

# **Random threshold graphs**

Rahul Roy

Indian Statistical Institute, New Delhi.

# Threshold Graphs

We want to construct a graph  $G_n$  on the vertex set  $V_n := \{1, 2, \dots, n\}$ .

Each vertex  $i$  comes with a weight  $X_i \in \mathbb{R}$  attached to it.

There is a threshold value  $\theta \in \mathbb{R}$  such that

$i$  and  $j$  are connected if  $X_i + X_j > \theta$ .

The **threshold graph**  $\mathbb{G}_n$  consists of the vertex set

$$V_n = \{1, 2, \dots, n\},$$

and the edge set

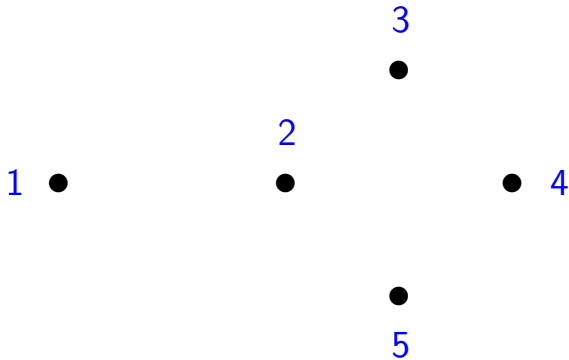
$$E_n = \{\langle i, j \rangle : X_i + X_j > \theta, 1 \leq i < j \leq n\}.$$

[Homophily graphs  $|X_i - X_j| < \theta'$ ]

# Equivalent characterisations

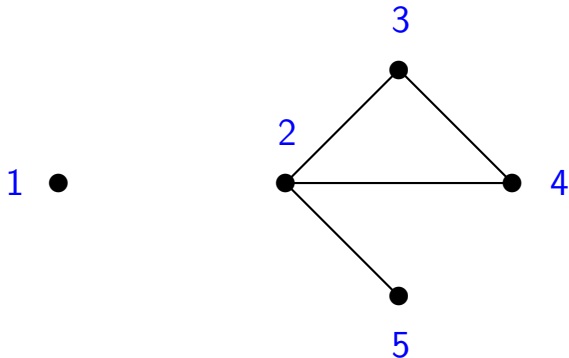
Each of the following is equivalent to a threshold graph (Mahadev and Peled 1995):

1. As a labelled graph,  $\mathbb{G}$  is uniquely determined by its degree sequence.
2.  $\mathbb{G}$  may be constructed by placing vertices one at a time such that each new vertex is either isolated or connected to all the previous vertices.



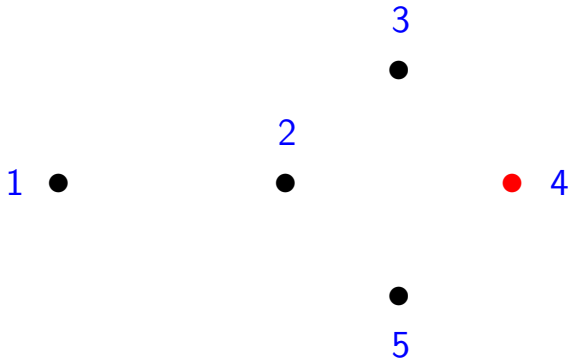
Weights:  $X_1 = 1$ ,  $X_2 = 4$ ,  $X_3 = 3$ ,  $X_4 = 3$ ,  $X_5 = 2$

$$\theta = 5.5$$



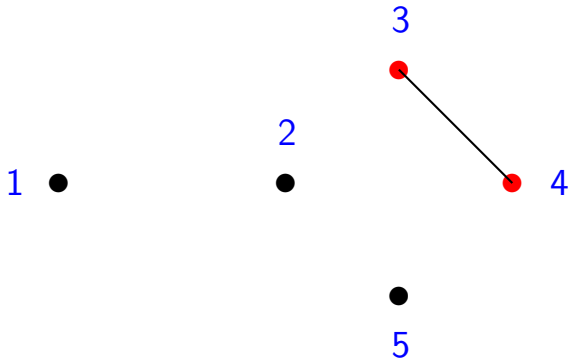
Weights:  $X_1 = 1$ ,  $X_2 = 4$ ,  $X_3 = 3$ ,  $X_4 = 3$ ,  $X_5 = 2$

