Statistical Properties of Networks

Network Science '21: Session 2

Claudio J. Tessone

Blockchain & Distributed Ledger Technologies Group, UZH Blockchain Center

Lecture Objectives

1. Which kinds of distributions will we see?
2. We will go beyond degree distributions to *higher order network properties*: Assortativity and clustering
3. We will see which networks display different of these properties
4. We will start to understand the necessity for benchmark models to compare our data with

Outlook

1

2

3 From Local to Global Properties

Shortest path

4 Directed Networks

Reciprocity

Network Motifs

5 Undirected networks

Clustering coefficient

6 Assortativity

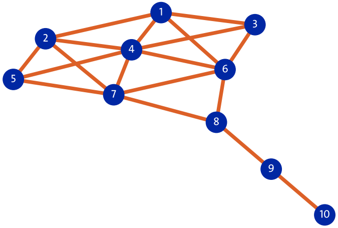
7 Network randomisation

Recap Lecture 1

Network representation

Points to...

Points from ...	1	2	3	4	5	6	7	8	9	10
1		1	1	1		1				
2	1			1	1		1			
3	1			1		1				
4	1	1	1		1	1	1			
5		1		1			1			
6	1		1	1			1	1		
7		1		1	1	1		1		
8						1	1		1	
9								1		1
10									1	



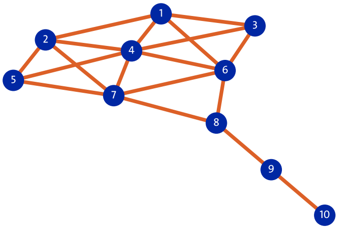
- The adjacency matrix $\mathbb{A} = \{a_{ij}\}_{i,j=1}^{N}$ is a representation of the network
- Symmetric for undirected networks

Node degree

Points to...

Points from node 4

	1	2	3	4	5	6	7	8	9	10
1		1	1	1		1				
2	1			1	1		1			
3	1			1		1				
4	1	1	1		1	1	1			
5		1		1			1			
6	1		1	1			1	1		
7		1		1	1	1		1		
8						1	1		1	
9								1		1
10									1	



The degree of a node can be computed as a sum over the adjacency matrix

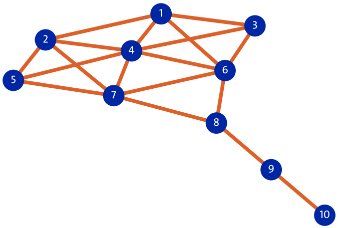
$$k_i = \sum_{j=1}^{N} a_{ij}$$

Node degree

Points to...

Points from node 4

	1	2	3	4	5	6	7	8	9	10
1		1	1	1		1				
2	1			1	1		1			
3	1			1		1				
4	1	1	1		1	1	1			
5		1		1			1			
6	1		1	1			1	1		
7		1		1	1	1		1		
8						1	1		1	
9								1		1
10									1	



The total number of links $L = |V|$ can be computed as a sum over the adjacency matrix,

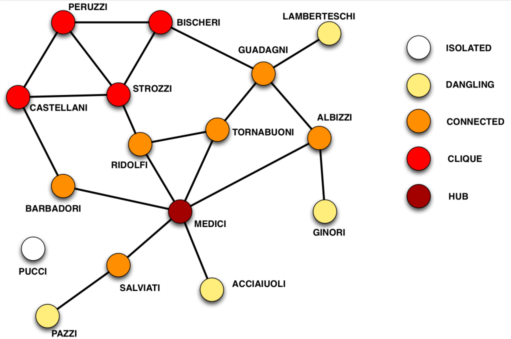
$$L = \sum_{i<j}^{N} a_{ij}$$

Degree is a first-order property of the nodes

$$k_i = \sum_{j=1}^{N} a_{ij}$$

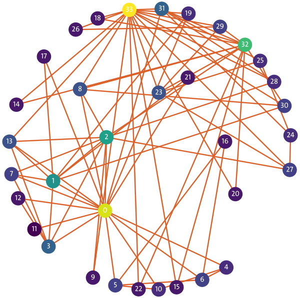
Node degree may hint at the importance of specific vertices

