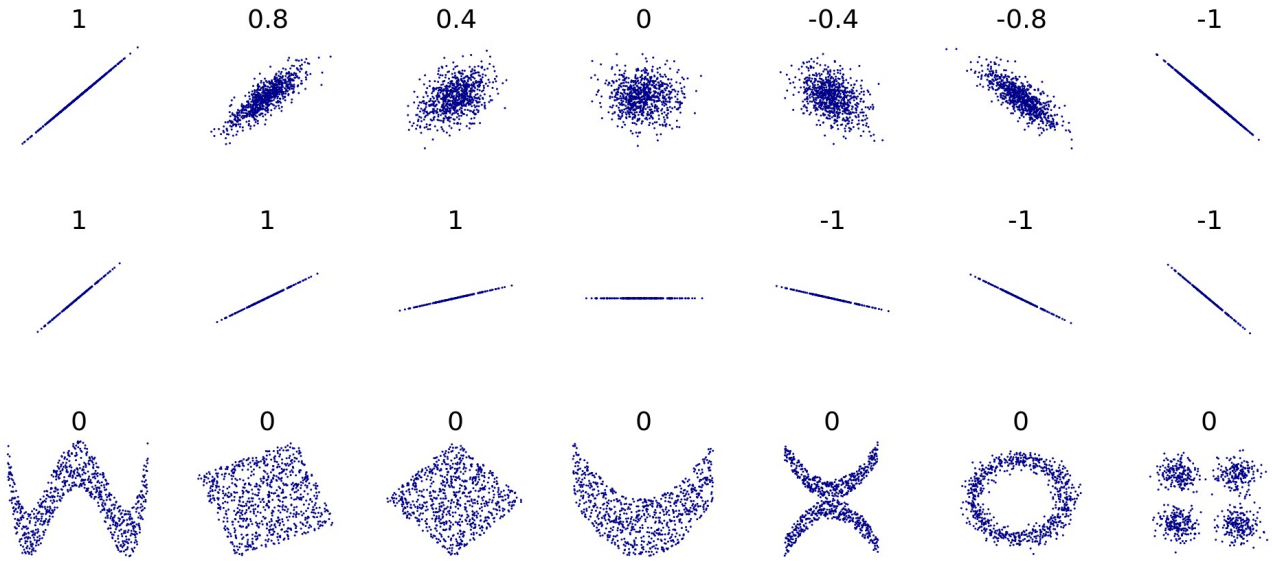
# Pearson Correlations on Complex Networks