$$\theta = 5.5$$



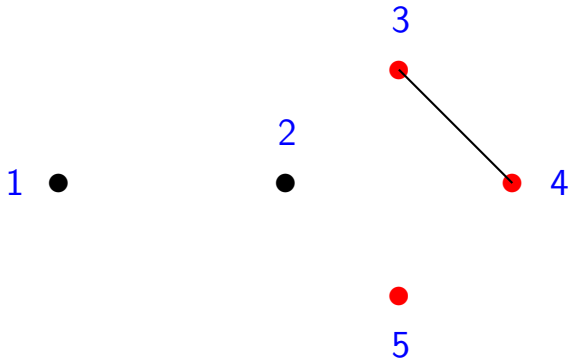
Weights:  $X_1 = 1$ ,  $X_2 = 4$ ,  $X_3 = 3$ ,  $X_4 = 3$ ,  $X_5 = 2$

$$\theta = 5.5$$



Weights:  $X_1 = 1$ ,  $X_2 = 4$ ,  $X_3 = 3$ ,  $X_4 = 3$ ,  $X_5 = 2$

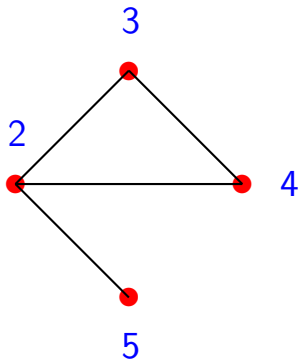
$$\theta = 5.5$$



Weights:  $X_1 = 1$ ,  $X_2 = 4$ ,  $X_3 = 3$ ,  $X_4 = 3$ ,  $X_5 = 2$

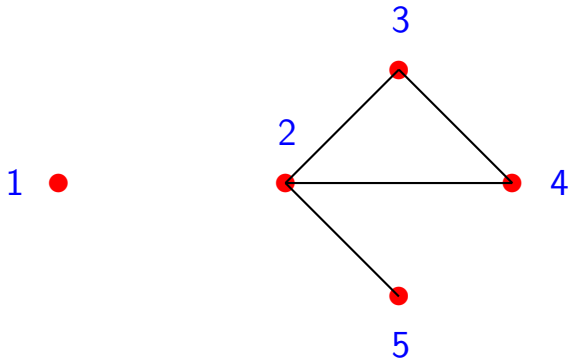
$$\theta = 5.5$$

1 ●



Weights:  $X_1 = 1$ ,  $X_2 = 4$ ,  $X_3 = 3$ ,  $X_4 = 3$ ,  $X_5 = 2$

$\theta = 5.5$



Weights:  $X_1 = 1$ ,  $X_2 = 4$ ,  $X_3 = 3$ ,  $X_4 = 3$ ,  $X_5 = 2$   
 $\theta = 5.5$

A threshold graph consists of one connected subgraph and each of the other components is an isolated vertex.

Note the ‘complement’ of a threshold graph is also a threshold graph, with the thresholding property being

$$X_i + X_j \leq \theta$$

# Random threshold graph

Now suppose that  $X_1, X_2, \dots, X_n$  are independent and identically distributed (iid) random variables with  $X_1$  having a distribution function  $F$ .

The resulting random graph  $\mathbb{G}_n$  is the random threshold graph.

We discuss some results regarding the asymptotic properties of the graph  $\mathbb{G}_n$ .

# Degree of a vertex

The degree of the vertex  $i$  in  $\mathbb{G}_n$  is given by

$$D_n(i) := \#\{j : \langle i, j \rangle \in E_n\}$$

Note that  $\{D_n(i) : i \geq 1\}$  are identically distributed, though not independent. Let  $D_n$  denote a random variable with this distribution.

To obtain the distribution of  $D_n$ , note that given  $n + 1$  vertices, and conditioned on the event  $X_1 = x$ , a vertex  $j \in \{2, \dots, n + 1\}$  is connected to the vertex 1 if and only if  $X_j > \theta - x$ .

