{bV_{i,j}^{t-1}}},$$

where $U_{i,j}^t = \sum_{k \neq i,j} X_{i,k}^t X_{j,k}^t$, $V_{i,j}^t = \sum_{k \neq i,j} \{X_{i,k}^t (1 - X_{j,k}^t) + (1 - X_{i,k}^t) X_{j,k}^t\}$.

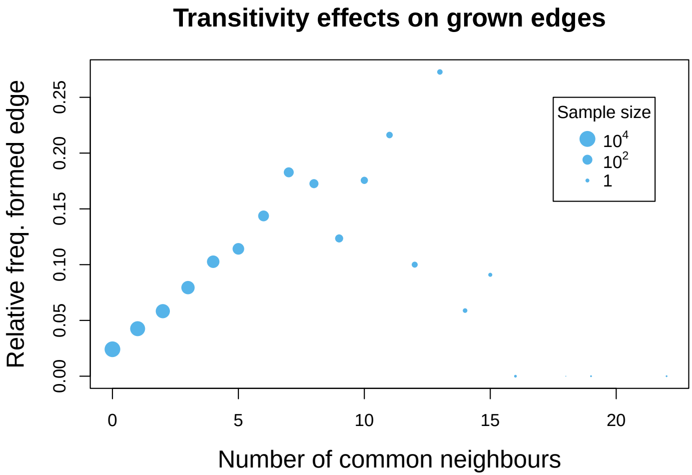
Let

$$\mathcal{U}_\ell = \big\{(i,j,t) : 1 \leq i < j \leq p,\ t \in [n] \setminus \{1\},\ X_{i,j}^{t-1} = 0,\ U_{i,j}^{t-1} = \ell\big\},$$

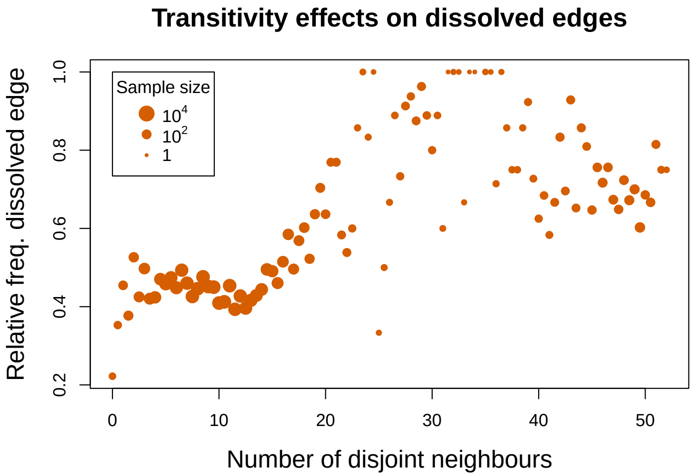
$$\mathcal{V}_\ell = \big\{(i,j,t) : 1 \leq i < j \leq p,\ t \in [n] \setminus \{1\},\ X_{i,j}^{t-1} = 1,\ V_{i,j}^{t-1} = \ell\big\},$$

$$\mathcal{U}_\ell^1 = \big\{(i,j,t) \in \mathcal{U}_\ell,\ X_{i,j}^t = 1\big\}, \quad \mathcal{V}_\ell^0 = \big\{(i,j,t) \in \mathcal{V}_\ell,\ X_{i,j}^t = 0\big\}.$$

Transitivity: both $|\mathcal{U}_\ell^1| / |\mathcal{U}_\ell|$ and $|\mathcal{V}_\ell^0| / |\mathcal{V}_\ell|$ $\nearrow$, as $\ell \nearrow$.



