Social Networks & Recommendation Systems

V. Static random graphs.

Grzegorz Siudem

Warsaw University of Technology

European Funds
Knowledge Education Development

Warsaw University of Technology

European Union
European Social Fund

MSc program in Data Science has been developed
as a part of task 10 of the project
„NERW PW. Science - Education - Development - Cooperation”
co-funded by European Union from European Social Fund.

Before classes

Reminder

- Bernoulli distribution $\mathbb{P}(X = k) = \binom{N}{k} p^k (1 - p)^{N-k}$,
- Piosson distribution $\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$,
- convergence of one to another (in what sense?)
- handshake lemma $\sum_i k_i = 2E$ $(\langle k \rangle = 2\frac{E}{N}.)$.

Question:

Why does the 2 factor appear in the formula?

Lecture

A reminder of what complex network science is

$$\text{complex network science} = \text{data} + \text{metrics} + \text{models}$$

A reminder of what complex network science is

complex network science = data + metrics + models

Why do we need models?

- for generating networks under controlled conditions,

A reminder of what complex network science is

complex network science $=$ data $+$ metrics $+$ models

Why do we need models?

- for generating networks under controlled conditions,
- allow to extract network features important from a given point of view,

complex network science $=$ data $+$ metrics $+$ models

Why do we need models?

- for generating networks under controlled conditions,
- allow to extract network features important from a given point of view,
- give a chance to build good statistics while analyzing the dynamics on networks (classes 11 and 12),

A reminder of what complex network science is

complex network science $=$ data $+$ metrics $+$ models

Why do we need models?

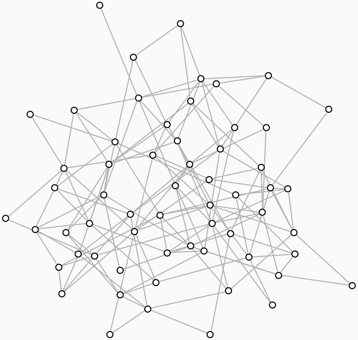
- for generating networks under controlled conditions,
- allow to extract network features important from a given point of view,
- give a chance to build good statistics while analyzing the dynamics on networks (classes 11 and 12),
- give us an insight into the mechanisms behind real processes (see next classes and BA networks),

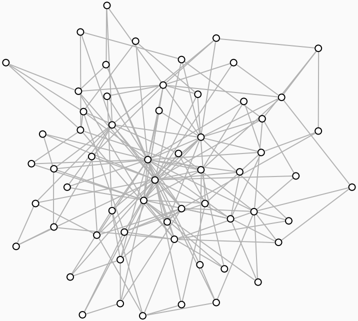
complex network science = data + metrics + models

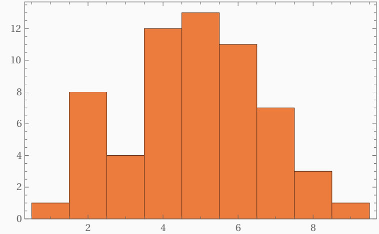
Why do we need models?

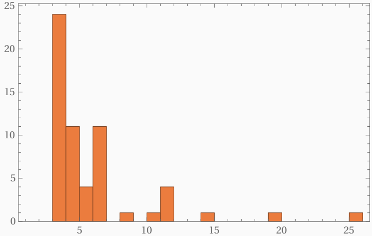
- for generating networks under controlled conditions,
- allow to extract network features important from a given point of view,
- give a chance to build good statistics while analyzing the dynamics on networks (classes 11 and 12),
- give us an insight into the mechanisms behind real processes (see next classes and BA networks),
- it is a very beautiful field of applied mathematics.

Attention! ER graphs are *normal*



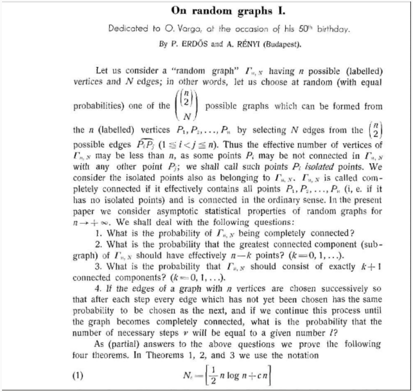






And really poissonian, as we will see in a moment...