Node degree: interpretation



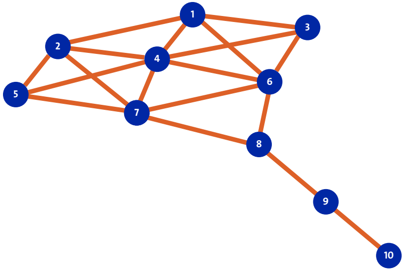
Graph of the relationships of the Florence families in in the Renaissance

Node degree: interpretation



Zachary Karate Club

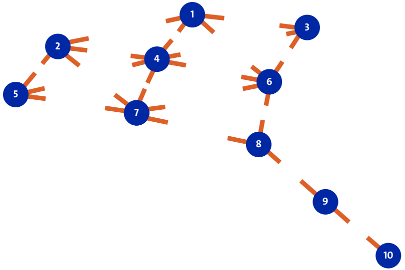
Degree distribution



Definition

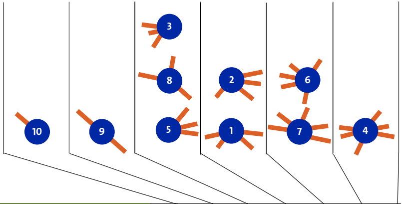
Denoted by $p(k)$. It is a function of the degree. Its value is the fraction of nodes with degree k

Degree distribution



This is the information we keep from the network

Degree distribution



Degree d	0	1	2	3	4	5	6
Degree distribution $P(d)$	0	0.1	0.1	0.3	0.2	0.2	0.1

These are the values of the degree distributions for the Kite network

Degree distribution

$$p(k) = \text{Prob}(k_i = k)$$

In many contexts of interest, the degree distribution does not have a characteristic scale

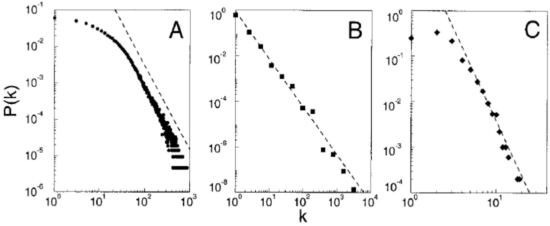
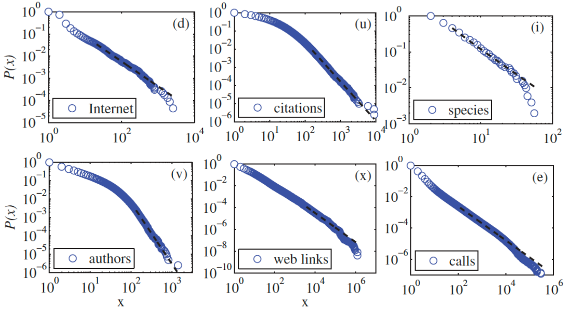


Fig. 1. The distribution function of connectivities for various large networks. (**A**) Actor collaboration graph with $N = 212{,}250$ vertices and average connectivity $\langle k \rangle = 28.78$. (**B**) WWW, $N = 325{,}729$, $\langle k \rangle = 5.46$ (6). (**C**) Power grid data, $N = 4941$, $\langle k \rangle = 2.67$. The dashed lines have slopes (A) $\gamma_{actor} = 2.3$, (B) $\gamma_{www} = 2.1$ and (C) $\gamma_{power} = 4$.

Degree distribution

$$p(k) = \text{Prob}(k_i = k)$$

In many contexts of interest, the degree distribution does not have a characteristic scale



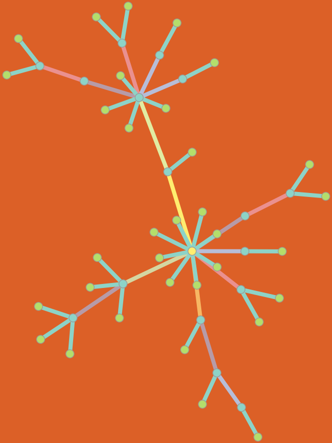
The reason for this universality will be explained in Session 6

A note on nomenclature

The notation $\langle \cdot \rangle$ represents the average of the argument.