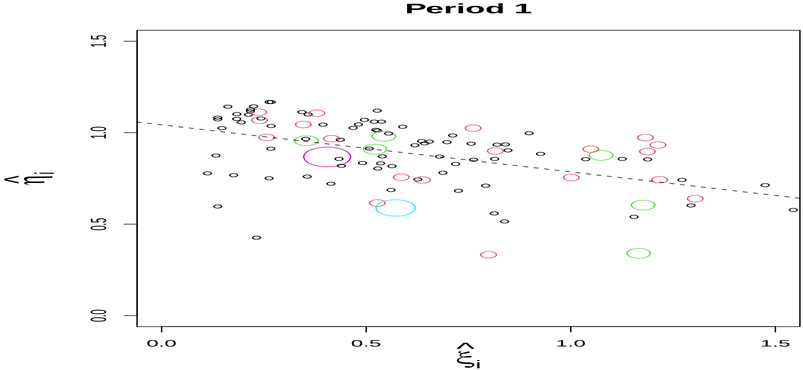
Plot of relative edge frequency $|\mathcal{U}_\ell^1|/|\mathcal{U}_\ell|$ against ℓ for $\ell = 0, 1, \cdots$.



Plot of relative non-edge frequency $|\mathcal{V}_\ell^0|/|\mathcal{V}_\ell|$ against ℓ for $\ell = 0, 1, \cdots$.

Fitting for Period 1

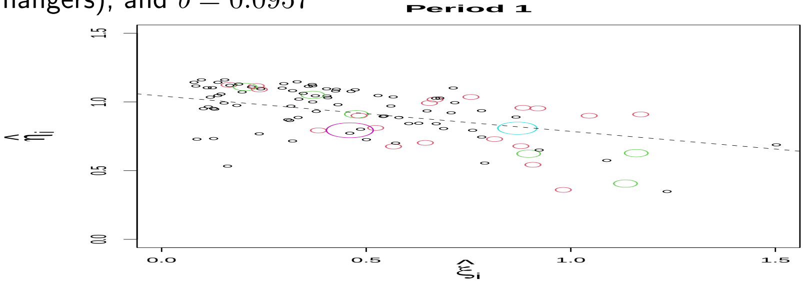
$\widehat{a} = 0.1273$ and $\widehat{b} = 0.0916$



$\widehat{\xi}_i$ and $\widehat{\eta}_i$ are negatively correlated: employees who tend to grow new edges also tend to maintain existing edges.

Fitting for Period 2

$\widehat{a} = 0.2099$ – stronger transitivity effect (more email activities among mangers), and $\widehat{b} = 0.0957$



Circles are sized and coloured according to hierarchical levels in the company: the smallest black circles have no direct reports, while the largest purple circle is CEO.

The means of $\widehat{\xi}_i$ for managers and non-managers are, respectively, 0.68 and 0.42: managers are more likely to grow edges. However, this increasing pattern does not continue at higher levels.

Stronger transitivity and lower edge density: concentration of email activities among a smaller group of employees, many of them managers.

Comparison with other models by AIC & BIC

Global AR model:

$$P(X_{i,j}^t = 1 | X_{i,j}^{t-1} = 0) = \alpha, \quad P(X_{i,j}^t = 0 | X_{i,j}^{t-1} = 1) = \beta$$

Edgewise AR model:

$$P(X_{i,j}^t = 1 | X_{i,j}^{t-1} = 0) = \alpha_{i,j}, \quad P(X_{i,j}^t = 0 | X_{i,j}^{t-1} = 1) = \beta_{i,j}$$

Edgewise mean model: $X_{i,j}^t \stackrel{\text{iid}}{\sim} \text{Bernoulli}(P_{i,j})$

Degree parameter mean model: $X_{i,j}^t \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\nu_i \nu_j)$

No edge dependence in the above 4 models

No dynamic dependence in the last 2 models

No. of parameters is, respectively, 2, $p(p-1)$, $\frac{1}{2}p(p-1)$ and p.

AR transitivity model has $2p+2$ parameters.

	Period 1		**Period 2**	
Model	**AIC**	**BIC**	**AIC**	**BIC**
Transitivity AR model	**33227**	**35176**	52548	**54654**
Global AR model	36309	36327	58267	58287
Edgewise AR model	42717	144102	55840	165394
Edgewise mean model	33248	83941	**47133**	101910
Degree parameter mean model	41730	42695	68969	70013

For Period 1, AR transitivity model achieves the lowest AIC and BIC.

For Period 2, it achieves the lowest BIC, and the 2nd lowest AIC (behind the edgewise mean model).

