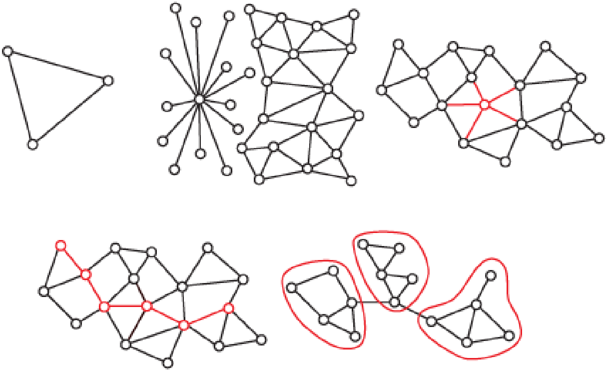
Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#281 Networks

été 2020

Networks



Networks

Let $V = \{1, \cdots, n_V\}$ denote either nodes, or vertices (n_V is the order). Let $E \in \{0,1\}^{n_V \times n_V}$ represents the relationships, through an adjacency matrix $\boldsymbol{A}$, $A_{i,j} = 1$ indicates a link - or edge - between i and j, or a collection of links $\{e_1, \cdots, e_{n_E}\}$. Let $n_E = |E|$ denote the number of edges, called size.
The degree $d(\cdot)$ of a vertice v is its number of incident edges.
A network is a pair (V, E)

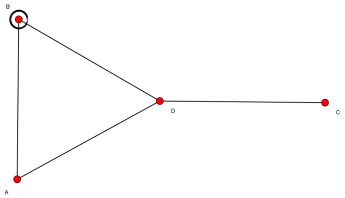
- $G =$ Internet, $V =$ computers, $E =$ IP network adjacency
- $G =$ World Wide Web, $V =$ web pages, $E =$ hyperlink
- $G =$ Articles, $V =$ authors, $E =$ citations
- $G =$ Friendship Network, $V =$ persons, $E =$ friendship
- $G =$ Airport Network, $V =$ airports, $E =$ non-stop flight

Networks

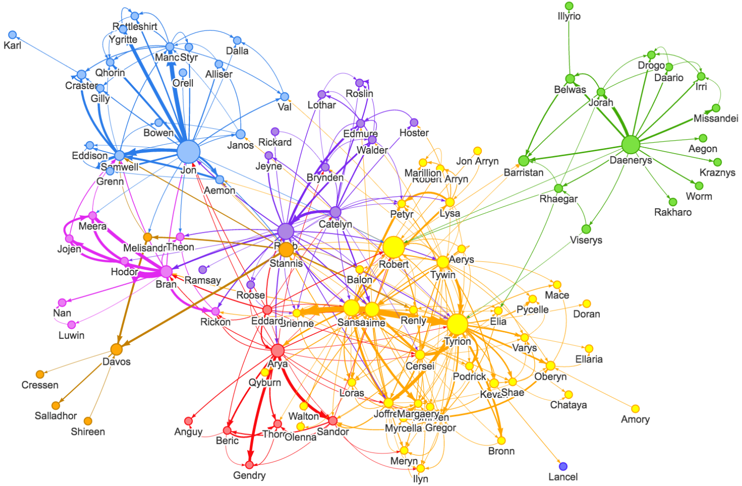
Adjacency matrix $\boldsymbol{A}$, $A_{i,j} = 1$ indicates a link between i and j
The degree of a vertice v is $d(v)$, the number of vertices in V incident to v (i.e. the number of neighbors of v)
Row i contains list of vertices connected to vertice i.