If the index is obvious, it is omitted

So far we have seen networks ...



... like this

From Local to Global Properties

Some examples

+ **Average network degree:**

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i$$

+ **Average link strength (for weighted networks):**

$$\langle s \rangle = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} s_{ij}$$

Remember that N represents always the number of nodes

Shortest path

- Shortest path $\iff$ distance between two nodes;
- By definition, shortest path *between a node and itself* is zero;
- If there is *no path* connecting two nodes, shortest path is defined to be *infinite*
- $\mathbb{A}^r$ gives the number of walks of length **r** between each pair of nodes

Shortest path

$$d_{i,j} = \text{argmin}_r \left([\mathbb{A}^r]_{i,j} \neq 0 \right) \quad (1)$$

Shortest path

- Shortest path $\iff$ distance between two nodes;
- By definition, shortest path *between a node and itself* is zero;
- If there is *no path* connecting two nodes, shortest path is defined to be *infinite*
- $\mathbb{A}^r$ gives the number of walks of length **r** between each pair of nodes

Shortest path

$$d_{i,j} = \text{argmin}_r \left([\mathbb{A}^r]_{i,j} \neq 0 \right) \quad (1)$$

Shortest path

- Shortest path $\iff$ distance between two nodes;
- By definition, shortest path *between a node and itself* is zero;
- If there is *no path* connecting two nodes, shortest path is defined to be *infinite*
- $\mathbb{A}^r$ gives the number of walks of length **r** between each pair of nodes

Shortest path

$$d_{i,j} = \text{argmin}_r \left([\mathbb{A}^r]_{i,j} \neq 0 \right) \quad (1)$$

Shortest path

- Shortest path $\iff$ distance between two nodes;
- By definition, shortest path *between a node and itself* is zero;
- If there is *no path* connecting two nodes, shortest path is defined to be *infinite*
- $\mathbb{A}^r$ gives the number of walks of length **r** between each pair of nodes

Shortest path

$$d_{i,j} = \text{argmin}_r \left([\mathbb{A}^r]_{i,j} \neq 0 \right) \quad (1)$$

Shortest path

- Shortest path $\iff$ distance between two nodes;
- By definition, shortest path *between a node and itself* is zero;
- If there is *no path* connecting two nodes, shortest path is defined to be *infinite*
- $\mathbb{A}^r$ gives the number of walks of length **r** between each pair of nodes

Shortest path

$$d_{i,j} = \text{argmin}_r \left([\mathbb{A}^r]_{i,j} \neq 0 \right) \tag{1}$$

Related concepts

- **Average distance** in a network is the average length of the shortest path between two nodes.

$$\langle d_{i,j} \rangle = \frac{1}{(N-1)^2} \sum_{i,j=1}^{N} d_{i,j}$$

- **Eccentrity** of a vertex $\epsilon(v)$ is the greatest distance between v and any other vertex
- **Diameter** of a graph is maximum eccentrity of any vertex in the graph (or the largest distance between pairs of nodes)

$$D = \max_{i,j} \left(d_{i,j} \right)$$

- A **peripheral vertex** is the one whose eccentrity achieves the diameter

Related concepts

- **Average distance** in a network is the average length of the shortest path between two nodes.

$$\langle d_{i,j} \rangle = \frac{1}{(N-1)^2} \sum_{i,j=1}^{N} d_{i,j}$$

- **Eccentrity** of a vertex $\epsilon(v)$ is the greatest distance between v and any other vertex
- **Diameter** of a graph is maximum eccentrity of any vertex in the graph (or the largest distance between pairs of nodes)

$$D = \max_{i,j} \left(d_{i,j} \right)$$

- A **peripheral vertex** is the one whose eccentrity achieves the diameter

Related concepts

+ **Average distance** in a network is the average length of the shortest path between two nodes.

$$\langle d_{i,j} \rangle = \frac{1}{(N-1)^2} \sum_{i,j=1}^{N} d_{i,j}$$

+ **Eccentrity** of a vertex $\epsilon(v)$ is the greatest distance between v and any other vertex
+ **Diameter** of a graph is maximum eccentrity of any vertex in the graph (or the largest distance between pairs of nodes)

$$D = \max_{i,j} \left(d_{i,j} \right)$$

+ A **peripheral vertex** is the one whose eccentrity achieves the diameter

Related concepts

- **Average distance** in a network is the average length of the shortest path between two nodes.