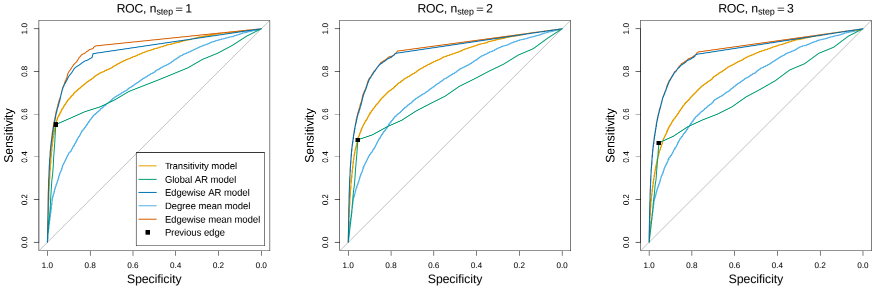
Post-sample edge forecasting

For 26 networks in Period 2, train models based on the first n_{train} data for $n_{\text{train}} = 10, \ldots, 23$.

Based on the fitted model, we make n_{step}-step forward prediction for $\mathbf{X}_{n_{\text{train}}+n_{\text{step}}}$ for $n_{\text{step}} = 1, 2, 3$.

The combined results are presented in ROC curves.

ROC curves: Sensitivity $= \frac{TP}{TP+FN}$, Specificity $= \frac{TN}{TN+FP}$



The two edgewise models (with $O(p^2)$ parameters) perform about the same, are clearly better than all the other models.

The transitivity model (with $O(p)$ parameters) outperform the other three models.

Setting

Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be available observations from a Markov chain with order $r(\geq 1)$, where $\mathbf{X}_t$ can be a vector or a matrix.

Let $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, be a parametric family of Markov models with order r.

Let $P_{\widehat{\boldsymbol{\theta}}}$ denote a fitted model based on the data $\mathbf{X}_1, \cdots, \mathbf{X}_n$. We assume that the estimated parameter can be expressed as

$$\widehat{\boldsymbol{\theta}} \equiv \widehat{\boldsymbol{\theta}}(\mathbf{Y}_r, \cdots, \mathbf{Y}_n),$$

where $\mathbf{Y}_t = (\mathbf{X}_t, \mathbf{X}_{t-1}, \cdots, \mathbf{X}_{t-r})$.

We denote by P^0 the true underlying distribution of Markov chain $\{\mathbf{X}_t\}$.

The diagnostic checking for the goodness-of-fit of the model is often via a statistical test for the hypothesis

$$H_0 : P^0 \in \{P_{\boldsymbol{\theta}}, \, \boldsymbol{\theta} \in \boldsymbol{\Theta}\} \quad \text{against} \quad H_1 : P^0 \notin \{P_{\boldsymbol{\theta}}, \, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}.$$

Adversarial Approach

1. Generate a Markov chain $\mathbf{X}_1^*, \cdots, \mathbf{X}_m^*$ from the fitted model $P_{\widehat{\boldsymbol{\theta}}}$. Let $\mathbf{Y}_t^* = (\mathbf{X}_t^*, \mathbf{X}_{t-1}^*, \cdots, \mathbf{X}_{t-r}^*)$.
2. Construct an optimum classification rule $\psi \in [0, 1]$ which assigns the true data $\{\mathbf{Y}_t\}$ and the simulated data $\{\mathbf{Y}_t^*\}$ into two different classes.

Then

$$\text{Hardness(Classification)} = \text{Goodness-of-Fit}(P_{\boldsymbol{\theta}})$$

If P^0 and $P_{\widehat{\boldsymbol{\theta}}}$ differ substantially from each other, we may find a ψ such that $\psi(\mathbf{Y}_t) = 1$ and $\psi(\mathbf{Y}_t^*) = 0$.

If the model is a correct one (i.e. H_0 holds) and $P_{\widehat{\boldsymbol{\theta}}} \xrightarrow{D} P^0$, it will be extremely hard to separate two sets of samples.