Erdős-Rényi model



On random graphs I.

Dedicated to O. Varga, at the occasion of his 50th birthday.

By P. ERDŐS and A. RÉNYI (Budapest).

Let us consider a "random graph" $\Gamma_{n,N}$ having n possible (labelled) vertices and N edges; in other words, let us choose at random (with equal probabilities) one of the $\binom{\binom{n}{2}}{N}$ possible graphs which can be formed from the n (labelled) vertices $P_1, P_2, \ldots, P_n$ by selecting N edges from the $\binom{n}{2}$ possible edges $\widehat{P_i P_j}$ $(1 \leqq i < j \leqq n)$. Thus the effective number of vertices of $\Gamma_{n,N}$ may be less than n, as some points P_i may be not connected in $\Gamma_{n,N}$ with any other point P_j; we shall call such points P_i *isolated* points. We consider the isolated points also as belonging to $\Gamma_{n,N}$. $\Gamma_{n,N}$ is called completely connected if it effectively contains all points $P_1, P_2, \ldots, P_n$ (i, e. if it has no isolated points) and is connected in the ordinary sense. In the present paper we consider asymptotic statistical properties of random graphs for $n \to +\infty$. We shall deal with the following questions:

1. What is the probability of $\Gamma_{n,N}$ being completely connected?

2. What is the probability that the greatest connected component (subgraph) of $\Gamma_{n,N}$ should have effectively $n-k$ points? $(k=0, 1, \ldots)$.

3. What is the probability that $\Gamma_{n,N}$ should consist of exactly $k+1$ connected components? $(k=0, 1, \ldots)$.

4. *If the edges of a graph with n vertices are chosen successively so that after each step every edge which has not yet been chosen has the same probability to be chosen as the next, and if we continue this process until the graph becomes completely connected, what is the probability that the number of necessary steps ν will be equal to a given number l?*

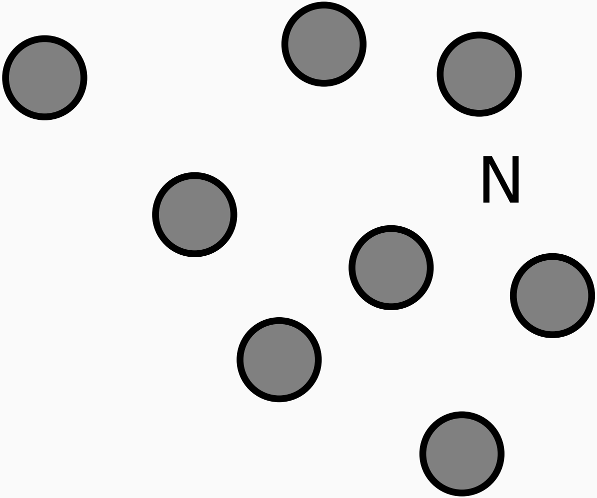
As (partial) answers to the above questions we prove the following four theorems. In Theorems 1, 2, and 3 we use the notation

$$N_c = \left[\frac{1}{2} n \log n + c n\right] \qquad (1)$$

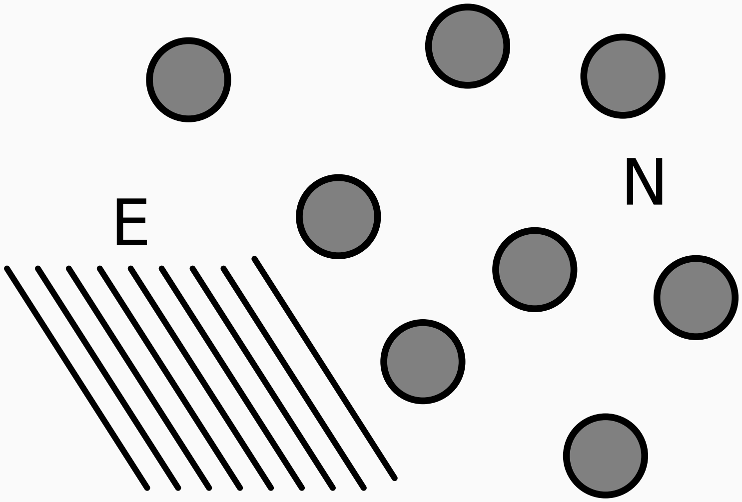
Two variants of the model:

- $G_{N,E}$ (Erdős-Rényi),
- $G_{N,p}$ (E. Gilbert)

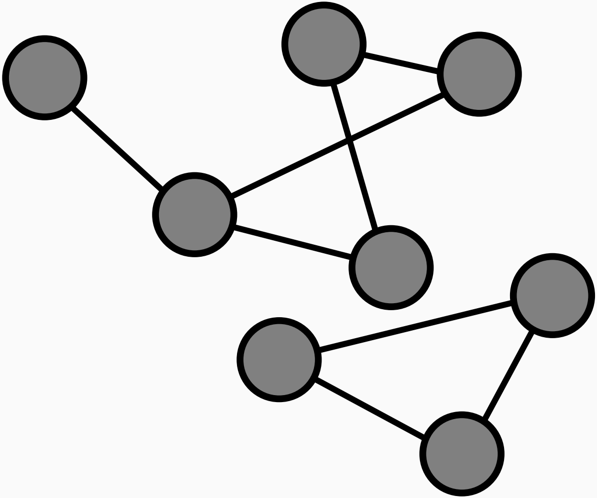
Erdős-Rényi model in the $G_{N,E}$ form



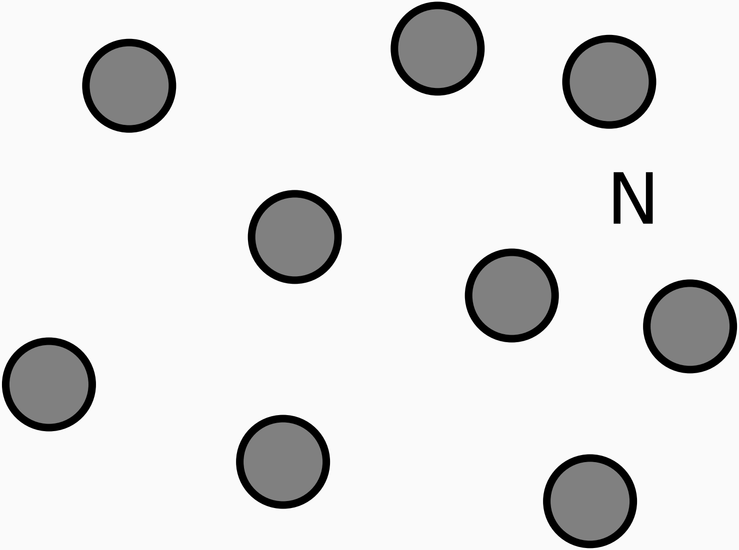
Erdős-Rényi model in the $G_{N,E}$ form



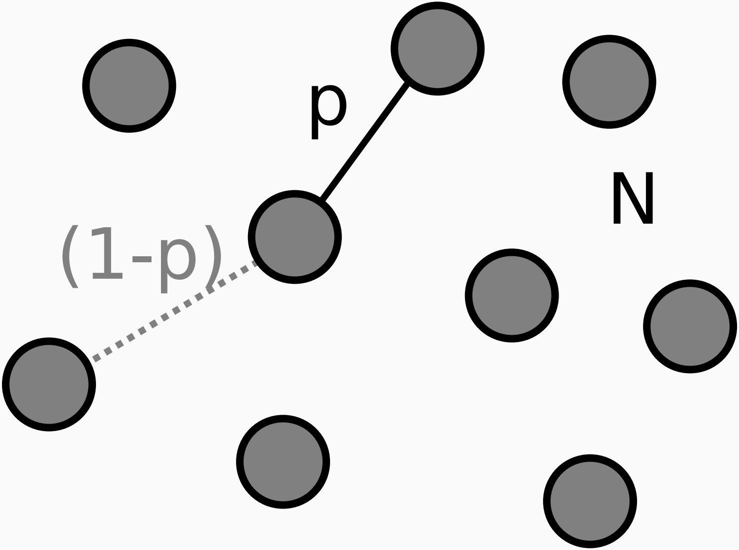
Erdős-Rényi model in the $G_{N,E}$ form



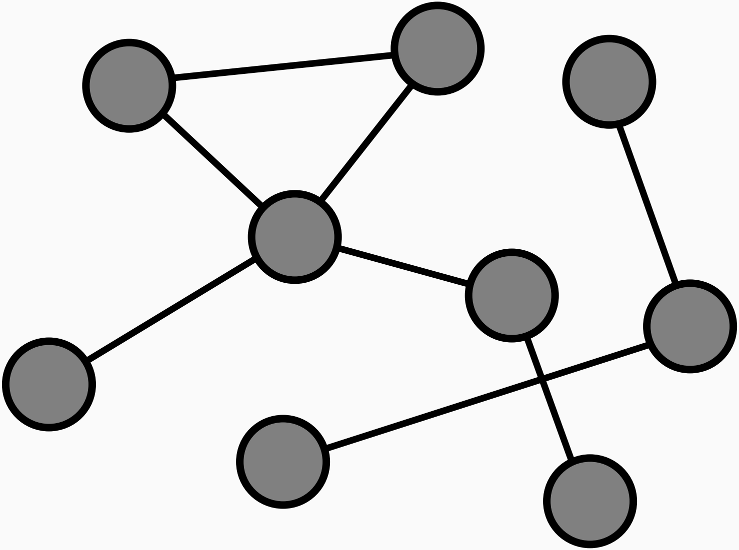
Erdős-Rényi model in the $G_{N,p}$ form



Erdős-Rényi model in the $G_{N,p}$ form



Erdős-Rényi model in the $G_{N,p}$ form