Michele Coscia

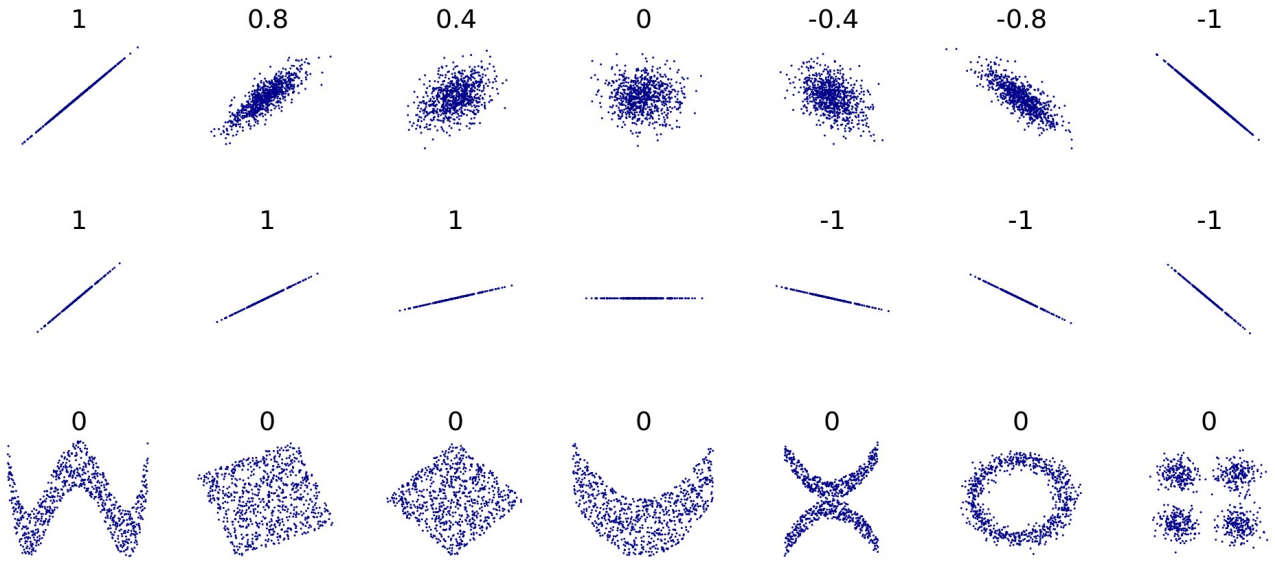
NERDS Meeting
Jan 19th, 2022

IT UNIVERSITY OF CPH

# Background

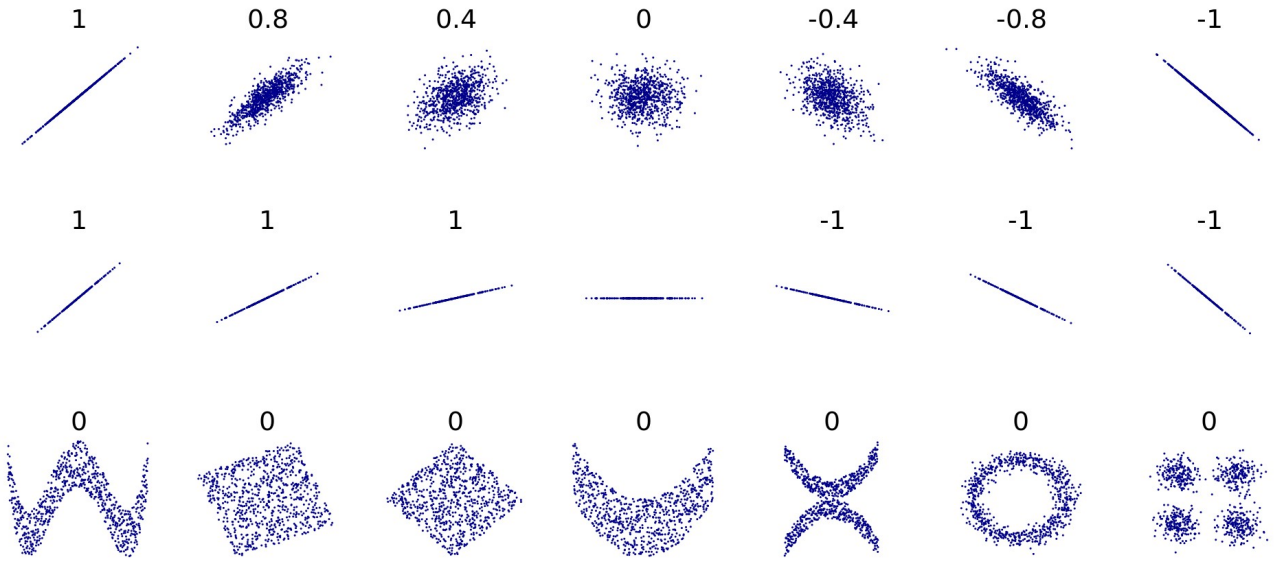


# Background



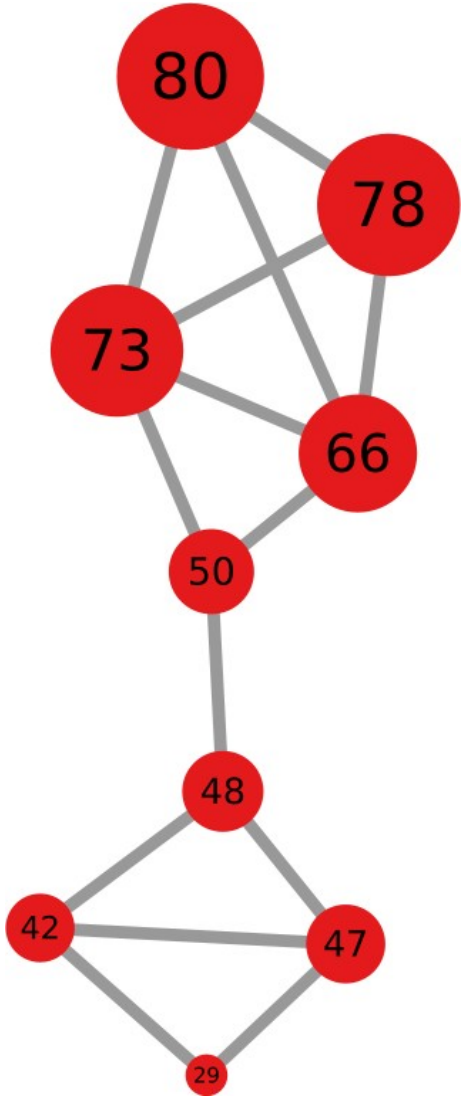

- Correlation = normalized covariance

# Background



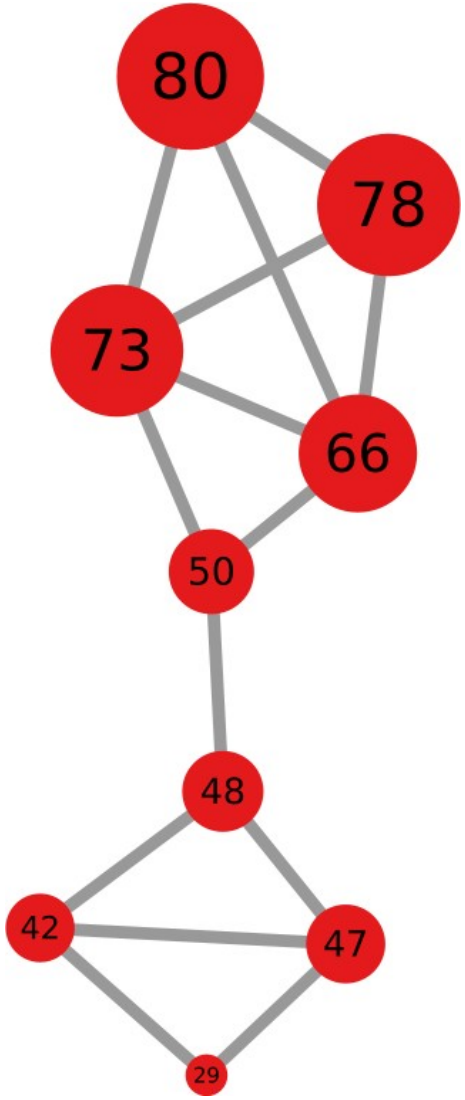

- Correlation = normalized covariance
- How well a linear equation describes the relationship between X and Y

# Motivation



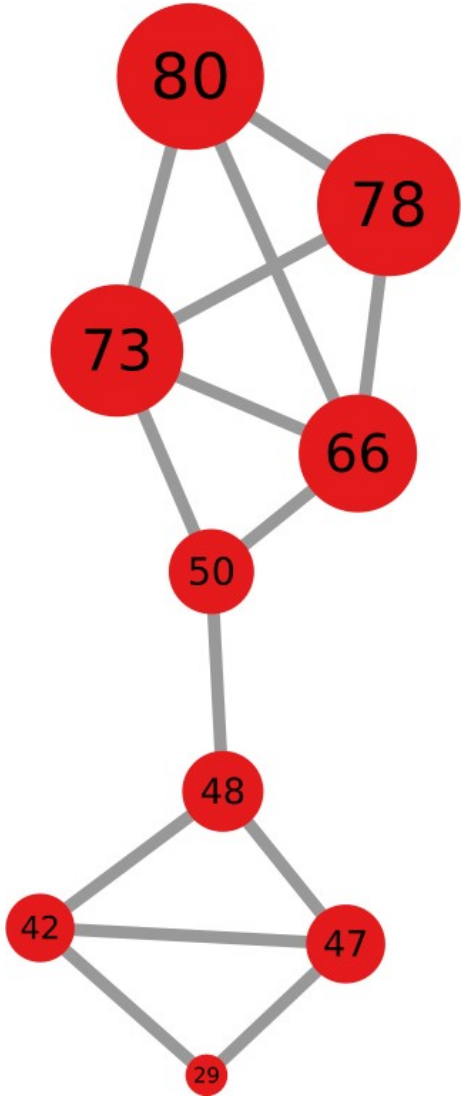
# Motivation

- Represent node properties as numeric vectors



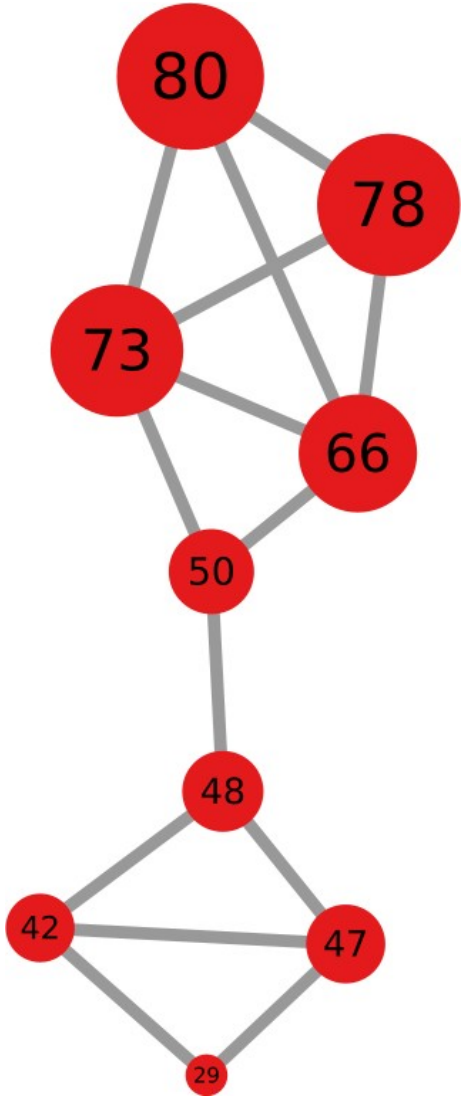
# Motivation

- Represent node properties as numeric vectors
- Many ways in which a network structure influences their values



# Motivation

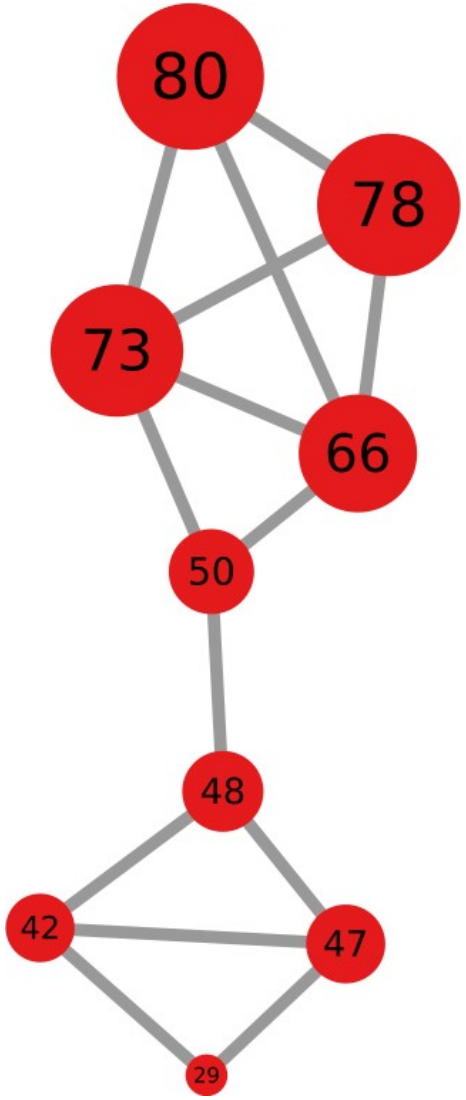
- Represent node properties as numeric vectors
- Many ways in which a network structure influences their values
- E.g.: SI/SIS/SIR, homphily/heterophily, etc



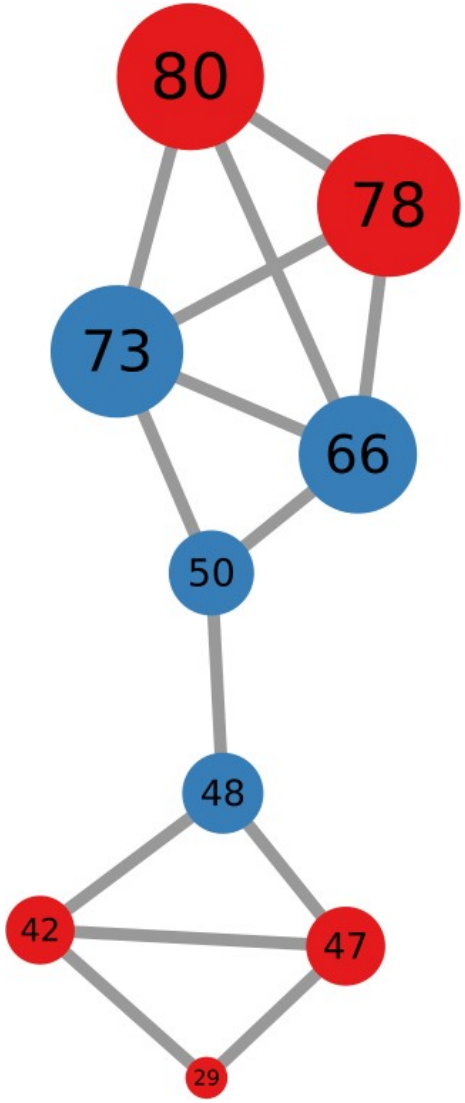
# Motivation

- Represent node properties as numeric vectors
- Many ways in which a network structure influences their values
- E.g.: SI/SIS/SIR, homphily/heterophily, etc