$$d(i) = \sum_{j=1}^{n_V} A_{i,j} = \boldsymbol{A}_{i,\cdot}^{\top}\mathbf{1}.$$

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

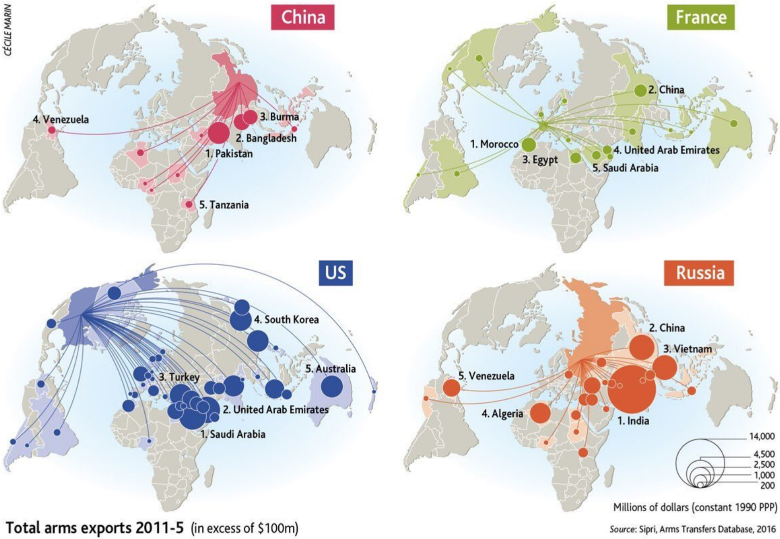


Networks

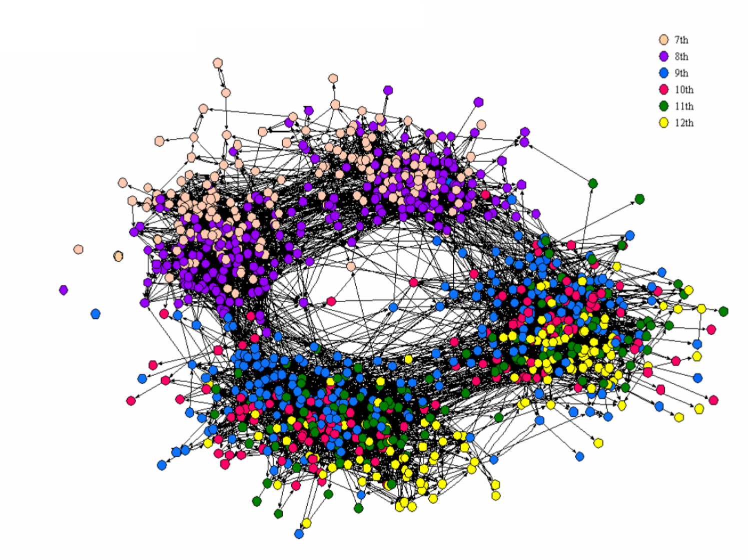


Karl
Rattleshirt
Ygritte
Manc
Styr
Dalla
Qhorin
Craster
Gilly
Orell
Alliser
Val
Bowen
Janos
Eddison
Samwell
Grenn
Jon
Aemon
Rickard
Jeyne
Lothar
Roslin
Edmure
Walder
Hoster
Brynden
Catelyn
Meera
Jojen
Melisandre
Theon
Hodor
Bran
Ramsay
Stannis
Roose
Nan
Luwin
Rickon
Eddard
Brienne
Davos
Cressen
Salladhor
Shireen
Arya
Qyburn
Anguy
Beric
Thoros
Gendry
Walton
Olenna
Sandor
Marillion
Jon Arryn
Robert Arryn
Petyr
Lysa
Robert
Aerys
Tywin
Balon
Renly
Tyrion
Cersei
Loras
Joffrey
Margaery
Myrcella
Gregor
Meryn
Ilyn
Podrick
Shae
Bronn
Lancel
Mace
Pycelle
Doran
Elia
Varys
Ellaria
Oberyn
Chataya
Amory
Illyrio
Belwas
Jorah
Drogo
Daario
Irri
Missandei
Aegon
Kraznys
Worm
Rakharo
Daenerys
Barristan
Rhaegar
Viserys