So, for  $0 \leq k \leq n$  we have

$$P(D_{n+1} = k) = \int_{-\infty}^{\infty} \binom{n}{k} [1 - F(\theta - x)]^k [F(\theta - x)]^{n-k} F(dx).$$

With this and studying the characteristic function of  $D_n/n$  we get

## Theorem

As  $n \rightarrow \infty$ ,

$\frac{D_n}{n}$  converges in distribution to  $1 - F(\theta - X_1)$ .

For statistical purposes, and in particular in a data analysis context, we also need to understand the correlation of the degrees of two vertices **1** and **2** (say).

We have

## Theorem

$\frac{D_n(1)}{n}$  and  $\frac{D_n(2)}{n}$  are asymptotically independent.

However this independence breaks under the condition that the vertices **1** and **2** are connected.

## Theorem

Given  $\langle \mathbf{1}, \mathbf{2} \rangle \in E_n$ ,  $\frac{D_n(1)}{n}$  and  $\frac{D_n(2)}{n}$  are not asymptotically independent.

The previous theorem needs a caveat –

We need the distribution function  $F$  to be such that

$$P(\langle \mathbf{1}, \mathbf{2} \rangle \in E_n \mid X_1 < \theta/2, X_2 > \theta/2) > 0,$$

i.e., there exists  $u$  and  $v$  in the support of  $F$  such that  $u < \theta/2 < v$  and  $u + v > \theta$ .

Otherwise  $\mathbb{G}_n$  has one connected subgraph, which is complete, and the other vertices are all isolated, i.e.,

To understand clustering we need to study the number of triangles in this random graph.

Formally,

$$T_n := \#\{(i, j, k) : 1 \leq i < j < k \leq n, \\ \langle i, j \rangle, \langle j, k \rangle \text{ and } \langle k, i \rangle \in E_n\}.$$

Let  $h : \mathbb{R}^3 \rightarrow \mathbb{R}$  be given by

$$h(x, y, z) := I_{\{x+y>\theta, y+z>\theta, z+x>\theta\}}$$

and

$$\begin{aligned} F_3(\theta) &:= E(h(X_1, X_2, X_3)) \\ &= P(\langle 1, 2 \rangle, \langle 2, 3 \rangle, \langle 3, 1 \rangle \in E_n) \end{aligned}$$

$$\begin{aligned} \zeta(F) &:= \\ &E \left[ \left( \int_{\mathbb{R}} \int_{\mathbb{R}} F(dx_2) F(dx_3) h(X_1, x_2, x_3) \right)^2 \right] \\ &\quad - (F_3(\theta))^2 \\ &> 0. \end{aligned}$$

We have

## Theorem

As  $n \rightarrow \infty$ ,

- (a)  $\frac{6}{n^3}T_n \rightarrow F_3(\theta)$  almost surely;
- (b)  $\sqrt{n} \left[ \frac{6}{n^3}T_n \rightarrow F_3(\theta) \right]$  converges in distribution to  $\sqrt{3\zeta(F)} \mathbf{Z}$ , where  $\mathbf{Z}$  is a standard normal random variable.

For the number of triangles with  $\mathbf{1}$  being a vertex, i.e.

$$T_n(\mathbf{1}) :=$$

$$\#\{(i,j) : 2 \leq i < j < n + 1, h(X_1, X_i, X_j) = 1\}$$

## Theorem

As  $n \rightarrow \infty$ ,

$$\frac{2}{n^2} T_n \Rightarrow \int_{\mathbb{R}} \int_{\mathbb{R}} F(dx_2) F(dx_3) h(X, x_2, x_3)$$

where  $X$  is an independent random variable identical in distribution to  $X_1$ .

Besides triangles, we can also count subgraphs in  $\mathbb{G}_n$  isomorphic to a given subgraph  $H$  and obtain similar results.

In a data analysis context a fixed subgraph is called a motif of a graph. Depending on the types of real networks (e.g., internet, gene networks, neural networks, social networks), there are some small motifs which appear in an entire graph significantly more than in random graphs. These motifs are relevant to functional roles, e.g., signal transduction, information processing, etc.