**~ 0.74**

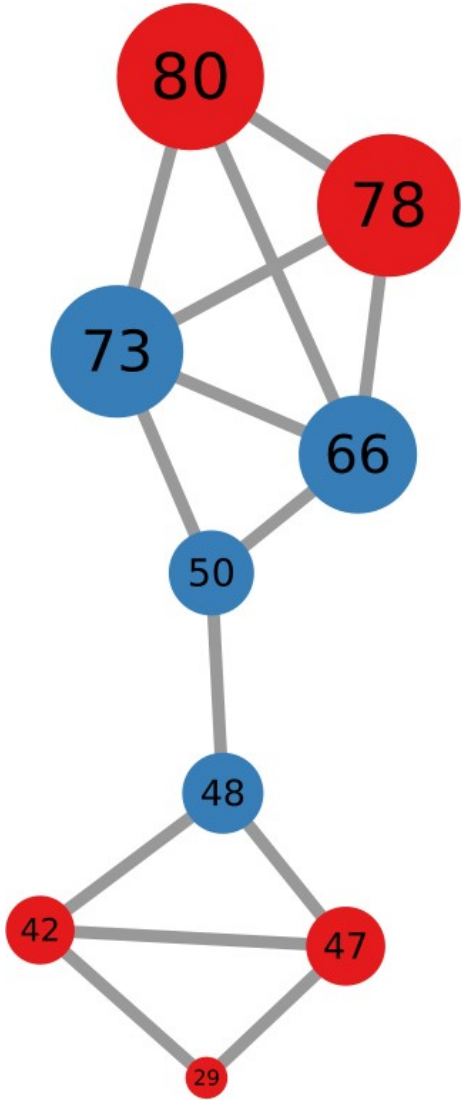


# Motivation



# Motivation

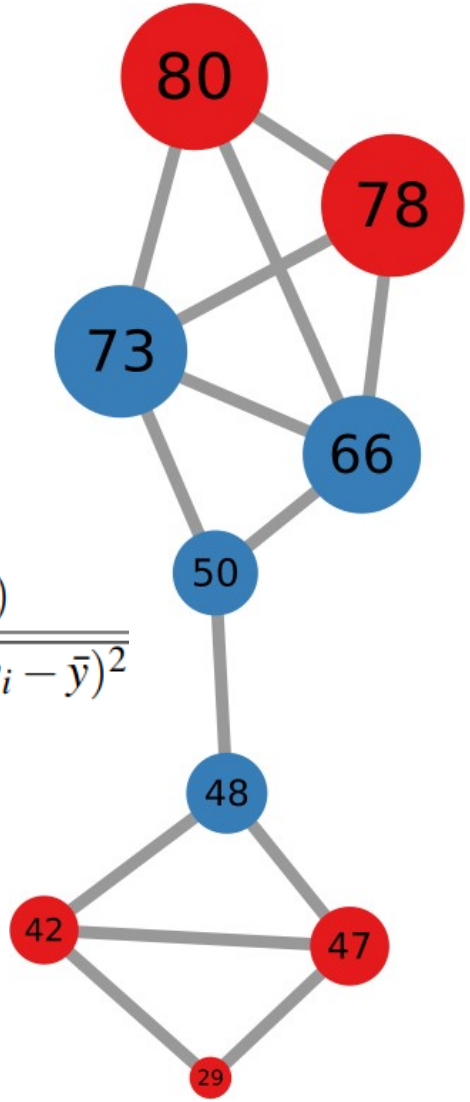
- What if we have more than one variable?



# Motivation

- What if we have more than one variable?
- Pearson doesn't see the network

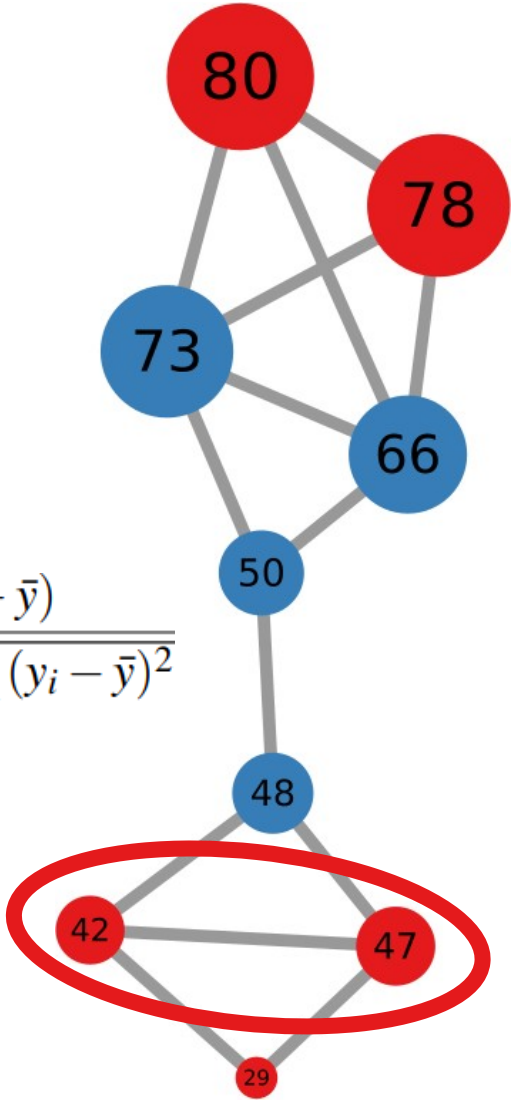
$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



# Motivation

- What if we have more than one variable?
- Pearson doesn't see the network
- Assortativity ignores higher-order relationships

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



# Research Question

How can we define a Pearson Correlation Coefficient that takes into account the network structure?

# Pearson

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

# Pearson

**Covariance of X and Y**

$$\rho_{x,y} = \frac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

# Pearson

Covariance of X and Y

$$\rho_{x,y} = \frac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}{\boxed{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

StDev of X

# Pearson

Covariance of X and Y

$$\rho_{x,y} = \frac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}{\boxed{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\boxed{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}}$$

StDev of X

StDev of Y

# Pearson

Covariance of X and Y

$$\rho_{x,y} = \frac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}{\boxed{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\boxed{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}}$$

StDev of X StDev of Y

- Max when Cov(x,y) = SD(x) * SD(y)

# Pearson

Covariance of X and Y

$$\rho_{x,y} = \frac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}{\boxed{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\boxed{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}}$$

StDev of X StDev of Y

- Max when Cov(x,y) = SD(x) * SD(y)
- Each change in x is proportional to a change in y

# Pearson

Covariance of X and Y

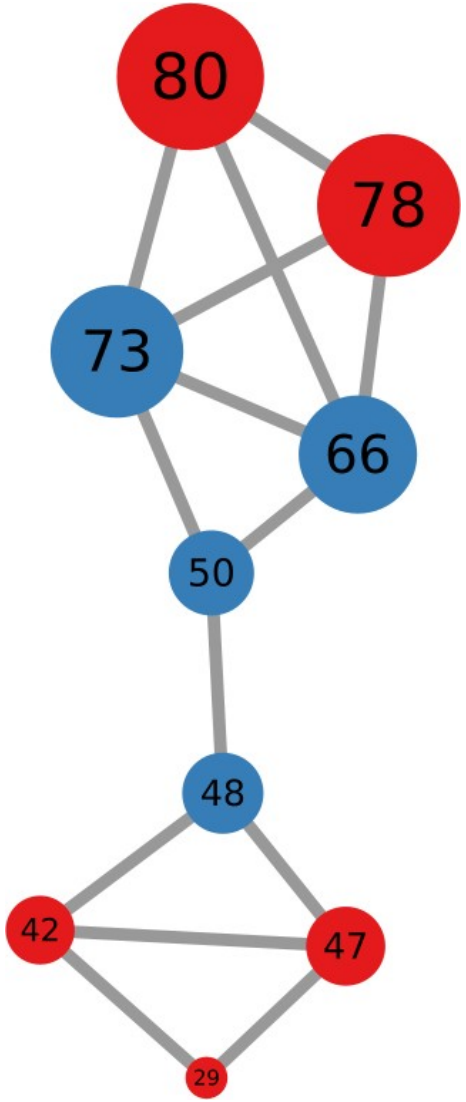
$$\rho_{x,y} = \frac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}{\boxed{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\boxed{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}}$$

StDev of X StDev of Y

- Max when Cov(x,y) = SD(x) * SD(y)
- Each change in x is proportional to a change in y
- Example: Cov(x,x) = SD(x) * SD(x) = Var(x)