Let Ψ denote the set of candidate classification rules $\psi \in [0,1]$. Define the 'hardness':

$$G_n = \min_{\psi \in \Psi} \Big(-\frac{1}{n-r} \sum_{t=r+1}^{n} \log\{\psi(\mathbf{Y}_t)\} - \frac{1}{m-r} \sum_{t=r+1}^{m} \log\{1-\psi(\mathbf{Y}_t^*)\} \Big).$$

Then G_n is always non-negative, and it attains the minimum value 0 when $\psi(\mathbf{Y}_t) \equiv 1$ and $\psi(\mathbf{Y}_t^*) \equiv 0$.

The larger G_n is, the more likely $P_{\boldsymbol{\theta}}$ is an adequate model for data $\mathbf{X}_1, \cdots, \mathbf{X}_n$.

Assume in general $P_{\widehat{\boldsymbol{\theta}}} \xrightarrow{D} P^1$ and

$$\frac{1}{n-r}\sum_{t=r+1}^{n}\log\{\psi(\mathbf{Y}_t)\} \xrightarrow{P} E_{P^0}[\log\{\psi(\mathbf{Y}_t)\}],$$

$$\frac{1}{m-r}\sum_{t=r+1}^{m}\log\{1-\psi(\mathbf{Y}_t^*)\} \xrightarrow{P} E_{P^1}[\log\{1-\psi(\mathbf{Y}_t^*)\}],$$

as $n, m \to \infty$. Then the population counterpart of G_n is of the form

$$G_0 = \min_{\psi\in\Psi}\big(-E_{P^0}[\log\{\psi(\mathbf{Y}_t)\}] - E_{P^1}[\log\{1-\psi(\mathbf{Y}_t^*)\}]\big).$$

Let Ψ contains all possible classifiers taking values on the interval $[0,1]$. The minimizer in G_0 is the Bayesian rule:

$$\psi(\mathbf{y}) = p_0(\mathbf{y})/\{p_0(\mathbf{y}) + p_1(\mathbf{y})\},$$

where $p_0(\cdot)$ and $p_1(\cdot)$ are the PDFs of P^0 and P^1.

When $p_0 \equiv p_1$ (i.e. $P_{\boldsymbol{\theta}}$ is the correct model), the minimizer is $\psi(\mathbf{y}) \equiv 1/2$, and $G_0 = 2\log 2$.

Testing via sample splitting for $H_0 : P^0 \in \{P_{\boldsymbol\theta}, \boldsymbol\theta \in \boldsymbol\Theta\}$

We split the sample $\{\mathbf{Y}_{r+1}, \cdots, \mathbf{Y}_n\}$ into three parts:
$\mathcal{Y}_1 = \{\mathbf{Y}_{r+1}, \ldots, \mathbf{Y}_{r+n_1}\}$, $\mathcal{Y}_2 = \{\mathbf{Y}_{r+n_1+1}, \ldots, \mathbf{Y}_{r+n_1+n_2}\}$ and
$\mathcal{Y}_3 = \{\mathbf{Y}_{r+n_1+n_2+1}, \ldots, \mathbf{Y}_n\}$.

For convenience, we write $\mathcal{Y}_i = \{\mathbf{Y}_{t,i} : t = 1, \ldots, n_i\}, \ i = 1, 2, 3.$

The test is defined as follows:

- Based on $\mathcal{Y}_1$: estimate $\widehat{\boldsymbol\theta} = \widehat{\boldsymbol\theta}(\mathcal{Y}_1)$. Generate two independent synthetic samples $\{\mathbf{Y}^*_{1,1}, \cdots, \mathbf{Y}^*_{m_1,1}\}$ and $\{\mathbf{Y}^*_{1,2}, \cdots, \mathbf{Y}^*_{m_2,2}\}$ from $P_{\widehat{\boldsymbol\theta}}$ with $m_1 = n_2$ and $m_2^{-1} = o(n_3^{-1})$.