$$\langle d_{i,j} \rangle = \frac{1}{(N-1)^2} \sum_{i,j=1}^{N} d_{i,j}$$

- **Eccentrity** of a vertex $\epsilon(v)$ is the greatest distance between v and any other vertex
- **Diameter** of a graph is maximum eccentrity of any vertex in the graph (or the largest distance between pairs of nodes)

$$D = \max_{i,j} \left(d_{i,j} \right)$$

- A **peripheral vertex** is the one whose eccentrity achieves the diameter

Directed Networks

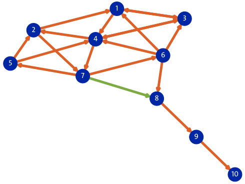
Are relationships mutual? If i claims to be friend of j, does j think the same about i?

Reciprocity

For directed networks, reciprocity measures the degree of how mutual relationships are

... to...

points from ...	1	2	3	4	5	6	7	8	9	10
1	0	0	1	1	0	0	0	0	0	0
2	1	0	0	0	0	0	1	0	0	0
3	1	0	0	1	0	0	0	0	0	0
4	0	1	1	0	0	0	1	0	0	0
5	0	1	0	1	0	0	0	0	0	0
6	1	0	1	1	0	0	0	1	0	0
7	0	0	0	0	1	1	0	1	0	0
8	0	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	0	0



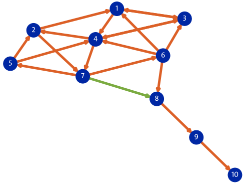
$$R = \frac{\sum_{u,v \in V} a_{u,v} a_{v,u}}{2 \sum_{u,v \in V} a_{u,v}} \quad (2)$$

Reciprocity

For directed networks, reciprocity measures the degree of how mutual relationships are

... to...

points from ...	1	2	3	4	5	6	7	8	9	10
1	0	0	1	1	0	0	0	0	0	0
2	1	0	0	0	0	0	1	0	0	0
3	1	0	0	1	0	0	0	0	0	0
4	0	1	1	0	0	0	1	0	0	0
5	0	1	0	1	0	0	0	0	0	0
6	1	0	1	1	0	0	0	1	0	0
7	0	0	0	0	1	1	0	1	0	0
8	0	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	0	0



$$R = \frac{\sum_{u,v \in V} a_{u,v} a_{v,u}}{2 \sum_{u,v \in V} a_{u,v}} \tag{2}$$

Why does it work?

- $a_{u,v}a_{v,u}$ is only equal to 1 if both edges exist

Reciprocity is a second-order property because it involves two nodes at the same time

Conversely, degree is a first order property

How is friendship data collected? Surveys!

Empirical results

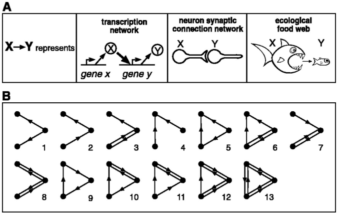
Reciprocity by Gender and Race (percentages

	Males		Females	
	Reciprocal	Nonreciprocal	Reciprocal	Nonreciprocal
Interracial	48.91%	51.09%	65.29%	34.71%
Intraracial	58.87	41.13	72.61	27.61
White	61.00	39.00	75.78	24.22
Black	46.17	53.83	62.39	37.61
Hispanic	48.75	51.25	65.35	34.65
Asian	59.44	40.56	69.09	30.91
Native American	52.22	47.78	59.04	40.96
Overall	56.98	43.02	71.49	28.51
N	10,635	8,030	15,743	6,278

"Friendship Reciprocity and Its Effects on School Outcomes among Adolescents" [E. Vaquera, G. Kao, *Social Science Research* (2008)]

- Many friendship ties are non-reciprocal
- Reciprocity decreases across self-perceived groups
- In organisations, reciprocity is higher within individuals of the same rank, much lower across ranks

Network motifs: reconstructing the network



It is possible to decompose the network into smaller sub-graphs representing different smaller-scale processes

Undirected networks

How likely that two of my friends are friends among themselves?