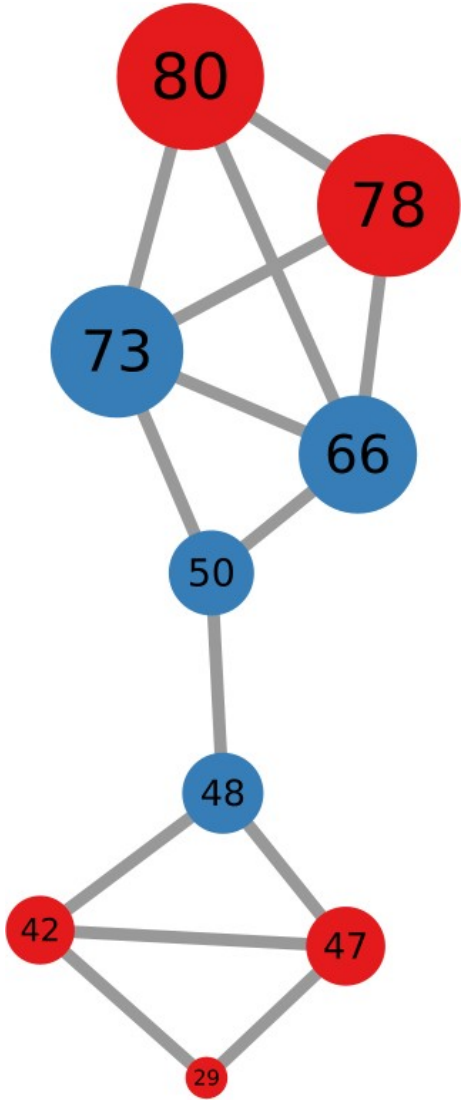
# Pearson & Networks

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



# Pearson & Networks

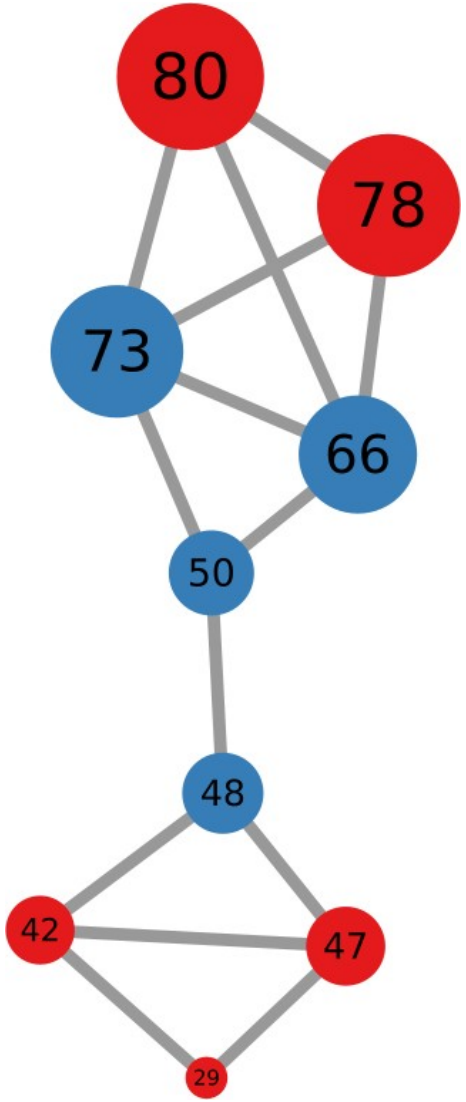
$$\rho_{x,y} = \frac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



# Pearson & Networks

$$\rho_{x,y} = \frac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

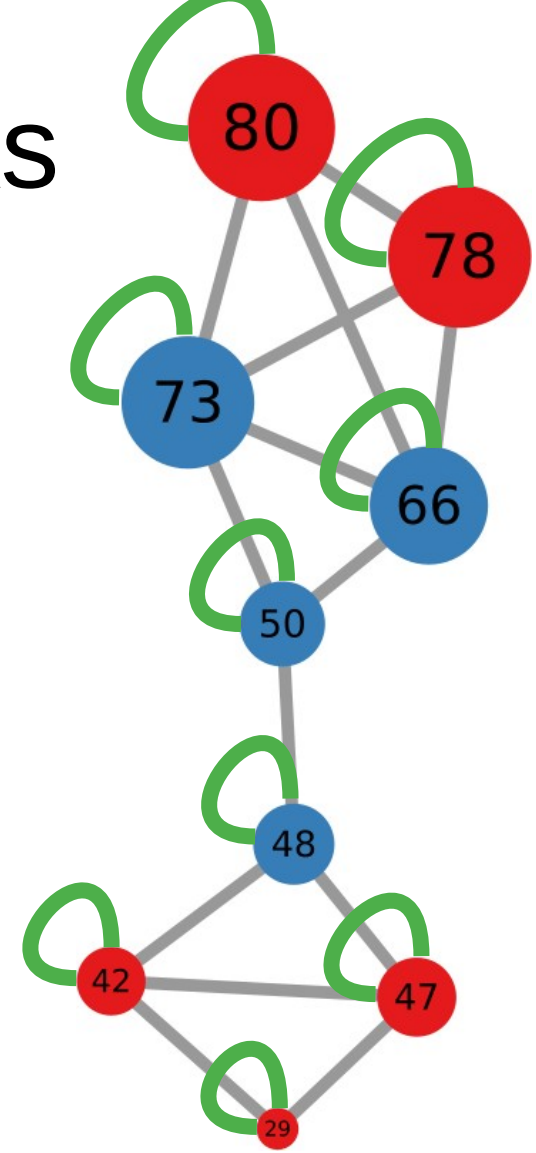
| | |
|---|---|
| 23 | 0.4 |
| 21 | 0.4 |
| 16 | -0.5 |
| 9 | -0.5 |
| -7 | -0.5 |
| -9 | -0.5 |
| -10 | 0.4 |
| -15 | 0.4 |
| -28 | 0.4 |



# Pearson & Networks

$$\rho_{x,y} = \frac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

| | |
|---|---|
| 23 | 0.4 |
| 21 | 0.4 |
| 16 | -0.5 |
| 9 | -0.5 |
| -7 | -0.5 |
| -9 | -0.5 |
| -10 | 0.4 |
| -15 | 0.4 |
| -28 | 0.4 |

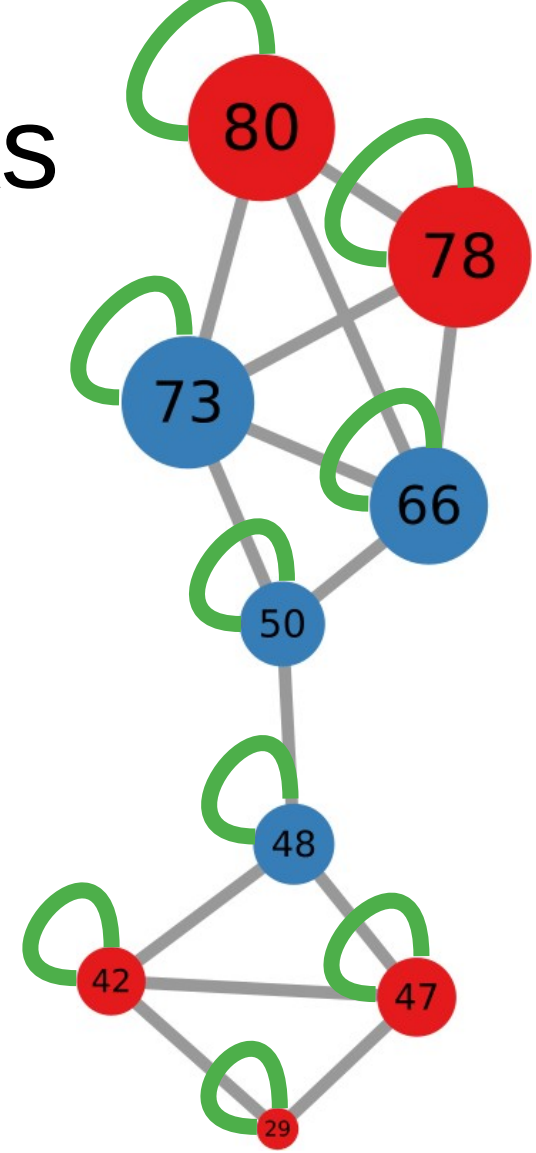


# Pearson & Networks

$$\rho_{x,y} = \frac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

| | |
|---|---|
| 23 | 0.4 |
| 21 | 0.4 |
| 16 | -0.5 |
| 9 | -0.5 |
| -7 | -0.5 |
| -9 | -0.5 |
| -10 | 0.4 |
| -15 | 0.4 |
| -28 | 0.4 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |



# Pearson & Networks

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

# Pearson & Networks

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

**sum(10, 9, -9, -5, 4, 5, -4, -7, -12)**

# Pearson & Networks

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

**sum(10, 9, -9, -5, 4, 5, -4, -7, -12)**

$$\rho_{x,y} = \frac{\mathrm{sum}(I \times (\hat{x} \otimes \hat{y}))}{\sigma_x \sigma_y}$$

# Pearson & Networks

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

**sum(10, 9, -9, -5, 4, 5, -4, -7, -12)**

$$\rho_{x,y} = \frac{\mathrm{sum}(I \times (\hat{x} \otimes \hat{y}))}{\sigma_x \sigma_y}$$

| | 0.4 | 0.4 | -0.5 | -0.5 | -0.5 | -0.5 | 0.4 | 0.4 | 0.4 |
|---|---|---|---|---|---|---|---|---|---|
| **23** | 10 | 10 | -13 | -13 | -13 | -13 | 10 | 10 | 10 |
| **21** | 9 | 9 | -12 | -12 | -12 | -12 | 9 | 9 | 9 |
| **16** | 7 | 7 | -9 | -9 | -9 | -9 | 7 | 7 | 7 |
| **9** | 4 | 4 | -5 | -5 | -5 | -5 | 4 | 4 | 4 |
| **-7** | 3 | -3 | 4 | 4 | 4 | 4 | -3 | -3 | -3 |
| **-9** | 4 | -4 | 5 | 5 | 5 | 5 | -4 | -4 | -4 |
| **-10** | 4 | -4 | 6 | 6 | 6 | 6 | -4 | -4 | -4 |
| **-15** | 7 | -7 | 8 | 8 | 8 | 8 | -7 | -7 | -7 |
| **-28** | -12 | -12 | 16 | 16 | 16 | 16 | -12 | -12 | -12 |