- Based on $\mathcal{Y}_2$: let

$$\widehat{\psi}_n = \arg\min_{\psi\in\Psi} \Big[-\frac{1}{n_2}\sum_{t=1}^{n_2} \log\{\psi(\mathbf{Y}_{t,2})\} - \frac{1}{m_1}\sum_{t=1}^{m_1} \log\{1-\psi(\mathbf{Y}^*_{t,1})\}\Big].$$

- Based on $\mathcal{Y}_3$: define

$$\widehat{G}_n = -\frac{1}{n_3}\sum_{t=1}^{n_3}\log\{\widehat{\psi}_n(\mathbf{Y}_{t,3})\} - \frac{1}{m_2}\sum_{t=1}^{m_2}\log\{1-\widehat{\psi}_n(\mathbf{Y}^*_{t,2})\}\,.$$

- The test statistic is defined as

$$\widehat{T}_n = \frac{\sqrt{n_3}(\widehat{G}_n - 2\log 2)}{\widehat{\sigma}\vee\delta_n}\,,$$

  where $\delta_n = \log(n)^{1.5} * n^{-1/8}$ (to control Type I error), and $\widehat{\sigma}^2$ is an estimator for $\sigma^2 = \mathrm{Var}[n_3^{-1/2}\sum_{t=1}^{n_3}\log\{\widehat{\psi}_n(\mathbf{Y}_{t,3})\}\,|\,\mathcal{Y}_1\cup\mathcal{Y}_2]$.
- We reject H_0 when $\widehat{T}_n < -z_\alpha$, where z_α is the top α-percentile of $N(0,1)$.

We let Ψ_n consist of the multilayer perceptron (MLP) classifiers of the form:

$$\Lambda(w_L^{\mathrm{T}}\sigma(w_{L-1}^{\mathrm{T}}\sigma(\cdots w_1^{\mathrm{T}}\sigma(w_0^{\mathrm{T}}X))))\,,$$

where $\Lambda\in[0,1]$ is 1-Lipschitz continous, σ is ReLu, and $|w|_\infty \le C_0$. We typically set $L=2$.

Estimation of σ^2.

Let $\widehat{Z}_t = \log \widehat{\psi}_n(\mathbf{Y}_{t,3})$ and $\tilde{Z} = n_3^{-1} \sum_{t=1}^{n_3} \widehat{Z}_t$. Define a kernel-type estimator fo σ^2:

$$\widehat{\sigma}_n^2 = \sum_{l=-n_3+1}^{n_3-1} \mathcal{K}\left(\frac{l}{b_n}\right) \widehat{H}_l \,,$$

where $\widehat{H}_l = n_3^{-1} \sum_{t=l+1}^{n_3} (\widehat{Z}_t - \tilde{Z})(\widehat{Z}_{t-l} - \tilde{Z})$ for $l \geq 0$ and $\widehat{H}_l = n_3^{-1} \sum_{t=-l+1}^{n_3} (\widehat{Z}_{t+l} - \tilde{Z})(\widehat{Z}_t - \tilde{Z})$ otherwise. Here $\mathcal{K}(\cdot)$ is a symmetric kernel function, and b_n is the bandwidth diverging with n.

Theoretical results

Condition 1. Assume $\{\mathbf{X}_t\}$ and $\{\mathbf{X}_t^*\}$ are strictly stationary β-mixing sequences with mixing-coefficients $\{\beta(k)\}_{k\geq 1}$ and $\{\beta^*(k)\}_{k\geq 1}$, respectively. There exists some universal constants $K_1, K_2 > 0$ and $\gamma > 0$ such that $\max\{\beta(k), \beta^*(k)\} \leq K_1 \exp(-K_2 k^{\gamma})$ for any $k \geq 1$.

Condition 2. The kernel function $\mathcal{K}(\cdot) : \mathbb{R} \to [-1, 1]$ is continuously differentiable on $\mathbb{R}$ and satisfies (i) $\mathcal{K}(0) = 1$, (ii) $\mathcal{K}(x) = \mathcal{K}(-x)$ for any $x \in \mathbb{R}$, and (iii) $\int_{-\infty}^{\infty} |\mathcal{K}(x)| \, \mathrm{d}x < \infty$. Let $b_n \asymp n_3^{1/4}$, $n_i = n/3$ for $i = 1, 2, 3$, and $m_1 = n_2$, $m_2^{-1} = o(n_3^{-1})$.

