

Chapter 6

Centrality & Assortative Mixing

Summary

- Measuring Node importance
- Do Birds of a feather flock together?

Reading

- Chapters 3 & 4 of Kleinberg's book



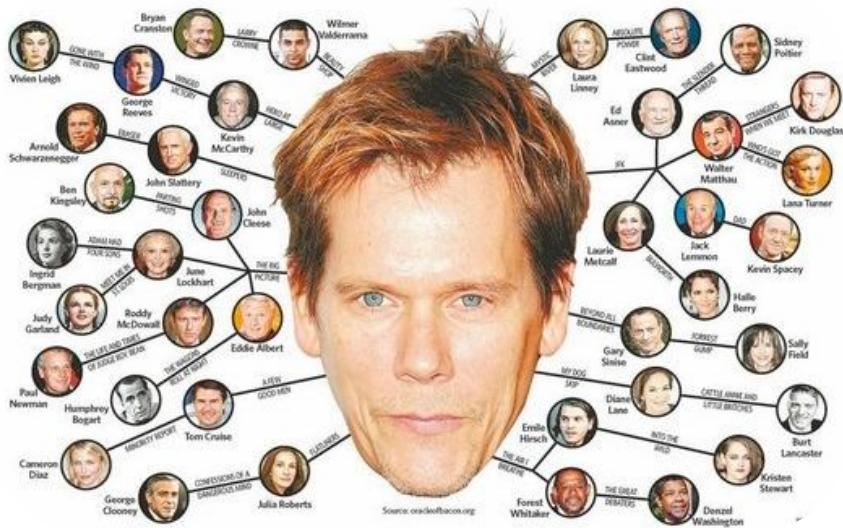
How important is a node in a network?

We can measure nodes importance using so-called **centrality**.

Bad term:
nothing to do with being central in general

Usage:

- Some centralities have straightforward interpretation
- Centralities can be used as node features for machine learning on graph



<https://oracleofbacon.org/>

Where are you?

It is always possible, once *fixed a context*, to measure our distance from a “focal” node.

For instance:

Movie Stars:

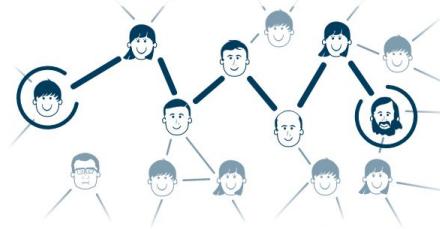
- Bacon number

Researchers:

- Erdos number

Are such “focal” nodes really different from the others?

Co-authorship distance computation



Paul Erdős
co-authored 2 papers with
Shlomo Moran
co-authored 7 papers with
Ronny Lempel
co-authored 4 papers with
Fabrizio Silvestri
co-authored 2 papers with
Giulio Rossetti
distance = 4

Name	Erdős number	Bacon number	Erdős–Bacon number
Daniel Kleitman	1	2	3 ^[4]
Bruce Reznick	1	2	3 ^[69]
Albert M. Chan	3 ^{[21][22][23]}	1 ^[24]	4
Nicholas Metropolis	2 ^[10]	2 ^[68]	4
Steven Strogatz	3 ^{[79][80][81]}	1 ^{[a][c][82]}	4 ^{[a][c]}
Robert J. Marks II	3 ^{[44][45][46]}	2 ^{[61][62][63]}	5
Tom Porter	3 (in two ways) ^{[6][7]}	2 ^{[a][c][8][9]}	5 ^{[a][c]}
Richard Thaler	3 ^{[86][87][88]}	2 ^{[89][a][90]}	5
Doron Zeilberger	2 ^{[93][35]}	3 ^{[a][94][95][96]}	5 ^{[a][c]}
Misha Collins	4 ^{[25][26][27][28]}	2 ^{[29][30]}	6
William A. Dembski	4 ^{[44][45][46][47]}	2 ^{[a][48]}	6 ^[a]
Richard Feynman	3	3	6 ^[10]
Ken Goldberg	3 ^{[49][50][51]}	3 ^{[52][53][54]}	6
Stephen Hawking	4	2 ^[a]	6 ^[20]

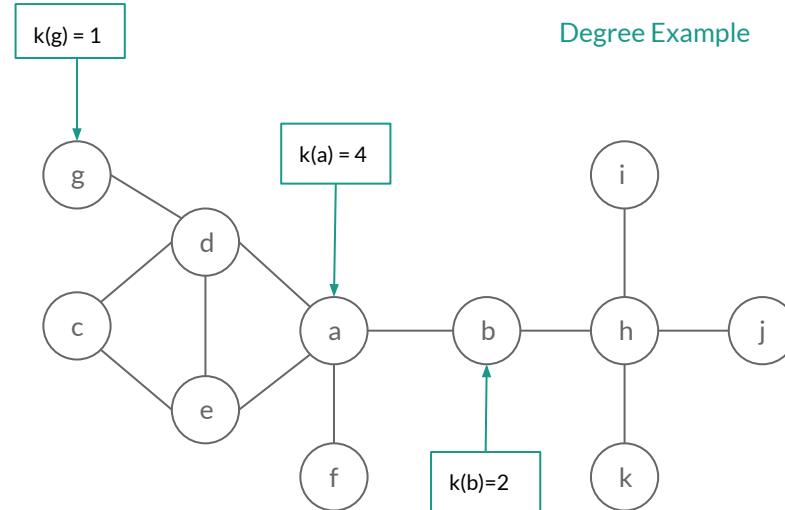
Degree Centrality

How many neighbors does a node have?