# Pearson & Networks

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

**sum(10, 9, -9, -5, 4, 5, -4, -7, -12)**

Outer

$$\rho_{x,y} = \frac{\mathrm{sum}(I \times (\hat{x} \otimes \hat{y}))}{\sigma_x \sigma_y}$$

| | 0.4 | 0.4 | -0.5 | -0.5 | -0.5 | -0.5 | 0.4 | 0.4 | 0.4 |
|---|---|---|---|---|---|---|---|---|---|
| **23** | 10 | 10 | -13 | -13 | -13 | -13 | 10 | 10 | 10 |
| **21** | 9 | 9 | -12 | -12 | -12 | -12 | 9 | 9 | 9 |
| **16** | 7 | 7 | -9 | -9 | -9 | -9 | 7 | 7 | 7 |
| **9** | 4 | 4 | -5 | -5 | -5 | -5 | 4 | 4 | 4 |
| **-7** | 3 | -3 | 4 | 4 | 4 | 4 | -3 | -3 | -3 |
| **-9** | 4 | -4 | 5 | 5 | 5 | 5 | -4 | -4 | -4 |
| **-10** | 4 | -4 | 6 | 6 | 6 | 6 | -4 | -4 | -4 |
| **-15** | 7 | -7 | 8 | 8 | 8 | 8 | -7 | -7 | -7 |
| **-28** | -12 | -12 | 16 | 16 | 16 | 16 | -12 | -12 | -12 |

# Pearson & Networks

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

**sum(10, 9, -9, -5, 4, 5, -4, -7, -12)**

Outer

$$\rho_{x,y} = \frac{\mathrm{sum}(I \times (\hat{x} \otimes \hat{y}))}{\sigma_x \sigma_y}$$

| | 0.4 | 0.4 | -0.5 | -0.5 | -0.5 | -0.5 | 0.4 | 0.4 | 0.4 |
|---|---|---|---|---|---|---|---|---|---|
| **23** | 10 | | | | | | | | |
| **21** | | 9 | | | | | | | |
| **16** | | | -9 | | | | | | |
| **9** | | | | -5 | | | | | |
| **-7** | | | | | 4 | | | | |
| **-9** | | | | | | 5 | | | |
| **-10** | | | | | | | -4 | | |
| **-15** | | | | | | | | -7 | |
| **-28** | | | | | | | | | -12 |

# Pearson & Networks

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

**sum(10, 9, -9, -5, 4, 5, -4, -7, -12)**

**Elementwise** **Outer**

$$\rho_{x,y} = \frac{\mathrm{sum}(I \times (\hat{x} \otimes \hat{y}))}{\sigma_x \sigma_y}$$

| | 0.4 | 0.4 | -0.5 | -0.5 | -0.5 | -0.5 | 0.4 | 0.4 | 0.4 |
|---|---|---|---|---|---|---|---|---|---|
| **23** | 10 | | | | | | | | |
| **21** | | 9 | | | | | | | |
| **16** | | | -9 | | | | | | |
| **9** | | | | -5 | | | | | |
| **-7** | | | | | 4 | | | | |
| **-9** | | | | | | 5 | | | |
| **-10** | | | | | | | -4 | | |
| **-15** | | | | | | | | -7 | |
| **-28** | | | | | | | | | -12 |

# Pearson & Networks

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

**sum(10, 9, -9, -5, 4, 5, -4, -7, -12)**

**Elementwise** **Outer**

$$\rho_{x,y} = \frac{\mathrm{sum}(I \times (\hat{x} \otimes \hat{y}))}{\sigma_x \sigma_y}$$

$$\sigma_x = \sqrt{\mathrm{sum}(I \times (\hat{x} \otimes \hat{x}))}$$

| | 0.4 | 0.4 | -0.5 | -0.5 | -0.5 | -0.5 | 0.4 | 0.4 | 0.4 |
|---|---|---|---|---|---|---|---|---|---|
| **23** | 10 | | | | | | | | |
| **21** | | 9 | | | | | | | |
| **16** | | | -9 | | | | | | |
| **9** | | | | -5 | | | | | |
| **-7** | | | | | 4 | | | | |
| **-9** | | | | | | 5 | | | |
| **-10** | | | | | | | -4 | | |
| **-15** | | | | | | | | -7 | |
| **-28** | | | | | | | | | -12 |

# Pearson & Networks

# Pearson & Networks

- We don’t have to use I

# Pearson & Networks

- We don’t have to use I
- We can use W, a matrix of node **proximities**

# Pearson & Networks

- We don’t have to use I
- We can use W, a matrix of node **proximities**
- My choice: $W = e^{-kL}$

# Pearson & Networks

- We don’t have to use I
- We can use W, a matrix of node **proximities**
- My choice: $W = e^{-kL}$
- L = shortest path length

# Pearson & Networks

- We don’t have to use I
- We can use W, a matrix of node **proximities**
- My choice: $W = e^{-kL}$
- L = shortest path length
- k = damping factor

# Pearson & Networks

- We don’t have to use I
- We can use W, a matrix of node **proximities**
- My choice: $W = e^{-kL}$
- L = shortest path length
- k = damping factor
- Self loop = 1, exponential decay with distance

# Pearson & Networks

$$\rho_{x,y} = \frac{\mathrm{sum}(I \times (\hat{x} \otimes \hat{y}))}{\sigma_x \sigma_y}$$

# Pearson & Networks

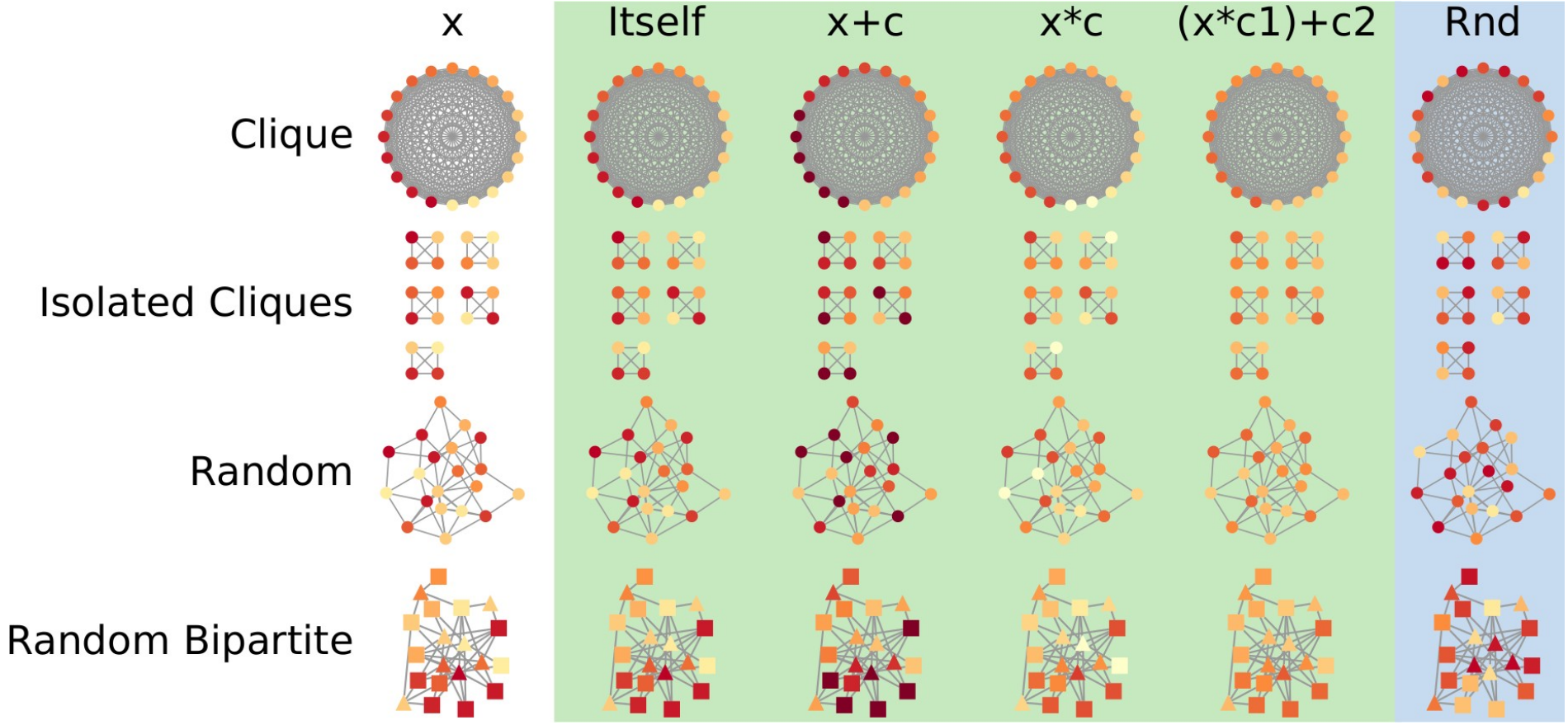
$$\rho_{x,y} = \frac{\mathrm{sum}(I \times (\hat{x} \otimes \hat{y}))}{\sigma_x \sigma_y} \longrightarrow \rho_{x,y,G} = \frac{\mathrm{sum}(W \times (\hat{x} \otimes \hat{y}))}{\sigma_{x,W}\, \sigma_{y,W}}$$

