

Náhodné siete

Beáta Stehlíková

2-EFM-155 Analýza sociálnych sietí

Fakulta matematiky, fyziky a informatiky, UK v Bratislave, 2019

Erdős–Rényi

Vplyv pravdepodobnosti p

Opakovanie: n vrcholov, každá hrana vznikne s pravdepodobnosťou p nezávisle od ostatných.

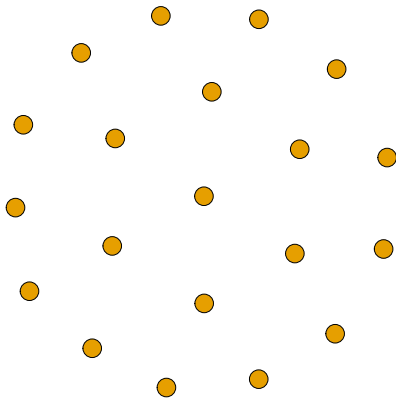
Pozrime sa na vplyv parametra p :

```
set.seed(123)
g1 <- erdos.renyi.game(20, p = 0.001)
g2 <- erdos.renyi.game(20, p = 0.003)
g3 <- erdos.renyi.game(20, p = 0.1)
g4 <- erdos.renyi.game(20, p = 0.25)
```

Veľmi malé p

S veľkou pravdepodobnosťou nevznikne žiadna hrana.

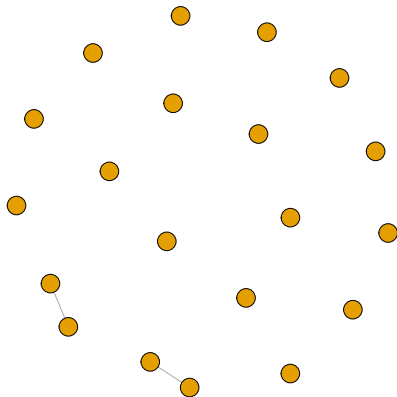
```
plot(g1, vertex.size = 10, vertex.label = NA)
```



Zväčšujeme p

S veľkou pravdepodobnosťou vznikne aspoň jedna hrana

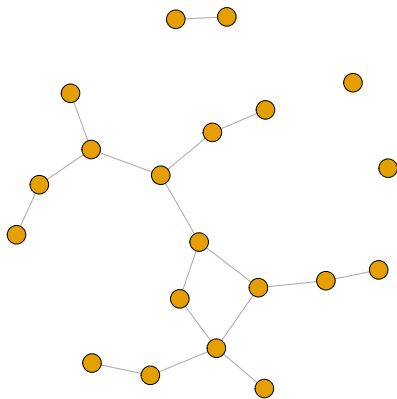
```
plot(g2, vertex.size = 10, vertex.label = NA)
```



Zväčšujeme p

S veľkou pravdepodobnosťou vznikne *giant component*

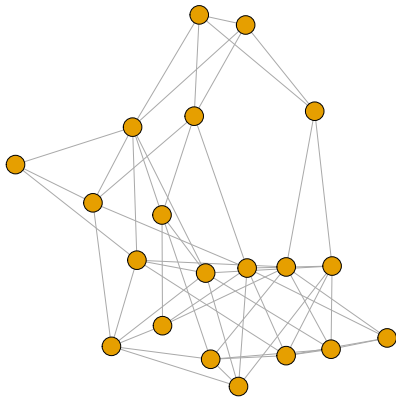
```
plot(g3, vertex.size = 10, vertex.label = NA)
```



Zväčšujeme p

S veľkou pravdepodobnosťou je sieť súvislá

```
plot(g4, vertex.size = 10, vertex.label = NA)
```



Monotónne vlastnosti - nepokazia sa pridaním hrany

Napríklad:

- ▶ sieť obsahuje aspoň jednu hranu
- ▶ sieť obsahuje cyklus
- ▶ sieť je súvislá

Čo napríklad nie je monotónna vlastnosť:

- ▶ sieť obsahuje vrchol s nepárnym stupňom

Prahová funkcia (threshold)