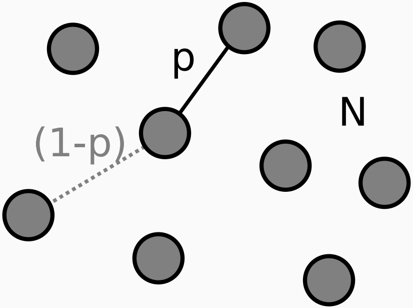


Let us focus on the $G_{N,p}$ model

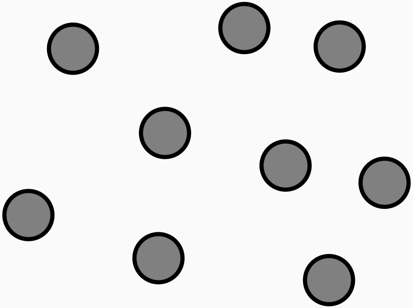
Assumptions

We analyze the ensemble described by two parameters:

- N niber of vertices,
- p probability of two vertices connection.



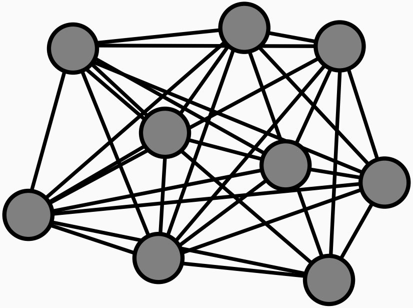
Limiting case: $p = 0$



Conclusion:

$p = 0$ is a trivial case.

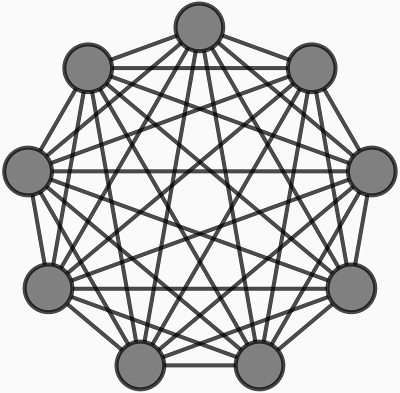
Limiting case: $p = 1$



Conclusion 1:

This is a very unsuccessful visualization...