# Pearson & Networks

$$\rho_{x,y} = \frac{\mathrm{sum}(I \times (\hat{x} \otimes \hat{y}))}{\sigma_x \sigma_y} \longrightarrow \rho_{x,y,G} = \frac{\mathrm{sum}(W \times (\hat{x} \otimes \hat{y}))}{\sigma_{x,W} \sigma_{y,W}}$$

$$\sigma_x = \sqrt{\mathrm{sum}(I \times (\hat{x} \otimes \hat{x}))} \longrightarrow \sigma_{x,W} = \sqrt{\mathrm{sum}(W \times (\hat{x} \otimes \hat{x}))}$$

# Validation: Expected Behavior

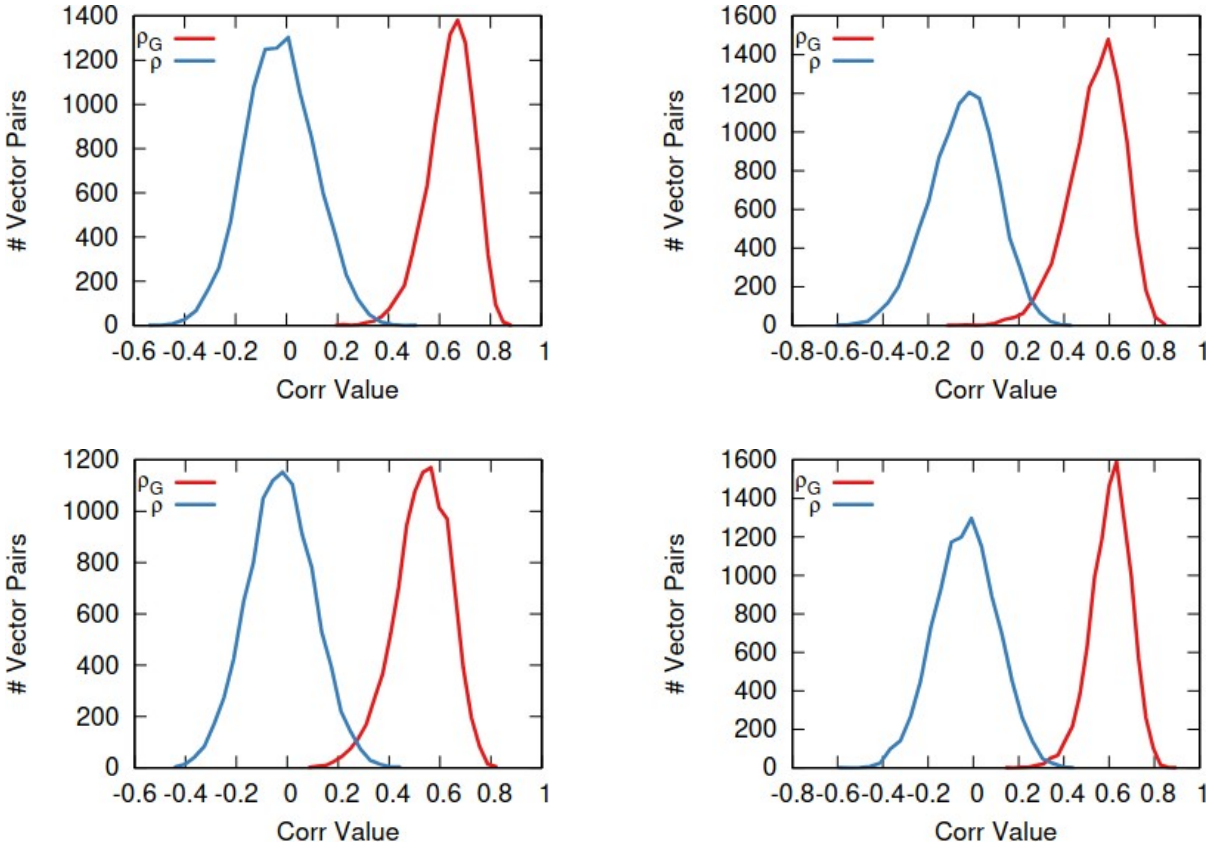


# Validation: Difference with Baseline

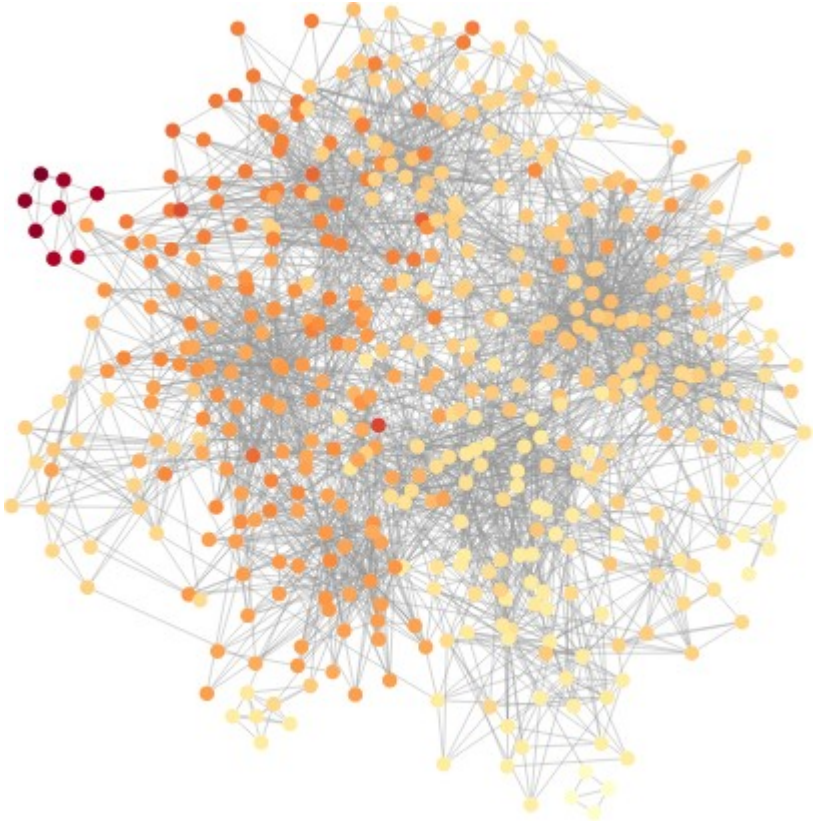