Networks

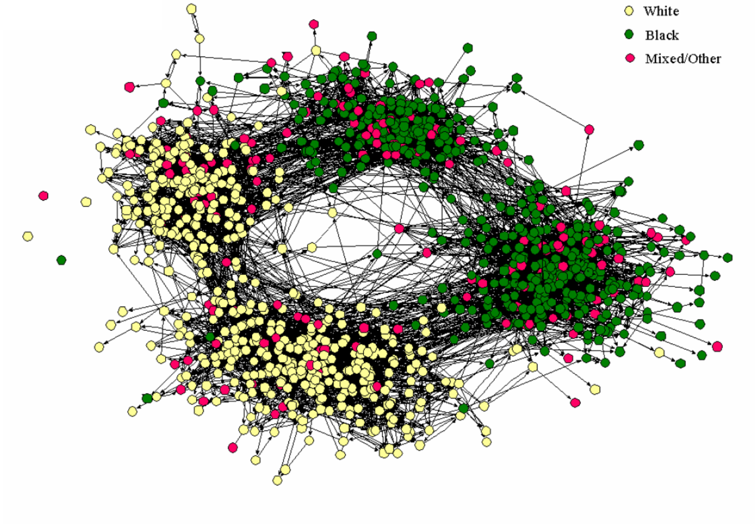

CÉCILE MARIN
China
4. Venezuela
3. Burma
2. Bangladesh
1. Pakistan
5. Tanzania
France
2. China
1. Morocco
3. Egypt
4. United Arab Emirates
5. Saudi Arabia
US
4. South Korea
5. Australia
3. Turkey
2. United Arab Emirates
1. Saudi Arabia
Russia
2. China
3. Vietnam
5. Venezuela
4. Algeria
1. India
14,000
4,500
2,500
1,000
200
Millions of dollars (constant 1990 PPP)
Total arms exports 2011-5 (in excess of $100m)
Source: Sipri, Arms Transfers Database, 2016

Networks



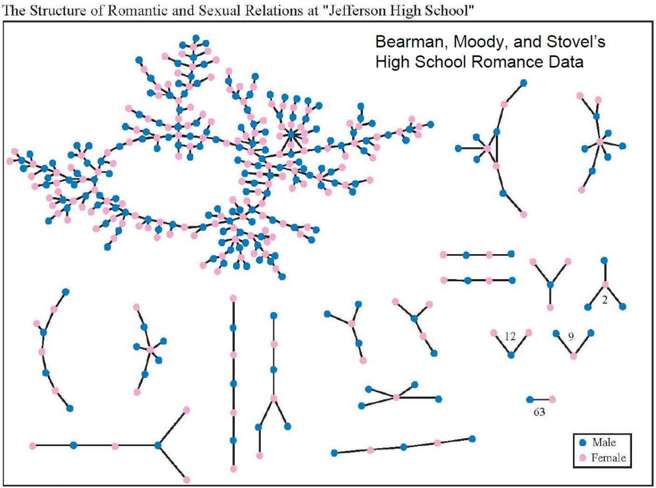
Race, School Integration and Friendship Segregation in America

Networks



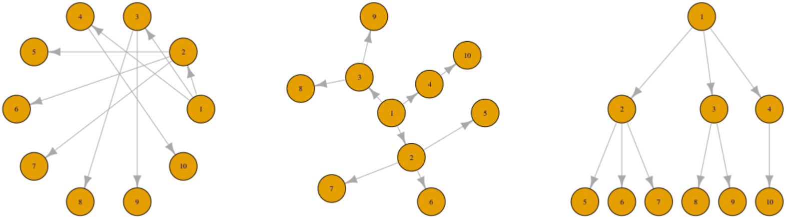
Race, School Integration and Friendship Segregation in America