Then it can be proved that

$$\sqrt{n_3}(\widehat{G}_n - 2\log 2)/\sigma \big| H_0 \xrightarrow{D} N(0, 1),$$

$$\lim_{n\to\infty} P(T_n < -z_\alpha | H_0) \leq \alpha, \text{ and}$$

$$\lim_{n\to\infty} P(T_n < -z_\alpha | H_1) = 1.$$

Permutation test

1. Compute $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{Y}_{r+1}, \cdots, \mathbf{Y}_n)$, and generate $\mathbf{Y}^*_{r+1}, \cdots, \mathbf{Y}^*_n$ from $P_{\widehat{\boldsymbol{\theta}}}$
2. Split each sample into two subsamples:
$$\{\mathbf{Y}_{1,i}, \cdots, \mathbf{Y}_{n_i,i}\} \quad \text{and} \quad \{\mathbf{Y}^*_{1,i}, \cdots, \mathbf{Y}^*_{n_i,i}\},$$
for $i = 1, 2$, and $n_1 = n_2 = (n - r)/2$.
3. Fit the logistic regression $\widehat{\psi}(\cdot)$ for classifying two classes $\{\mathbf{Y}_{1,1}, \cdots, \mathbf{Y}_{n_1,1}\}$ and $\{\mathbf{Y}^*_{1,1}, \cdots, \mathbf{Y}^*_{n_1,1}\}$, and compute
$$\widehat{G}_n = -\frac{1}{n_2}\sum_{t=1}^{n_2} \log\{\widehat{\psi}(\mathbf{Y}_{t,2})\} - \frac{1}{n_2}\sum_{t=1}^{n_2} \log\{1 - \widehat{\psi}(\mathbf{Y}^*_{t,2})\}.$$
4. Permute $\{\mathbf{Y}_{1,2}, \cdots, \mathbf{Y}_{n_2,2}, \mathbf{Y}^*_{1,2}, \cdots, \mathbf{Y}^*_{n_2,2}\}$, and re-calculate $\widehat{G}_n$ above using the first n_2 entries in the permuted sequence as new $\{\mathbf{Y}_{t,2}\}$, and the last n_2 entries as $\{\mathbf{Y}^*_{t,n_2}\}$. Denoted by $G^\star_n$ the resulting value of $\widehat{G}_n$.
5. Repeat 4. above B times, obtaining $G^\star_{n,1}, \cdots, G^\star_{n,B}$, where $B \geq 1$ is a large integer. We reject H_0 if $\widehat{G}_n$ is smaller than the α-th sample quantile of $\{G^\star_{n,1}, \cdots, G^\star_{n,B}\}$.

To validate the permutation test, it can be proved that

(i) $\mathcal{L}(\widehat{G}_n|H_0) = \mathcal{L}(G_n^\star|H_0)$ asymptotically, and

(ii) $P(\widehat{G}_n < G_n^\star|H_0) \to 1.$

Best approximation among a selection of candidate models

In absence of an appropriate model, we may choose one among a selection of candidate models based on the proposed adversarial measures:

For the i-th candidate model,

1. Fit the model $\widehat{\boldsymbol{\theta}}_i = \widehat{\boldsymbol{\theta}}_i(\mathbf{Y}_{r_i+1}, \cdots, \mathbf{Y}_n)$.
2. Generate $\mathbf{Y}^*_{r_i+1}, \cdots, \mathbf{Y}^*_n$ from $P_{\widehat{\boldsymbol{\theta}}_i}$. Split each of the two samples into two:
$$\{\mathbf{Y}_{1,j}, \cdots, \mathbf{Y}_{n_i,j}\} \quad \text{and} \quad \{\mathbf{Y}^*_{1,j}, \cdots, \mathbf{Y}^*_{n_i,j}\}, \quad j = 1, 2,$$
where $n_i = (n - r_i)/2$.
3. Fit a logistic regression $\widehat{\psi}_i$ to classify two data sets $\{\mathbf{Y}_{t,1}\}$ and $\{\mathbf{Y}^*_{t,1}\}$, and compute
$$\widehat{G}_{n,i} = -\frac{1}{n_i}\sum_{t=1}^{n_i}\log\{\widehat{\psi}_i(\mathbf{Y}_{t,2})\} - \frac{1}{n_i}\sum_{t=1}^{n_i}\log\{1 - \widehat{\psi}_i(\mathbf{Y}^*_{t,2})\}.$$

The best approximation model is the one which attains $\max_i \widehat{G}_{n,i}$.