$$y_v = \sum_{u \in N_v} x_u / |N_v|$$

# Validation: Difference with Baseline

$$y_v = \sum_{u \in N_v} x_u / |N_v|$$

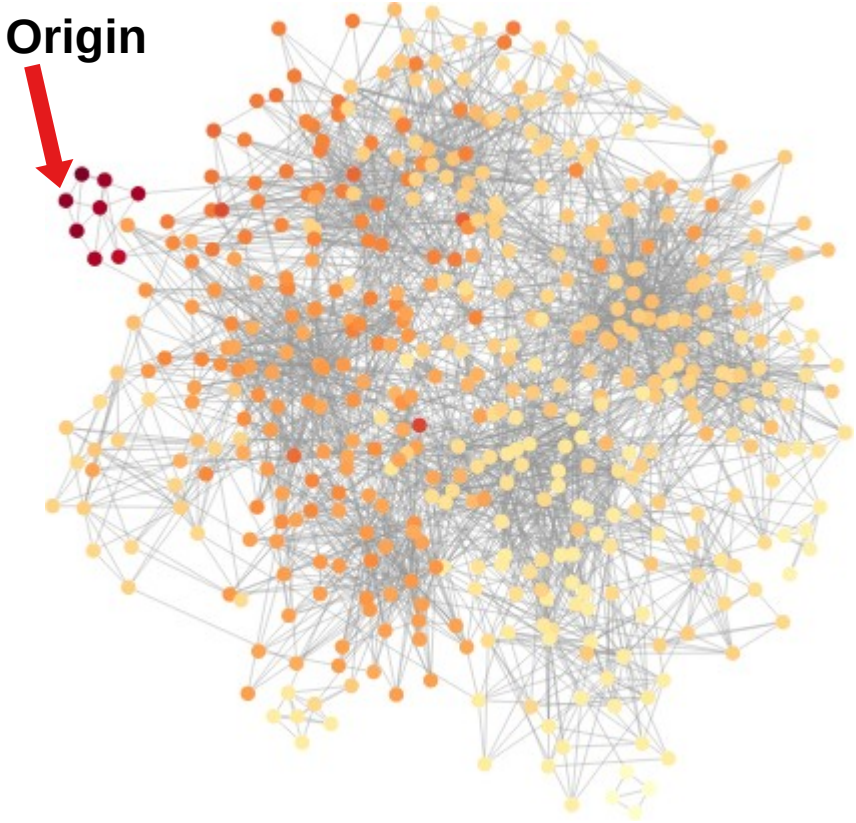


# Validation: Diffusion Processes



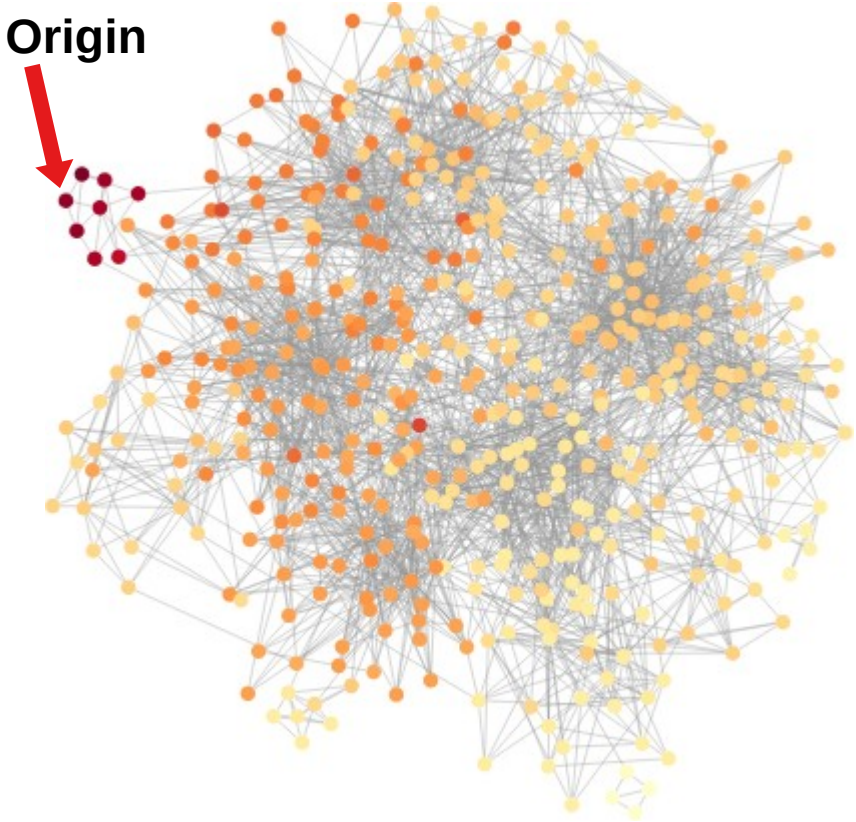
**x**

# Validation: Diffusion Processes

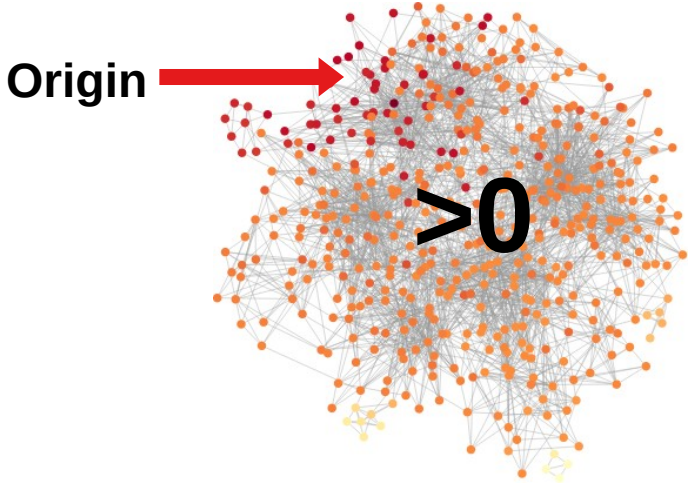



**X**

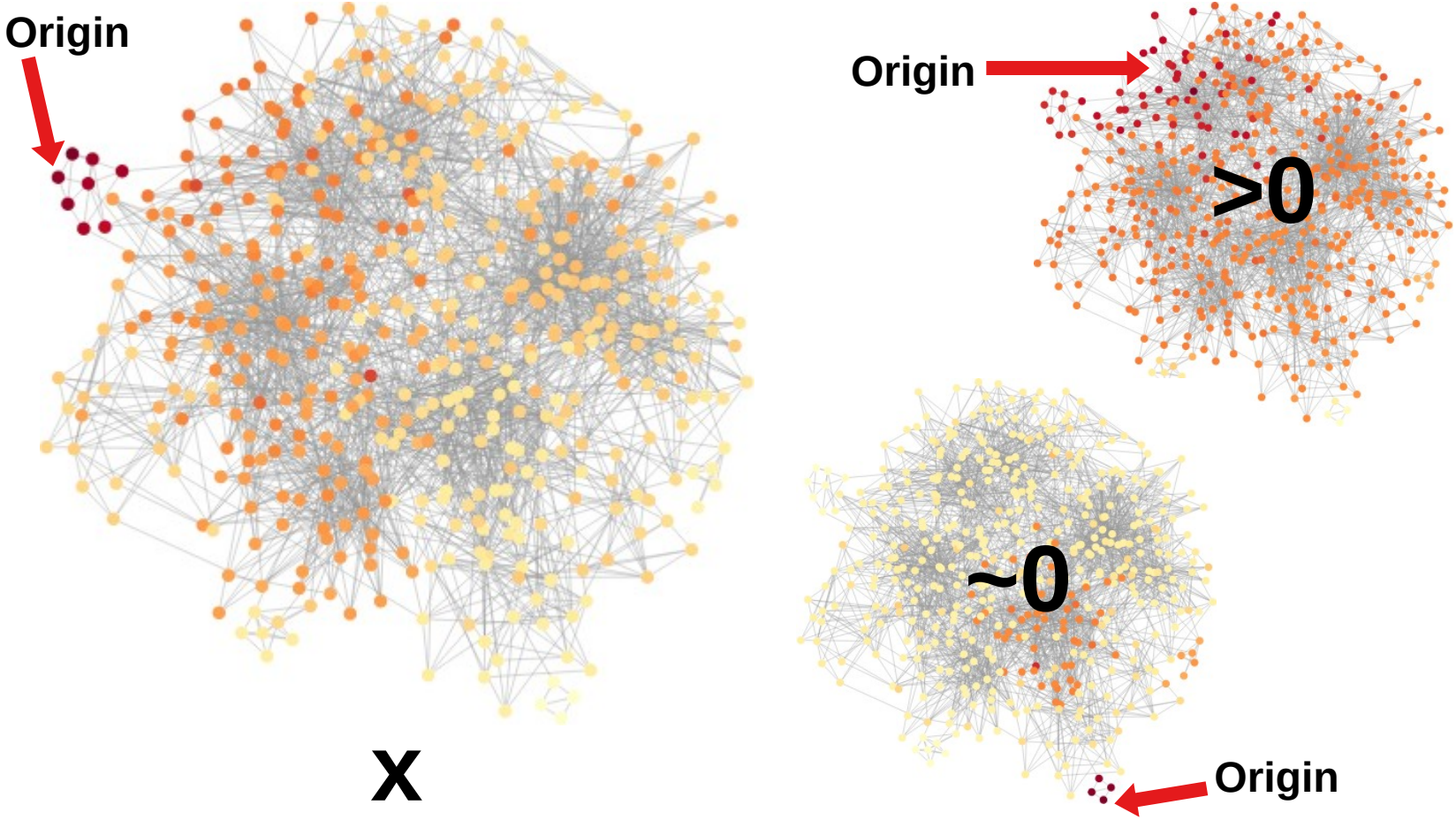
# Validation: Diffusion Processes



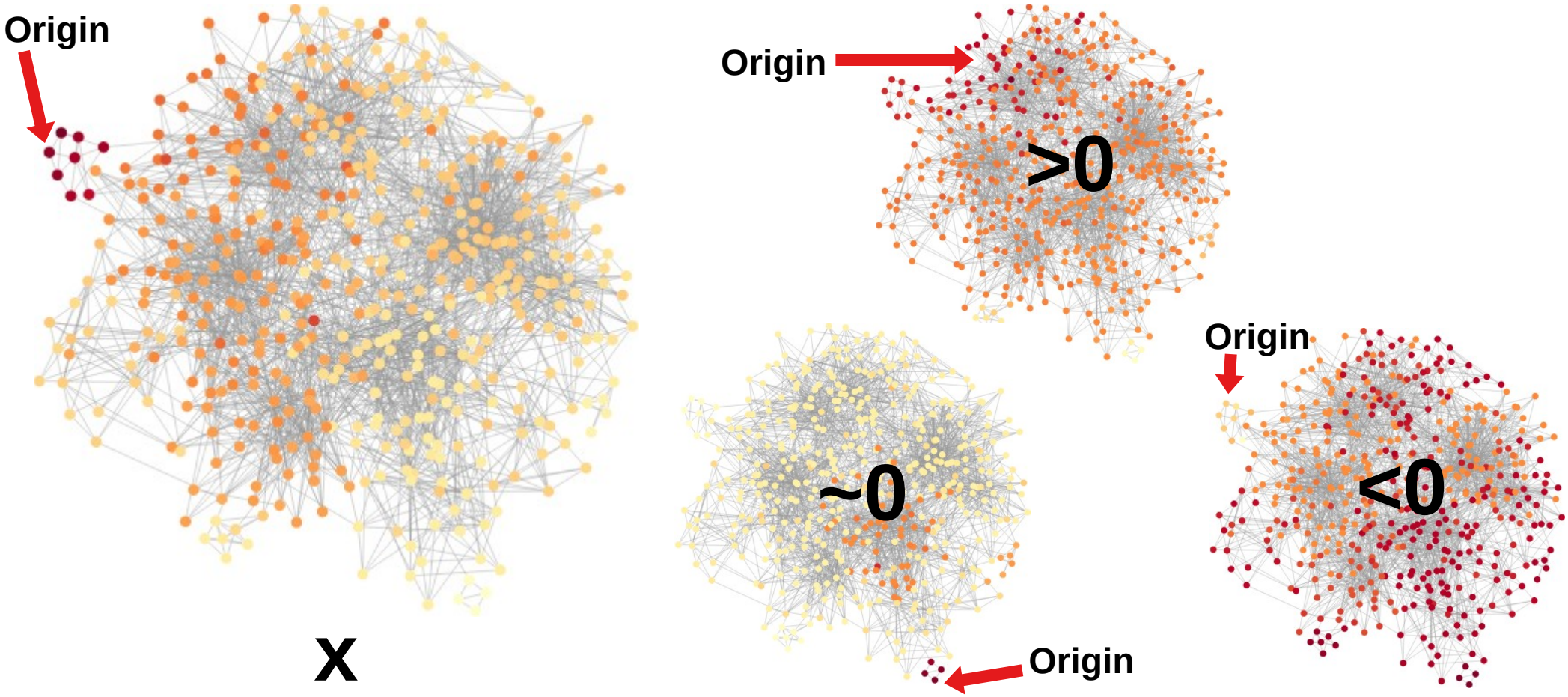

x



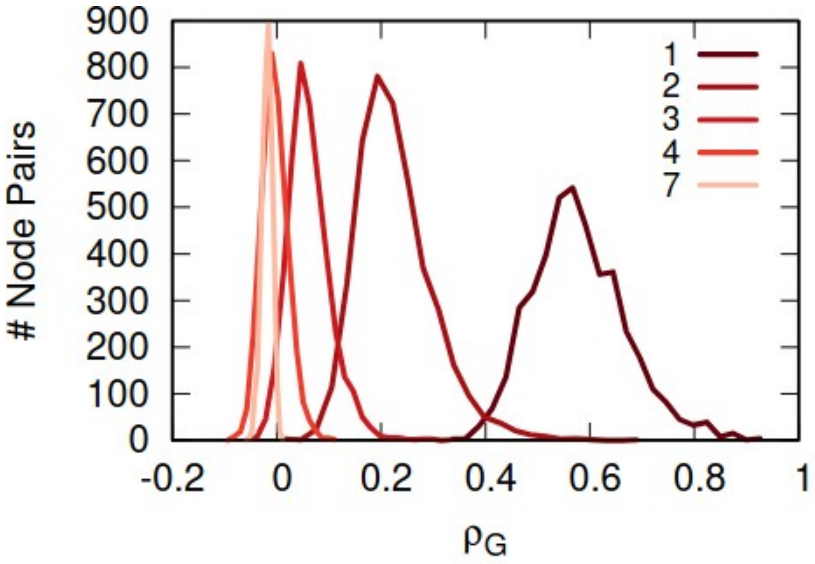