Limiting case: $p = 1$



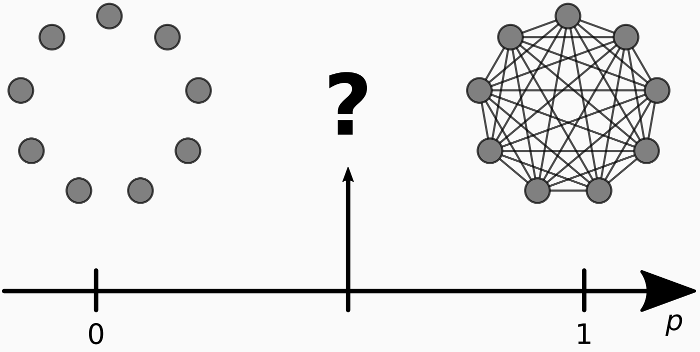
Conclusion 1:

~~This is a very unsuccessful visualization...~~

Conclusion 2:

$p = 1$ is a complete graphs.

ER graphs – dependence on the parameter p



Solution of the $G_{N,p}$ model

Question:

How many edges *on average* will appear?

Solution of the $G_{N,p}$ model

Question:

How many edges *on average* will appear?

Let's count

$$\langle E \rangle = p \times \frac{N(N-1)}{2},$$

which with the handshake lemma leads to

$$\langle k \rangle = p\frac{N(N-1)}{2}\frac{2}{N} = p(N-1) \approx pN.$$

Solution of the $G_{N,p}$ model

Question:

How many edges *on average* will appear?

Let's count

$$\langle E \rangle = p \times \frac{N(N-1)}{2},$$

which with the handshake lemma leads to

$$\langle k \rangle = p\frac{N(N-1)}{2}\frac{2}{N} = p(N-1) \approx pN.$$

So we know the average degree

What about the degree distribution of the vertices?

Solution of the $G_{N,p}$ model

Random variable describing vertex degree K

$$K = \sum_{i=1}^{N-1} X_k,$$

where X_k are iid variables with $\mathbb{P}(X_k = 1) = p$ $\mathbb{P}(X_k = 0) = 1 - p$.

Solution of the $G_{N,p}$ model

Random variable describing vertex degree K

$$K = \sum_{i=1}^{N-1} X_k,$$

where X_k are iid variables with $\mathbb{P}(X_k = 1) = p$ $\mathbb{P}(X_k = 0) = 1 - p$.

K has binomial distribution

$$\mathcal{P}(k) = \mathbb{P}(K = k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

Solution of the $G_{N,p}$ model

Random variable describing vertex degree K

$$K = \sum_{i=1}^{N-1} X_k,$$

where X_k are iid variables with $\mathbb{P}(X_k = 1) = p$ $\mathbb{P}(X_k = 0) = 1 - p$.

K has binomial distribution

$$\mathcal{P}(k) = \mathbb{P}(K = k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

Poissonian approximation

$$\mathcal{P}(k) \approx \frac{e^{-\langle k \rangle} \langle k \rangle^k}{k!},$$

because $\langle k \rangle \approx Np$.

Non-physicality of ER graphs

Let's calculate variance for the poissonian case

$$\mathbb{E}(K) = \sum_{k=0}^{\infty} \frac{k e^{-\langle k \rangle} \langle k \rangle^k}{k!} = \cdots = \langle k \rangle,$$

$$\mathbb{E}(K^2) = \sum_{k=0}^{\infty} \frac{k^2 e^{-\langle k \rangle} \langle k \rangle^k}{k!} = \cdots = \langle k \rangle + \langle k \rangle^2.$$

$$\mathrm{Var}(K) = \mathbb{E}(K^2) - [\mathbb{E}(K)]^2 = \langle k \rangle^2$$

Let's calculate variance for the poissonian case

$$\mathbb{E}(K) = \sum_{k=0}^{\infty} \frac{k e^{-\langle k \rangle} \langle k \rangle^k}{k!} = \cdots = \langle k \rangle,$$

$$\mathbb{E}(K^2) = \sum_{k=0}^{\infty} \frac{k^2 e^{-\langle k \rangle} \langle k \rangle^k}{k!} = \cdots = \langle k \rangle + \langle k \rangle^2.$$

$$\mathrm{Var}(K) = \mathbb{E}(K^2) - [\mathbb{E}(K)]^2 = \langle k \rangle^2$$

Details?