Clustering coefficient

How likely is it that if two nodes are common neighbours of another, they are connected to each other?

Each node i is endowed with a clustering coefficient C_i

$$C_i = \frac{\sum_{u,v \in V} a_{u,i} a_{v,i} a_{u,v}}{\sum_{u,v \in \mathcal{N}_i} a_{u,i} a_{v,i}} = \frac{2 \sum_{u,v \in \mathcal{N}_i} a_{u,v}}{k_i (k_i - 1)} \tag{3}$$

- $\mathcal{N}_i$ is the set of neighbours of node i
- *number of closed triangles divided over number of all possible triangles*

Clustering coefficient

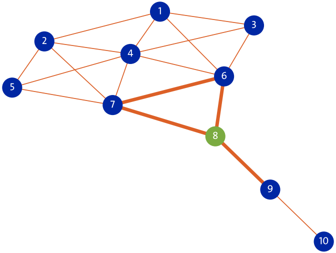
How likely is it that if two nodes are common neighbours of another, they are connected to each other?

Each node i is endowed with a clustering coefficient C_i

$$C_i = \frac{\sum_{u,v \in V} a_{u,i} a_{v,i} a_{u,v}}{\sum_{u,v \in \mathcal{N}_i} a_{u,i} a_{v,i}} = \frac{2 \sum_{u,v \in \mathcal{N}_i} a_{u,v}}{k_i (k_i - 1)} \quad (3)$$

- $\mathcal{N}_i$ is the set of neighbours of node i
- *number of closed triangles divided over number of all possible triangles*

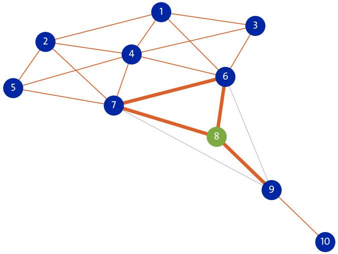
Clustering coefficient



For node 8, there is 1 triangle realised from 3 triangles possible

$$C_8 = \frac{1}{3}$$

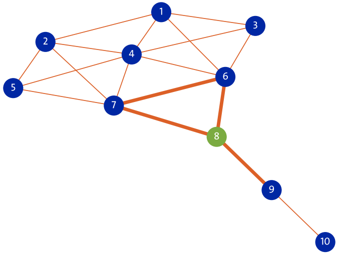
Clustering coefficient



For node 8, there is 1 triangle realised from 3 triangles possible

$$C_8 = \frac{1}{3}$$

Clustering coefficient



The list of all nodes' clustering

Node i	1	2	3	4	5	6	7	8	9	10
Clustering C_i	2/3	2/3	1	8/15	1/2	1	1/2	1/3	0	0

Global clustering coefficient

It gives a measure of the average value of the clustering coefficient in all the network

$$\langle C \rangle = \frac{1}{N} \sum_i C_i \tag{4}$$

For the example Kite network, $\langle C \rangle = 0.52$

Empirical values of the clustering coefficient

TABLE I. The general characteristics of several real networks. For each network we have indicated the number of nodes, the average degree $\langle k \rangle$, the average path length ℓ, and the clustering coefficient C. For a comparison we have included the average path length ℓ_{rand} and clustering coefficient C_{rand} of a random graph of the same size and average degree. The numbers in the last column are keyed to the symbols in Figs. 8 and 9.

Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference	Nr.
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999	1
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook *et al.*, 2001a, Pastor-Satorras *et al.*, 2001	2
Movie actors	225 226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz, 1998	3
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c	4
MEDLINE co-authorship	1 520 251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c	5
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c	6
NCSTRL co-authorship	11 994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c	7
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási *et al.*, 2001	8
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.*, 2001	9
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000	10
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000	11
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Solé, 2000	12
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Solé, 2000	13
Words, co-occurrence	460.902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Solé, 2001	14
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006	Yook *et al.*, 2001b	15
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998	16
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998	17

Empirical findings

- Real-world networks tend to have a large values of clustering

Assortativity

Homophily in social sciences refers to the propensity of an individual to link to others that are alike

Assortativity

- *Is is there homophily in the network? (nodes connect to others similar to them?)*
- Are nodes of low-degree connected with other low-degree nodes or are they mostly connected with high-degree nodes?
- Are nodes of high-degree connected with other low-degree nodes or are they mostly connected with high-degree nodes?