Networks



See Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks

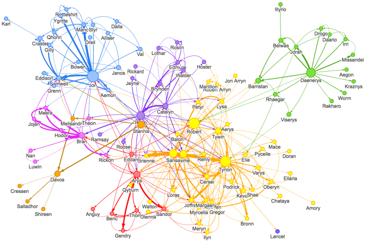
Networks

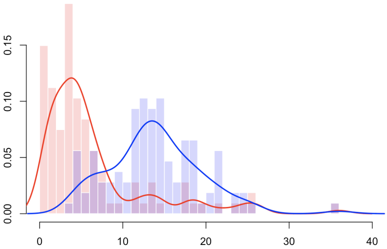


(same network, different representations)

Friendship Paradox

The average degree of a friend is strictly greater than the average degree of a random node. See Feld (1991) and Zuckerman & Jost (2001).





The average number of friends of a random person in the graph is

$$\mu = \frac{1}{n_V} \sum_{v \in V} d(v) = \frac{2n_E}{n_V} = \frac{1}{n_V} \|\boldsymbol{d}\|_1$$

where $\boldsymbol{d}$ is the vector of $d(v)$'s (or $\boldsymbol{d} = \boldsymbol{A}\mathbf{1}$).

Friendship Paradox

The average number of friends that a typical friend has is

$$\frac{1}{n_V}\sum_{v\in V}\left(\frac{1}{d(v)}\sum_{v':(v,v')\in E} d(v')\right)=\frac{1}{\|\boldsymbol{d}\|_1}\boldsymbol{d}^\top\boldsymbol{d}$$

We can prove that

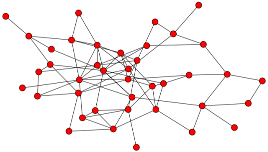
$$\frac{1}{\|\boldsymbol{d}\|_1}\boldsymbol{d}^\top\boldsymbol{d}\geq\frac{1}{n_V}\|\boldsymbol{d}\|_1$$

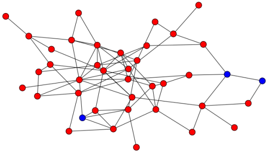
See also

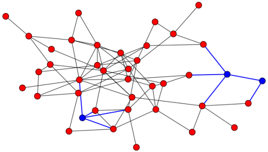
$$\frac{\mathbb{E}[D^2]}{\mathbb{E}[D]}=\mathbb{E}[D]+\frac{\text{Var}[D]}{\mathbb{E}[D]}\geq\mathbb{E}[D]$$

because $\text{Var}[D]\geq 0$.

Sampling & Networks







Surveying hard-to-reach groups through sampled respondents in a social network

Use Snowball Sampling (or Star Sampling)

Idea: start with an initial vertice sample V_0 and observe all incident edges, as well as all vertices sharing those edges. Iterate to have n edges.

Small Word Property

"*I read somewhere that everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice. Fill in the names. . . . How every person is a new door, opening up into other worlds. Six degrees of separation between me and everyone else on this planet. But to find the right six people...*" from Six Degrees of Separation

The small world property is mathematically intuitive: if the number of vertices within a given distance of a specific node growth exponentially with the distance, then the average path length increases a $\log n_V$.

Formally, we talk here about the average path length

$$\bar{\ell} = \binom{n_V}{2}^{-1} \sum_{u \neq v} d(u, v) = O(\log n_V)$$

Small Word Property

Intuition: if $d_v = d$, if we reach everyone after k hops $n_V \sim d^k$ i.e. $k = O(\log n_V)$

"*Each person in the world (at least among the 1.59 billion people active on Facebook) is connected to every other person by an average of three and a half other people. The average distance we observe is 4.57, corresponding to 3.57 intermediaries or 'degrees of separation*"', from Three and a half degrees of separation

An alternative is based on the harmonic mean, related to graph efficiency

$$\ell^{-1} = \frac{2}{n_V(n_V - 1)} \sum_{i>j} d_{i,j}^{-1}$$