During the project.

Non-physicality of ER graphs

Let's calculate variance for the poissonian case

$$\mathbb{E}(K) = \sum_{k=0}^{\infty} \frac{k e^{-\langle k \rangle} \langle k \rangle^k}{k!} = \cdots = \langle k \rangle,$$

$$\mathbb{E}(K^2) = \sum_{k=0}^{\infty} \frac{k^2 e^{-\langle k \rangle} \langle k \rangle^k}{k!} = \cdots = \langle k \rangle + \langle k \rangle^2.$$

$$\mathrm{Var}(K) = \mathbb{E}(K^2) - [\mathbb{E}(K)]^2 = \langle k \rangle^2$$

Details?

During the project.

Such small variance is not observed in the real life networks...

Non-physicality of ER graphs

Clustering coefficient: reminder

We are looking for the value of the clustering coefficient for ER graphs

$$C_i = \frac{2E_i}{k_i(k_i - 1)}.$$

Non-physicality of ER graphs

Clustering coefficient: reminder

We are looking for the value of the clustering coefficient for ER graphs

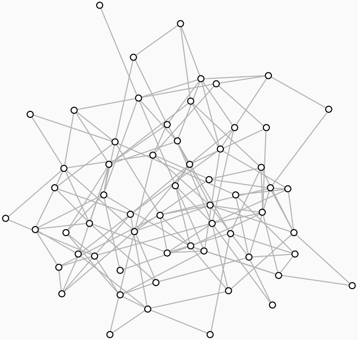
$$C_i = \frac{2E_i}{k_i(k_i - 1)}.$$

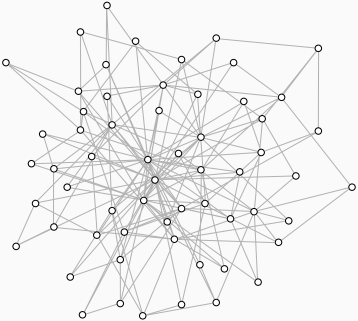
Answer:

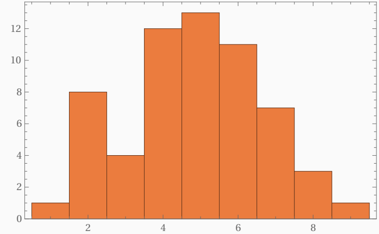
$$\langle C \rangle = \frac{p\langle k \rangle(\langle k \rangle - 1)}{\langle k \rangle(\langle k \rangle - 1)} = p.$$

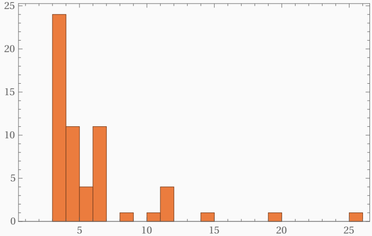
Such small clustering coefficient is not observed in the real life networks...

Non-physicality of ER graphs







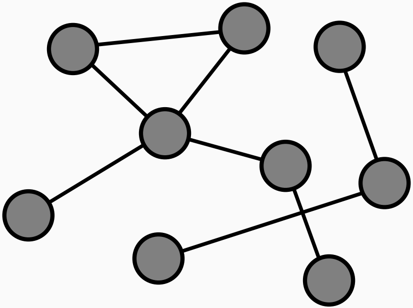


In the following classes, a more realistic BA model.↗

Configuration model

Idea:

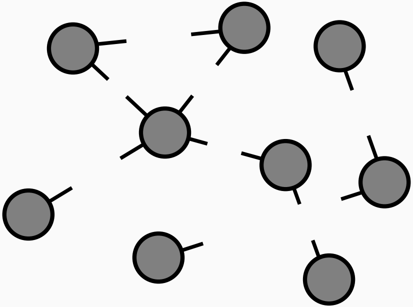
Let's build a graph from the bricks we have.