Assortativity

Assortativity is usually computed as the Pearson correlation coefficient of degree between pairs of linked nodes

$$r = \text{Corr}(k_i, k_j) = \frac{Cov(k_i, k_j)}{Var(k_i)^2} = \frac{\langle k_i k_j \rangle - \langle k_i \rangle \langle k_j \rangle}{\sigma^2} \quad (5)$$

where

$$\langle k_i k_j \rangle - \langle k_i \rangle \langle k_j \rangle = \sum_{(i,j)\in V} k_i k_j (P(k_i, k_j) - P(k_i)P(k_j)) \quad (6)$$

and

$$\sigma^2 = \sum_k k^2 P(k) - \left(\sum_k k P(k) \right)^2 \quad (7)$$

Assortativity

Assortativity is usually computed as the Pearson correlation coefficient of degree between pairs of linked nodes

$$r = \text{Corr}(k_i, k_j) = \frac{Cov(k_i, k_j)}{Var(k_i)^2} = \frac{\langle k_i k_j \rangle - \langle k_i \rangle \langle k_j \rangle}{\sigma^2} \tag{5}$$

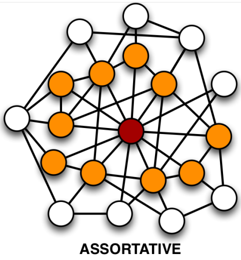
where

$$\langle k_i k_j \rangle - \langle k_i \rangle \langle k_j \rangle = \sum_{(i,j) \in V} k_i k_j (P(k_i, k_j) - P(k_i) P(k_j)) \tag{6}$$

and

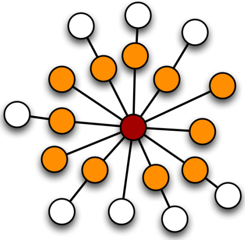
$$\sigma^2 = \sum_k k^2 P(k) - \left(\sum_k k P(k) \right)^2 \tag{7}$$

Interpretation: Assortativity



if $0 < r < 1$*, the network is said to be assortative: nodes with large (resp. low) degree have neighbours with large (resp. low) degree*

Interpretation: Dissortativity



if $-1 < r < 0$, the network is said to be dissortative: nodes with large (resp. low) degree have neighbours with low (resp. high) degree

Empirical values

Some real-world examples

Network	n	r
Physics coauthorship (a)	52 909	0.363
Biology coauthorship (a)	1 520 251	0.127
Mathematics coauthorship (b)	253 339	0.120
Film actor collaborations (c)	449 913	0.208
Company directors (d)	7 673	0.276
Internet (e)	10 697	−0.189
World-Wide Web (f)	269 504	−0.065
Protein interactions (g)	2 115	−0.156
Neural network (h)	307	−0.163
Marine food web (i)	134	−0.247
Freshwater food web (j)	92	−0.276

Empirical findings

- Technological networks → *dissortative*
- Social networks → *assortative*

Nearest neighbours' degree

Another means of capturing the degree correlation is by examining the properties of neighbour connectivity

$$k_{nn}(k_i) = \frac{1}{k_i} \sum_j a_{i,j} k_j \tag{8}$$

Nearest neighbours' degree

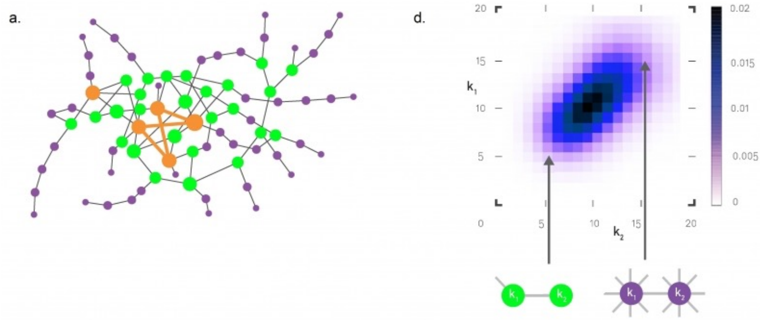
Another means of capturing the degree correlation is by examining the properties of neighbour connectivity