Funkcia $t(n)$ sa nazýva **prahová funkcia** pre určitú monotónnu vlastnosť siete, ak platí pre $n \rightarrow \infty$:

$$P[A(N)] \rightarrow 1 \text{ pre } p(n)/t(n) \rightarrow \infty,$$

$$P[A(N)] \rightarrow 0 \text{ pre } p(n)/t(n) \rightarrow 0,$$

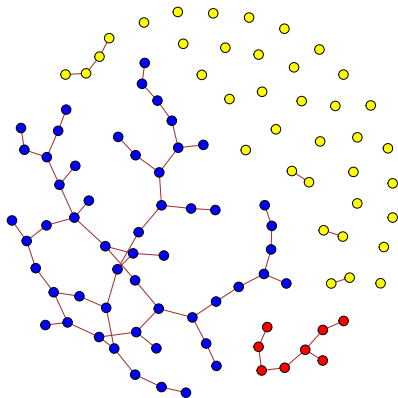
kde $A(N)$ znamená, že náhodne vygenerovaná sieť N má uvažovanú vlastnosť (napr. "sieť je súvislá")

Príklad: Pre existenciu cyklu je prahová funkcia $1/n$. Takže napríklad ak budeme generovať ER siete s $p(n) = 1/\sqrt{n}$, tak pre $n \rightarrow \infty$ bude pravdepodobnosť výskytu cyklu konvergovať k 1.

Cvičenie: Využitím matice susednosti navrhnete postup, ako zistiť, či sieť obsahuje cyklus. Naprogramujte generovanie sietí s $p(n) = 1/\sqrt{n}$ a zistite, v koľkých prípadoch sieť obsahovala cyklus.

Giant component

Cvičenie: Prečo názov “giant component”. Pre zadanú sieť vypočítajme podiel počtu vrcholov v najväčšom a druhom najväčšom komponente (ak sú dva najväčšie, bude to 1). Znovu spravíme štatistiku získaných hodnôt. (Obr. je pre $n = 100$ a $p = 0.015$)



Giant component

Cvičenie: AKý veľký je giant component Zobrazte štatistiku o veľkosti najväčšieho komponentu.

Na predchádzajúcom obrázku obsahuje najväčší komponent 54 vrcholov, teda 54 percent vrcholov siete.

Giant component

Intuitívne odvodíme **aproximáciu podielu vrcholov v najväčšom komponente**

- ▶ q = hľadaný podiel vrcholov
- ▶ pozrieme sa na náhodne zvolený vrchol, pravdepodobnosť, že je v najväčšom komponente je q
- ▶ pravd., že tento vrchol nie je v najväčšom komponente je pravd., že všetci jeho susedia sú mimo tohto komponentu
- ▶ a to je $(1 - q)^d$, kde d je stupeň vrchola

$$1 - q = \sum (1 - q)^d P(d),$$

kde $P(d)$ je pravdepodobnosť, že vrchol má stupeň d

V ER sieťach je rozdelenie binomické, ktoré sa aproximuje Poissonovým s parametrom $\lambda = (n - 1)p$ (preto sa im hovorí aj *Poisson networks*).

Giant component

Hľadané q je riešením rovnice

$$q = 1 - e^{-q(n-1)p}$$

Táto rovnica má riešenie z $(0, 1)$ práve vtedy, keď $p > 1/(n-1)$.
<https://press.princeton.edu/titles/8767.html>, ukážková kapitola 4,
poznámka pod čiarou 23

Cvičenie: Ako sa to zhoduje s našimi simuláciami?