Simulation models

We evaluate the performance of several network models through simulations, including: Erdős-Rényi (ER) Model, Stochastic Block Model (SBM), β-Model (Chatterjee et al., 2011), Two-way Heterogeneity Model (TWHM) by Jiang et al. (2023) and Transitivity Model (TRM) by Chang et al. (2024).

Features selection

- When the networks generated from the ER, SBM or β-Model, we use the q quantiles of the degree sequences $d_i^t = \sum_{j=1}^p X_{i,j}^t$ as features for $\mathbf{X}^t$ with $q = \min(p/2, n/2, 20)$.
- When the networks generated from the TWHM, we use the q quantiles of $\sum_{j=1}^p X_{i,j}^t$, $\sum_{j=1}^p X_{i,j}^t X_{i,j}^{t-1}$, and $\sum_{j=1}^p X_{i,j}^t (1 - X_{i,j}^{t-1})$ as features for $\mathbf{X}^t$ with $q = \min(p/2, n/4, 20)$.
- For the TRM, features are selected according to the methodology outlined in Chang et al. (2024).

Simulation setting

Set significance level $\alpha = 0.05$, permutation times B$= 1000$, number of synthetic samples $m_2 = 1000$, $n_i = 100, 200$ $(i = 1, 2, 3)$ and $p = 50, 100$, or 200. Data are generated based on the following four model settings:

1. SBM: two communities with the probability matrix: $\begin{bmatrix} 0.6 & 0.2 \\ 0.2 & 0.4 \end{bmatrix}$.
2. β-Model: The parameters $\beta_i \overset{\text{i.i.d.}}{\sim} U(-1, 1)$.
3. TWHM: The parameters $\beta_{i,0} \overset{\text{i.i.d.}}{\sim} U(-1, 1)$ and $\beta_{i,1} \overset{\text{i.i.d.}}{\sim} U(-1, 1)$.
4. TRM: The parameters $\xi_i \overset{\text{i.i.d.}}{\sim} U(0.5, 0.7)$ and $\eta_i \overset{\text{i.i.d.}}{\sim} U(0.5, 0.7)$ and $a = b = 5$.

Tests

- Test 1: $\widehat{T}_n$ with MLP as the classifier.
- Test 2: Permutation test with the Logistic Regression as the classifier.

Simulation result

Table: The Type I error under Test 1 and Test 2, where H_0, the data generated model corresponds to ER, SBM, β-Model, TWHM, or TRM.

		ER		SBM		β-Model	
		n=100	n=200	n=100	n=200	n=100	n=200
Test 1	p=50	0	0	0	0	0	0
	p=100	0	0	0	0	0	0
	p=200	0	0	0	0	0	0
Test 2	p=50	0.07	0.05	0.1	0.08	0.06	0.04
	p=100	0	0.06	0.08	0.07	0.07	0.06
	p=200	0.03	0.02	0.06	0.07	0.12	0.08
		TWHM		TRM			
Test 1	p=50	0	0	0	0		
	p=100	0	0	0.04	0.03		
	p=200	0	0	0.14	0.15		
Test 2	p=50	0.08	0.1	0.1	0.1		
	p=100	0.1	0	0.04	0.07		
	p=200	0.08	0.04	0.09	0.05		

Real Data: Email interactions

The email interactions in a medium-sized Polish manufacturing company in January -September 2010 (Michalski et al., 2014).

Consider $p = 106$ of the most active participants out of an original 167 employees.

$n = 39$ represents 39 weeks, and $X_{i,j}^t = 1$ if participants i and j exchanged at least one email during Week t.

A change-point at $t = 14$: Period 1 first 13 points, Period 2 last 26 points

Comparison with other models

- Global AR model:

$$P(X_{i,j}^t = 1 | X_{i,j}^{t-1} = 0) = \alpha, \quad P(X_{i,j}^t = 0 | X_{i,j}^{t-1} = 1) = \beta$$

- Edgewise AR model:

$$P(X_{i,j}^t = 1 | X_{i,j}^{t-1} = 0) = \alpha_{i,j}, \quad P(X_{i,j}^t = 0 | X_{i,j}^{t-1} = 1) = \beta_{i,j}$$

- Edgewise mean model: $X_{i,j}^t \overset{\text{i.i.d.}}{\sim}$ Bernoulli$(P_{i,j})$
- Degree parameter mean model: $X_{i,j}^t \overset{\text{i.i.d.}}{\sim}$ Bernoulli$(v_i v_j)$

No edge dependence in the above 4 models
No dynamic dependence in the last 2 models
No. of parameters is, respectively, 2, $p(p-1)$, $p(p-1)/2$ and p.
AR transitivity model has $2p+2$ parameters.