It is then averaged over all nodes with degree k

$$\langle k_{nn}(k) \rangle = \sum_q qP(q|k) \tag{8}$$

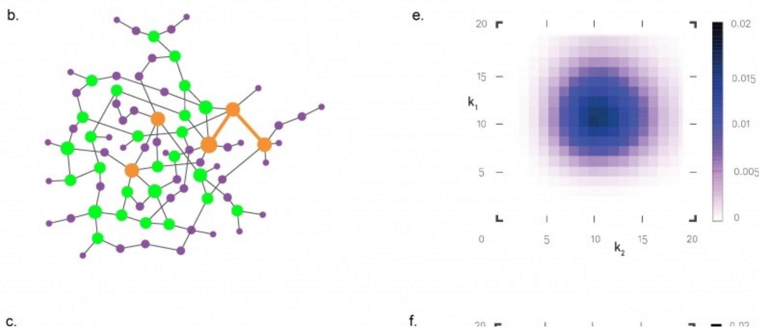
where $P(q|k)$ is the conditional probability that, given that the node degree is k, one of its edges points to a node whose degree is q

Assortativity



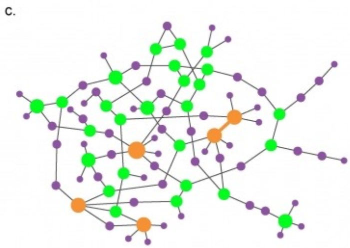
If $\langle k_{nn}(k) \rangle$ is increasing on k, the network is assortative

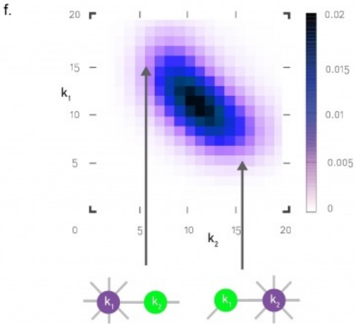
Assortativity



If $\langle k_{nn}(k) \rangle$ does not have a trend, the network is not assortative nor dissortative

Assortativity





If $\langle k_{nn}(k) \rangle$ is decreasing on k, the network is dissortative

Network randomisation

The typical research cycle

- Data collection
- Data analysis (e.g. measure clustering or assortativity)
- What is the meaning of the results? Is it relevant (statistically significant)?
- What is the origin of the result?

The approach taken is: compare the results with those coming from a benchmark model

The typical research cycle

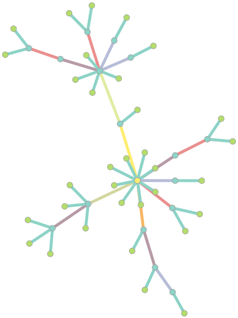
- Are the results of the benchmark model the same as in the data? Results are not significant
- Are the results of the benchmark model different from those in the data? Results are significant

Degree-preserving network randomisation

This approach asks the question if the results are purely the effect of the degree distribution

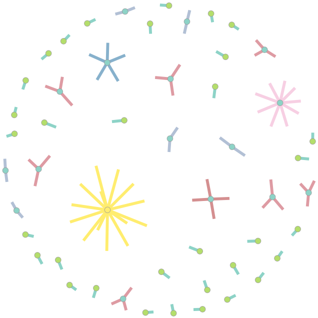
+ If we change the network completely, but leave the nodes' degrees unaltered… do we obtain the same results for the property we are measuring?
+ *If the result of our observation does not change… then, the result stems from the degree distribution*
+ If the result of our observation does change… then, the result comes from higher order network properties

Turning a network into pieces



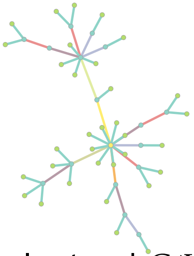
- Given a network $G(V, E)$ we cut the edges
- Each isolated node i has k_i *stubs*
- Randomly connecting pairs of stubs, we can build other networks with the same degree distribution...