Giant component - aplikácia

Model pre šírenie infekcie (*Jackson, kapitola 4.3*):

- ▶ n = celkový počet ľudí
- ▶ π = podiel tých, ktorí majú imunitu
- ▶ p = pravdepodobnosť kontaktu (kontakty modelujeme ER sieťou), pri nej sa človek bez imunity nakazí
- ▶ jeden vrchol sa nakazí sledovanou chorobou

Model v reči sietí:

- ▶ počet vrcholov n
- ▶ z nich náhodne odoberieme $100 \times \pi$ percent vrcholov
- ▶ medzi ostatnými vznikajú hrany (kontakty) podľa ER modelu

Giant component - aplikácia

Vieme:

- ▶ pre aké p, n, π vzniká veľký komponent
- ▶ ak nevzniká, nenastane epidémia
- ▶ odhadnúť podiel vrcholov vo veľkom komponente q , teda očakávaný podiel ľudí nakazených chorobou

Watts-Strogratz

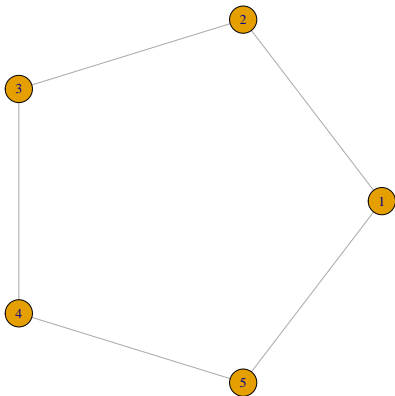
Začínáme zo symetrickej siete typu lattice (neskôr pridáme parameter p), napr.

```
g1 <- watts.strogatz.game(dim=1, size=5, nei=1, p=0)
g2 <- watts.strogatz.game(dim=1, size=7, nei=2, p=0)
g3 <- watts.strogatz.game(dim=2, size=4, nei=1, p=0)
```

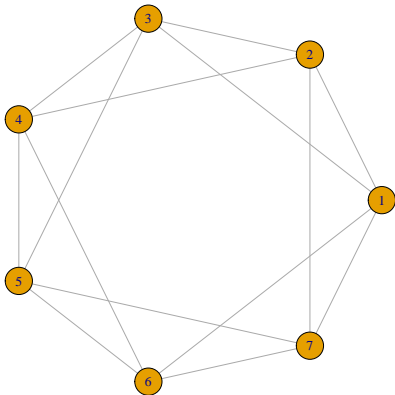
Parametre:

- ▶ `dim` - Integer constant, the dimension of the starting lattice.
size
- ▶ `size` - Integer constant, the size of the lattice along each dimension.
- ▶ `nei` - Integer constant, the neighborhood within which the vertices of the lattice will be connected

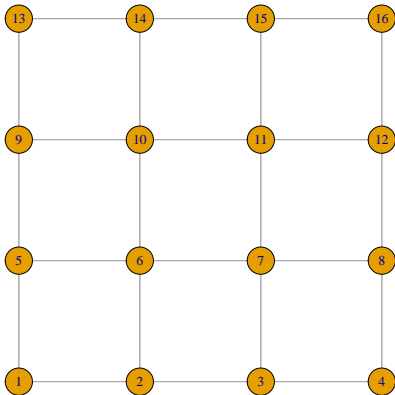
```
plot(g1, layout = layout_in_circle)
```



```
plot(g2, layout = layout_in_circle)
```



```
plot(g3, layout = layout_on_grid)
```



Typické vlastnosti pre veľké n

ER model

- ▶ malá miera zhlukovania (videli sme na obrázkoch)

Lattice

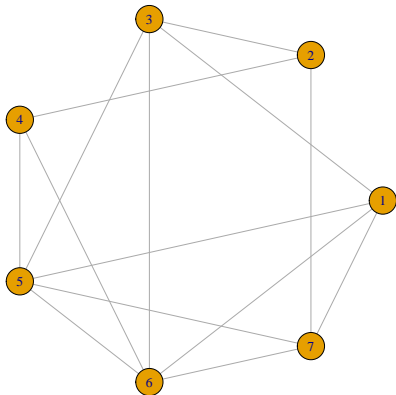
- ▶ vysoká miera zhlukovania
- ▶ ale príliš pravidelné
- ▶ a má veľký priemer (ako ďaleko sú dva najvzdialenejšie vrcholy)

Nový model - Watts-Strigatz

- ▶ v sieti lattice s pravdepodobnosťou p rovnomerne náhodne poprepájame spojené vrcholy
- ▶ pri dosť veľkom p sa zníži priemer
- ▶ ale nemalo by byť príliš veľké, aby sa zachovala vysoká miera zhlukovania