Configuration model

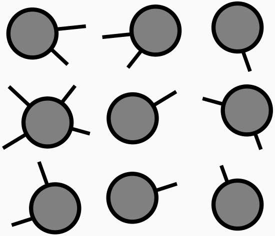
Idea:

Let's build a graph from the bricks we have.



Idea:

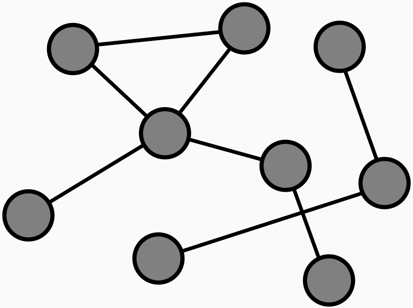
Let's build a graph from the bricks we have.



Configuration model

Idea:

Let's build a graph from the bricks we have.



What if you replace the actual bricks with selected *a priori*?

Configuration model

Let's choose thedegree configuration according to the expected distribution:

An example:

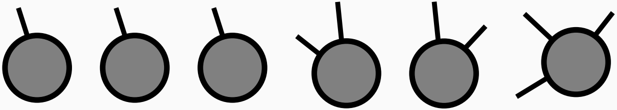
$$\{1, 1, 1, 2, 2, 3\}$$

Configuration model

Let's choose thedegree configuration according to the expected distribution:

An example:

$$\{1, 1, 1, 2, 2, 3\}$$



The only thing left to do is put these bricks together ...

Is it always possible?

Configuration model

Let's choose thedegree configuration according to the expected distribution:

An example:

$$\{1, 1, 1, 2, 2, 3\}$$

The only thing left to do is put these bricks together …

Is it always possible?

Question:

Find a (small) counterexample that the proposed procedure may not always be completed.

Stochastic block model

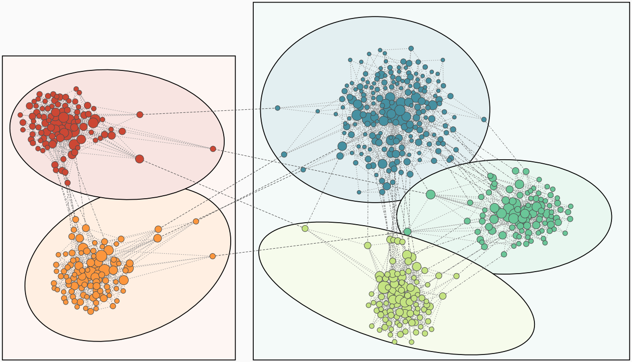
Generalization of the ER graphs

$$\begin{bmatrix} [p_{11}] & [p_{12}] & \dots & [p_{1N}] \\ [p_{21}] & [p_{22}] & \dots & [p_{2N}] \\ \dots & \dots & \ddots & \dots \\ [p_{N1}] & [p_{N2}] & \dots & [p_{NN}] \end{bmatrix}$$