For  $H_n = \#$  subgraphs in  $\mathbb{G}_n$  isomorphic to a fixed graph  $H$ , where  $H$  is a connected graph on  $k$  vertices, we have

## Theorem

As  $n \rightarrow \infty$

$$\frac{H_n}{n^k} \rightarrow F(H) \text{ almost surely}$$

and

$$\sqrt{n} \left[ \frac{H_n}{n^k} - F(H) \right] \Rightarrow \sqrt{k\sigma_H^2} Z$$

where  $F(H)$  and  $\sigma_H^2$  are quantities depending on the subgraph  $H$ .

To prove these convergence theorems we observe that there is an appropriate **U-statistic** which may be used to count triangles, fixed motifs, etc. This observation allows us to invoke the literature of U-statistics to obtain the results. The kernel function  $h$  is symmetric, so

$$\begin{aligned} T_n &:= \binom{n}{3} \frac{1}{\binom{n}{3}} \sum_{1 \leq i < j < k \leq n} h(X_i, X_j, X_k), \\ &= \binom{n}{3} U_n. \end{aligned}$$

The above study is dimension free – i.e. it does not take into account the location of the vertex in space.

A spatial (or geographical) model must be such that the connectivity of two vertices must be dependent on the distance between them.

# The Poisson model

Let  $\xi_0, \xi_1, \dots$  be a Poisson point process of intensity  $\lambda$  on  $\mathbb{R}^d$ , i.e., points are ‘randomly placed’ in space with an average of  $\lambda$  many points in a region of unit  $d$ -dimensional volume. We take  $\xi_0 \equiv \mathbf{O}$ , the origin.

Associated with  $\xi_i$  is a random variable  $X_i$ , where  $X_0, X_1, \dots$  are i.i.d. with a common distribution function  $F$ .

The random graph  $\mathbb{G}_{\theta, \beta}$  is obtained by

connecting  $\xi_i$  and  $\xi_j$  if and only if

$$X_i + X_j > \theta |\xi_i - \xi_j|^\beta$$

The degree,  $\Delta_r$ , of the origin in a sphere of radius  $r$  is

$$\Delta_r := \#\{i \geq 1 : (X_0 + X_i) > \theta |\xi_i|^\beta \text{ and } |\xi_i| \leq r\}.$$

For a fixed  $x \in \mathbb{R}$  and  $r > 0$ , let

$$\begin{aligned} f(r, x) &:= 1 - F(\theta r^\beta - x) \\ C_r(x) &:= \int_0^r t^{d-1} f(t, x) dt \end{aligned}$$

## Theorem

As  $r \rightarrow \infty$  if

$C_r(x) \rightarrow C(x) := \int_0^\infty t^{d-1} f(t, x) dt < \infty$  for every  $x \in \mathbb{R}$ , then we have

$$\Delta_r \Rightarrow \Delta$$

where the characteristic function of  $\Delta$  is given by

$$\phi_\Delta(t) = \int_{\mathbb{R}} F(dx) \exp(-\lambda \pi_{d-1} C(x) (1 - e^{it})).$$

## Theorem

Suppose there exists a sequence  $C_r$  such that  $C_r \rightarrow \infty$  and  $\frac{C_r(x) - C_r}{\sqrt{C_r}} \rightarrow g(x)$  as  $r \rightarrow \infty$  for every  $x \in \mathbb{R}$ . We have, as  $r \rightarrow \infty$ ,

$$\frac{\Delta_r - \lambda \pi_{d-1} C_r}{\sqrt{\lambda \pi_{d-1} C_r}} \Rightarrow Z + \sqrt{\lambda \pi_{d-1}} g(X_0).$$

# An example

For  $\beta = 1$  and  $d = 2$ , and  $F : [0, \infty) \rightarrow [0, 1]$  given by

$$F(x) = 1 - Kx^{-\alpha} \text{ for some } 0 < \alpha < 2, K > 0$$

we get

$$C_r = \frac{K\theta^{-\alpha}r^{2-\alpha}}{2 - \alpha}$$

and conditions of the last theorem are satisfied with  $g(x) = 0$  for all  $x \in \mathbb{R}$

Konno, Masuda, Roy and Sarkar (2006)

Ide, Konno and Masuda (2008)

Diaconis, Holmes and Janson (2010)

Farre and Roy (2012)