Ukážka

```
g <- watts.strogatz.game(dim=1, size=7, nei=2, p=0.1)
g <- simplify(g) # mozu vzniknut viacnasoby hrany a slucky
plot(g, layout = layout_in_circle)
```

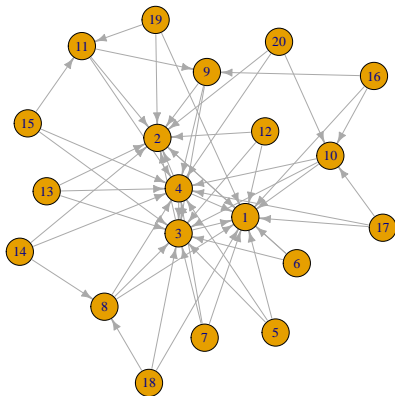


Barabasi - Albert

- ▶ Orientovaná sieť, ktorá vzniká postupne, v každom kroku pridáme m vrcholov
- ▶ Pravdepodobnosť, že nový vrchol bude spojený s niektorým zo starých, je úmerná tomu, koľko hrán už daný vrchol má
- ▶ Dá sa získať aj neorientovaná sieť

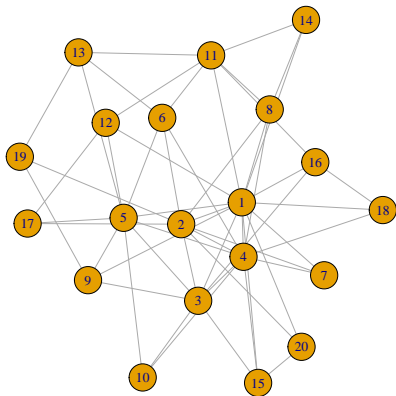
Príklad

```
g <- barabasi.game(n=20, m=3)  
plot(g, edge.arrow.size = 0.7)
```



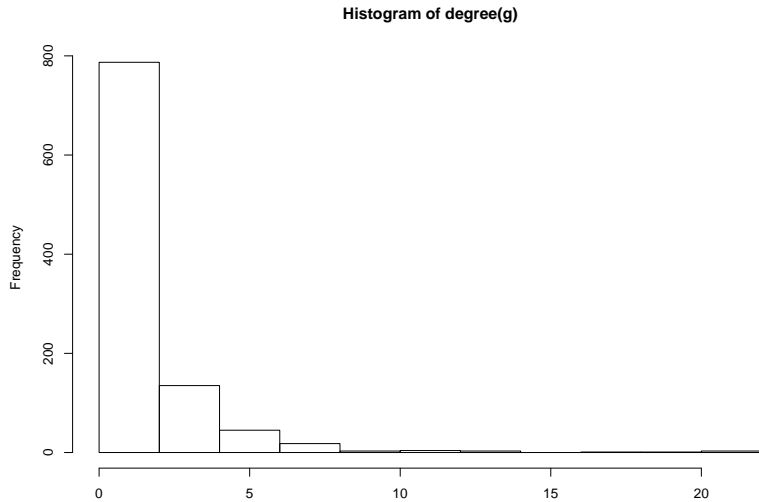
Príklad

```
g <- barabasi.game(n=20, m=3, directed = FALSE)  
plot(g)
```



Rozdelenie stupňa vrcholov

```
g <- barabasi.game(n=1000, directed = FALSE)  
hist(degree(g))
```



Štatistické modely

Základná myšlienka

Pravdepodobnosť, či hrana medzi dvoma vrcholmi vznikne, závisí od

- ▶ atribútov vrchola (prax, vek, ...)
- ▶ porovnanie atribútov vrcholov, medzi ktorými môže vzniknúť (oddelenie na ktorom pracujú, ...)
- ▶ charakteristík siete (celkový počet hrán, počet spoločných susedov danej dvojice vrcholov)

a parametre sa odhadujú štatistickými metódami

Tzv. **ERGM** (*Exponential random graph models*)

V R-ku: <https://cran.r-project.org/web/packages/ergm/index.html>