# Validation: Diffusion Processes



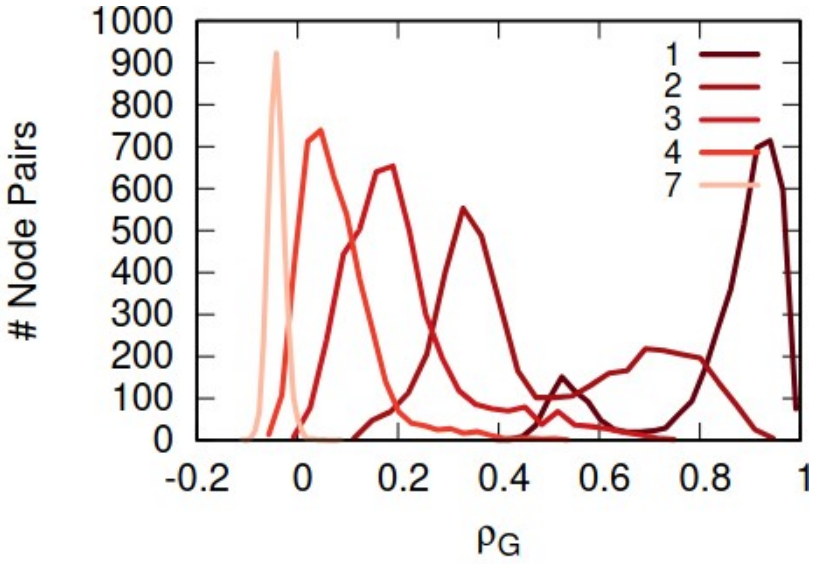
# Validation: Diffusion Processes



# Validation: SPL

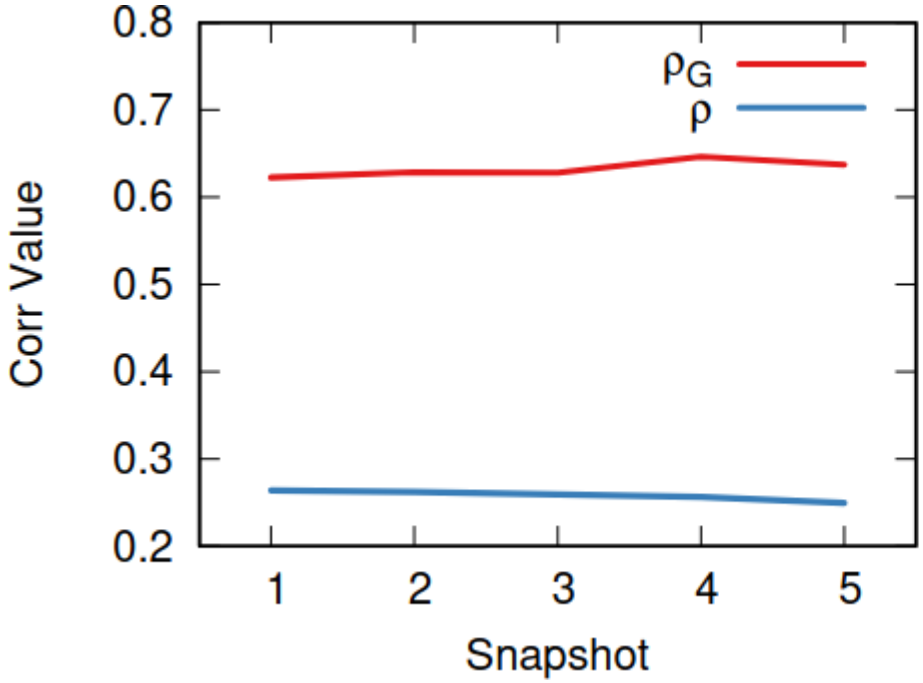



(a) $G_{n,m}$

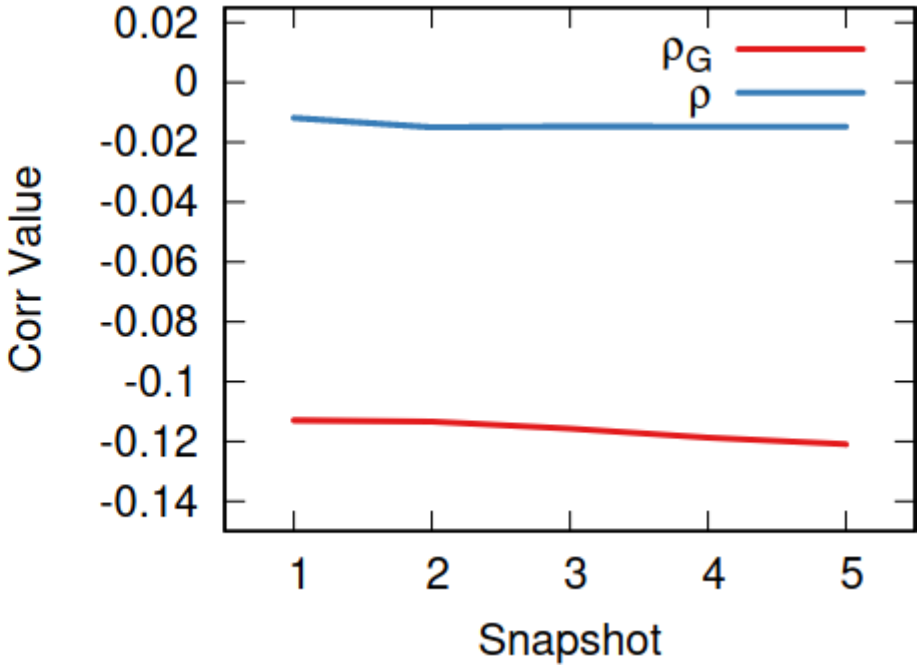



(b) Small World.

# Application




**Shelf size vs Wishlist size**




**Age vs # Tags used**

# Summary

- Pearson is useful but ignores networks
- We can plug a network in
- The resulting correlation is meaningful
- It has practical applications

# Thank you!

# Questions?

Michele Coscia

IT UNIVERSITY OF CPH

mcos@itu.dk