Significance level: $\alpha = 0.05$, permutation times: $B = 1000$.

The selected features include:

$$D_{1,t} = \frac{2}{p(p-1)} \sum_{i<j} (1 - X_{i,j}^{t-1}) X_{i,j}^{t}\,, \quad D_{0,t} = \frac{2}{p(p-1)} \sum_{i<j} X_{i,j}^{t-1} (1 - X_{i,j}^{t})\,,$$

$$W_{1,t} = \frac{2}{p(p-1)} \sum_{i<j} X_{i,j}^{t-1} X_{i,j}^{t}, \quad W_{0,t} = \frac{2}{p(p-1)} \sum_{i<j} (1 - X_{i,j}^{t-1})(1 - X_{i,j}^{t}),$$

$$U_t = \frac{2}{p(p-1)(p-2)} \sum_{i<j} X_{i,j}^{t} \sum_{k \neq i,j} X_{i,k}^{t-1} X_{j,k}^{t-1}\,,$$

$$V_t = \frac{2}{p(p-1)(p-2)} \sum_{i<j} (1 - X_{i,j}^{t}) \sum_{k \neq i,j} \{X_{i,k}^{t-1}(1 - X_{j,k}^{t-1}) + (1 - X_{i,k}^{t-1}) X_{j,k}^{t-1}\},$$

$$d_t = \frac{2}{p(p-1)} \sum_{i<j} X_{i,j}^{t}, \quad C_{3,t} = \frac{2}{p(p-1)(p-2)} \sum_{i,j,k \text{ different}} X_{i,j}^{t} X_{j,k}^{t}\,,$$

$$R_{3,t} = \frac{6}{p(p-1)(p-2)} \sum_{i,j,k \text{ different}} X_{i,j}^{t} X_{j,k}^{t} X_{k,i}^{t}\,.$$

Given the small sample sizes (e.g. Period 1: $n_1 = 7, n_2 = 6$), we performed feature selection using random forest variable importance (200 trees).

We retain the top d' features, where

$$d' = \arg \max_{1 \le i \le d-1} \frac{v_i + \epsilon}{v_{i-1} + \epsilon},$$

v_i is the variable importance, and $\epsilon = 10^{-4}$ is a small constant.

Test 2 repeated 100 times for stable results.

Test and model selection

Table: The testing results of Test 2 and the model selection results where ✓ means not rejected H_0.

	Period 1		Period 2	
Models	**Test**	**G-value**	**Test**	**G-value**
Transitivity model	✓	1.11	✓	**1.89**
Global AR model		0.02		0.44
Edgewise AR model	✓	1.16		0.28
Edgewise mean model	✓	**1.31**	✓	1.45
Degree parameter mean model		0.29		1.01

The edgewise mean model and transitivity model pass both tests and achieves the largest G-value for period 1 and 2, respectively. The edgewise AR model pass the

Reference

Chang, J., Fang, Q., Kolaczyk, E. D., MacDonald, P. W., and Yao, Q. (2024). Autoregressive networks with dependent edges. `ArXiv`.

Chatterjee, S., Diaconis, P., and Sly, A. (2011). Random graphs with a given degree sequence. *The Annals of Applied Probability*, 21(4):1400–1435.

Jiang, B., Leng, C., Yan, T., Yao, Q., and Yu, X. (2023). A two-way heterogeneity model for dynamic networks. *arXiv preprint arXiv:2305.12643*.

Michalski, R., Kajdanowicz, T., Bródka, P., and Kazienko, P. (2014). Seed selection for spread of influence in social networks: Temporal vs. static approach. *New Generation Computing*, 32:213–235.