Often enough to find important nodes
(e.g., main characters of a tv series talk with more people)

But not always

- Twitter users with the most contacts are spam
- Webpages/wikipedia pages with most links are often lists of references



k = number of links

$$A_{i,j} \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected,} \\ 0 & \text{otherwise} \end{cases}$$

$$k_i = \sum_{j=1}^n A_{ij}$$

Connectivity-based centralities

"influence based on the number of links a node has to other nodes in the network"



Recursive definitions

Recursive importance:

Important nodes are those connected to important nodes

Several centralities based on this idea:

- Eigenvector centrality
- PageRank
- Katz centrality
- ...

Idea:

1. Each node has a score (centrality)
2. If every node “sends” its score to its neighbors, the sum of all scores received by each node will be equal to its original score

$$x_i^{(t+1)} = \sum_{j=1}^n A_{ij} x_j^{(t)}$$

x_i is the centrality of node i

How to solve it (*power method*):

1. Initialize scores to random values
2. Update the values according to the desired rule

Does it converge?

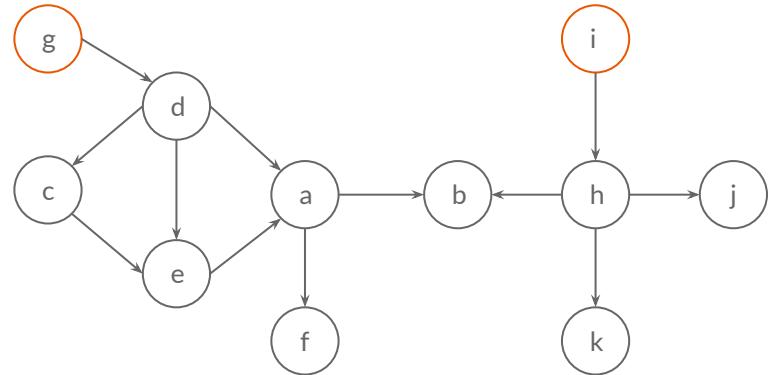
Yes, if the graph is *undirected* with a *single connected component* (Perron-Frobenius theorem)

Eigenvector Centrality

A pair of eigenvector (x) and eigenvalue (λ) is defined by the relation:

$$Ax = \lambda x$$

- x is a vector of size N that can be interpreted as the **nodes scores**
- Ax yield a new vector of the same size which corresponds for each node to **the sum of the received scores from its neighbors**
- the equality implies that the new scores are proportional to the previous ones



Problems:

In case of DiGraphs:

- Adjacency matrix is asymmetric
- 2 sets of eigenvectors
- 2 leading eigenvectors (use the incoming ones)

In presences of source nodes (0 in-degree):

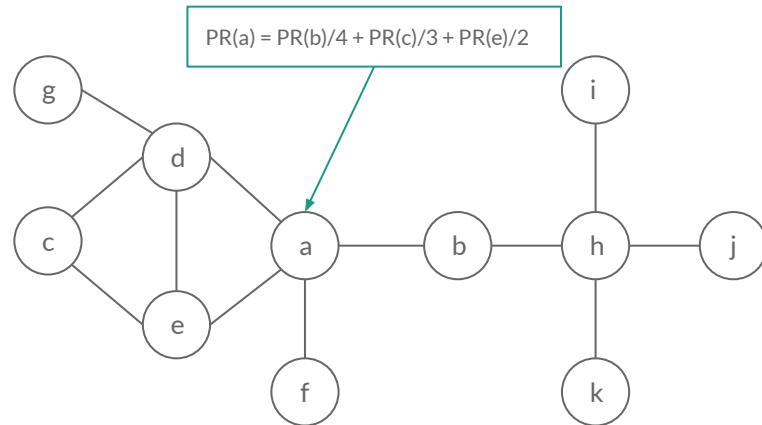
- $E(g) = 0$
- $E(d) = 0$ as well since its incoming link comes from A
- ...