Turning a network into pieces



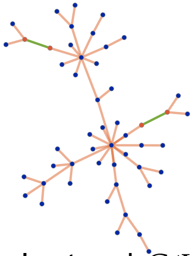
- Given a network $G(V, E)$ we cut the edges
- Each isolated node i has k_i *stubs*
- Randomly connecting pairs of stubs, we can build other networks with the same degree distribution...

A non-destructive approach



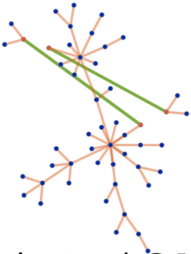
1. Start from the original network $G(V, E)$
2. Randomly select two edges
3. Exchange ends
4. Repeat steps 2-3 many times (at least, until each edge was selected - on average - once). We obtain G_{rnd}

A non-destructive approach



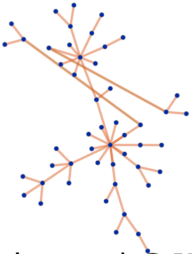
1. Start from the original network $G(V, E)$
2. Randomly select two edges
3. Exchange ends
4. Repeat steps 2-3 many times (at least, until each edge was selected - on average - once). We obtain G_{rnd}

A non-destructive approach



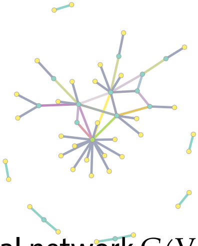
1. Start from the original network $G(V, E)$
2. Randomly select two edges
3. Exchange ends
4. Repeat steps 2-3 many times (at least, until each edge was selected - on average - once). We obtain G_{rnd}

A non-destructive approach



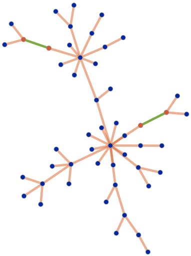
1. Start from the original network $G(V,E)$
2. Randomly select two edges
3. Exchange ends
4. Repeat steps 2-3 many times (at least, until each edge was selected - on average - once). We obtain G_{rnd}

A non-destructive approach



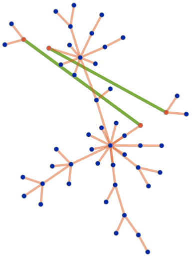
1. Start from the original network $G(V, E)$
2. Randomly select two edges
3. Exchange ends
4. Repeat steps 2-3 many times (at least, until each edge was selected - on average - once). We obtain G_{rnd}

No change in the degree distribution!



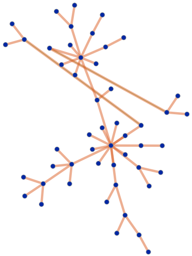
During an edge swap, the nodes affected do not change their degree!

No change in the degree distribution!



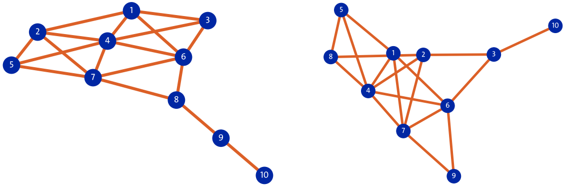
During an edge swap, the nodes affected do not change their degree!

No change in the degree distribution!



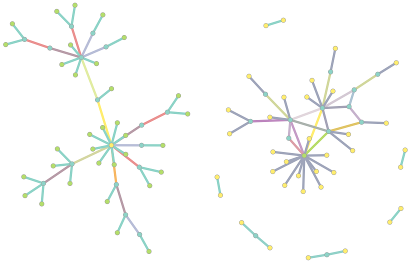
During an edge swap, the nodes affected do not change their degree!

Example in the kite network



Kite network before (left) and after (right) randomisation

Comparison of the results



Measure a property in G. Randomise. Measure the property in G_{rnd} and check if they are different

To attain statistical significance, multiple realisations of G_{rnd} are necessary

Claudio J. Tessone

**Blockchain & Distributed Ledger Technologies Group,
UZH Blockchain Center**

claudio.tessone@uzh.ch
http://www.ifi.uzh.ch/bdlt