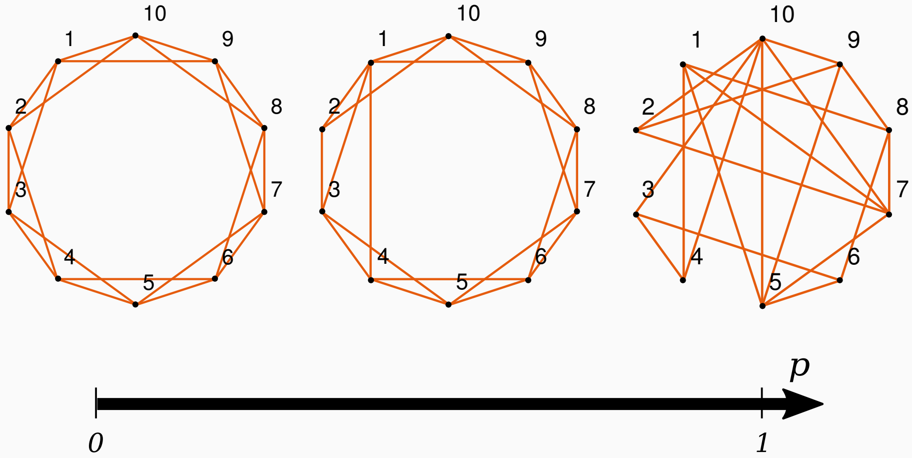
Stochastic block model

Generalization of the ER graphs

$$\begin{bmatrix} [p_{11}] & [p_{12}] & \dots & [p_{1N}] \\ [p_{21}] & [p_{22}] & \dots & [p_{2N}] \\ \dots & \dots & \ddots & \dots \\ [p_{N1}] & [p_{N2}] & \dots & [p_{NN}] \end{bmatrix}$$



Watts-Strogatz model



Networks with given hamiltonian

Let us considet the space of every possible graph with N vertices
i.e. set $M_N = \mathbb{M}^{N\times N}(\{0,1\})$. We want to define a probability distribution on it.

Networks with given hamiltonian

Let us considet the space of every possible graph with N vertices i.e. set $M_N = \mathbb{M}^{N\times N}(\{0,1\})$. We want to define a probability distribution on it.

We maximize entropy:

$$-\sum_{G\in M_N} \mathcal{P}(G) \ln \mathcal{P}(G),$$

Networks with given hamiltonian

Let us considet the space of every possible graph with N vertices i.e. set $M_N = \mathbb{M}^{N\times N}(\{0,1\})$. We want to define a probability distribution on it.

We maximize entropy:

$$-\sum_{G\in M_N} \mathcal{P}(G)\ln\mathcal{P}(G),$$

Under certain condition $f(\mathcal{P}(G)) = 0$,

Networks with given hamiltonian

Let us considet the space of every possible graph with N vertices i.e. set $M_N = \mathbb{M}^{N \times N}(\{0,1\})$. We want to define a probability distribution on it.

We maximize entropy:

$$-\sum_{G \in M_N} \mathcal{P}(G) \ln \mathcal{P}(G),$$

Under certain condition $f(\mathcal{P}(G)) = 0$,

Which leads us to Lagrange multipliers

$$\mathcal{L}[\mathcal{P}(G)] = -\sum_{G \in M_N} \mathcal{P}(G) \ln \mathcal{P}(G) + \lambda f(\mathcal{P}(G))$$

Networks with given hamiltonian

Let us considet the space of every possible graph with N vertices i.e. set $M_N = \mathbb{M}^{N\times N}(\{0,1\})$. We want to define a probability distribution on it.

We maximize entropy:

$$-\sum_{G\in M_N} \mathcal{P}(G)\ln\mathcal{P}(G),$$

Under certain condition $f(\mathcal{P}(G)) = 0$,

Which leads us to Lagrange multipliers

$$\mathcal{L}[\mathcal{P}(G)] = -\sum_{G\in M_N} \mathcal{P}(G)\ln\mathcal{P}(G) + \lambda f(\mathcal{P}(G))$$

It *only* remains to solve this equation

$$\frac{\partial\mathcal{L}}{\partial\mathcal{P}(G)} = 0.$$

Summary

Homework:

Read about:

- Dulbecco rule,
- Matthew effect,
- rich get richer rule,
- Yule processes.

Suggested source:

- M. Perc, *Journal of The Royal Society Interface* **11** (2014)

European Funds
Knowledge Education Development

Warsaw University of Technology

European Union
European Social Fund

MSc program in Data Science has been developed
as a part of task 10 of the project
„NERW PW. Science - Education - Development - Cooperation"
co-funded by European Union from European Social Fund.

Thank you for your attention!