PageRank

Main idea: The PageRank computation can be interpreted as a Random Walk process with restart

Probability that the RW will be in node i next step depends only on the current node j and the transition probability
 $j \rightarrow i$ determined by the stochastic matrix

- Consequently this is a first-order Markov process
- Stationary probabilities (i.e., when walk length tends towards ∞) of the RW to be in node i gives the PageRank of the node



$$PR(x) = \frac{1 - \alpha}{N} + \alpha \left(\sum_{k=1}^n \frac{PR(k)}{C(k)} \right)$$

Teleportation probability: the parameter α gives the probability that in the next step of the RW will follow a Markov process or with probability $1-\alpha$ it will jump to a random node

- $\alpha < 1$, it assures that the RW will never be stuck at nodes with $k_{out} = 0$, but it can restart the RW from a randomly selected other node (usually $\alpha=0.85$)

PageRank (cont'd)

PageRank can also be interpreted as the **dominant eigenvector** of the **normalized adjacency matrix**

Two improvements w.r.t. eigenvector centrality:

- Add a constant centrality gain for every node (**solves source node problem** in digraphs)
- Nodes with very high centralities give high centralities to their neighbors

$$\mathbf{R} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

where R is the solution of the equation

$$\mathbf{R} = \begin{bmatrix} (1 - \alpha)/N \\ (1 - \alpha)/N \\ \vdots \\ (1 - \alpha)/N \end{bmatrix} + \alpha \begin{bmatrix} \ell(p_1, p_1) & \ell(p_1, p_2) & \cdots & \ell(p_1, p_N) \\ \ell(p_2, p_1) & \ddots & & \vdots \\ \vdots & & \ell(p_i, p_j) & \\ \ell(p_N, p_1) & \cdots & & \ell(p_N, p_N) \end{bmatrix} \mathbf{R}$$

$$\sum_{i=1}^N \ell(p_i, p_j) = 1$$

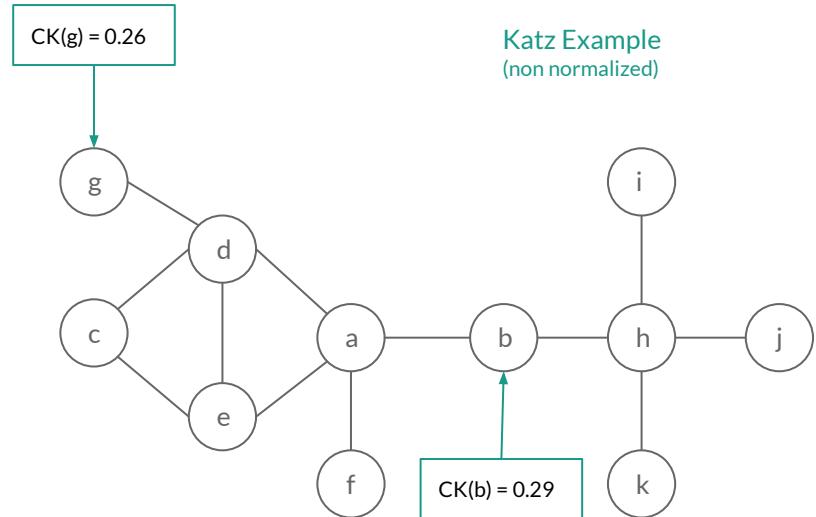
$\ell(p_i, p_j)$ is the ratio between number of links outbound from page j to page i to the total number of outbound links of page j

Katz Centrality

Measuring the relative **degree of influence** of a node within a network

$CK(i)$:

for all distances k , for all nodes j , sum the number of different paths $i \rightarrow j$ of length k (multiplied by an attenuation factor α)



$$CK(i) = \sum_{k=1}^{\infty} \sum_{j=1}^n \alpha^k (A^k)_{ij}$$

NB: Attenuation factor α must be smaller than $1/|\lambda|$, i.e. the reciprocal of the absolute value of the largest eigenvalue of A .

Geometric Centralities

"importance of a node depends on some function of its distances w.r.t. other nodes"



Closeness Centrality

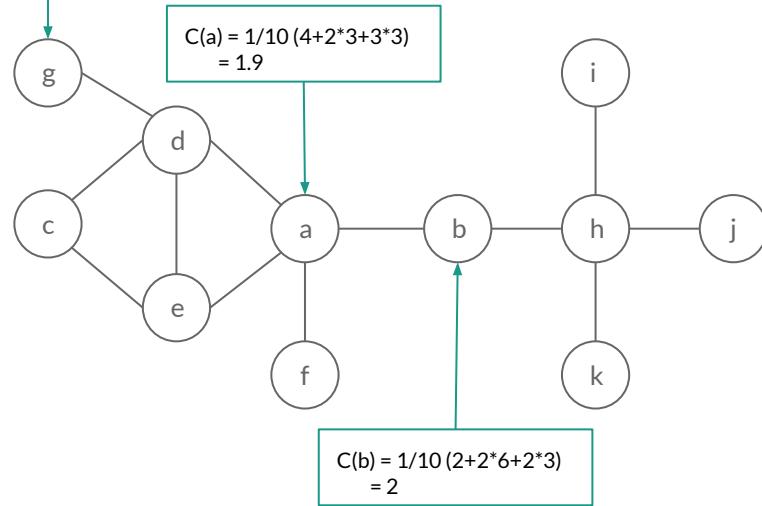
Farness: average of length of shortest paths to all other nodes

Closeness: inverse of the Farness
(normalized by number of nodes)

- Highest closeness = More central
- Closeness=1: directly connected to all other nodes
- Well defined only on connected networks

Farness Example

$$C(g) = 1/10 (1+2*3+2*3+4+3*5) \\ = 3.2$$

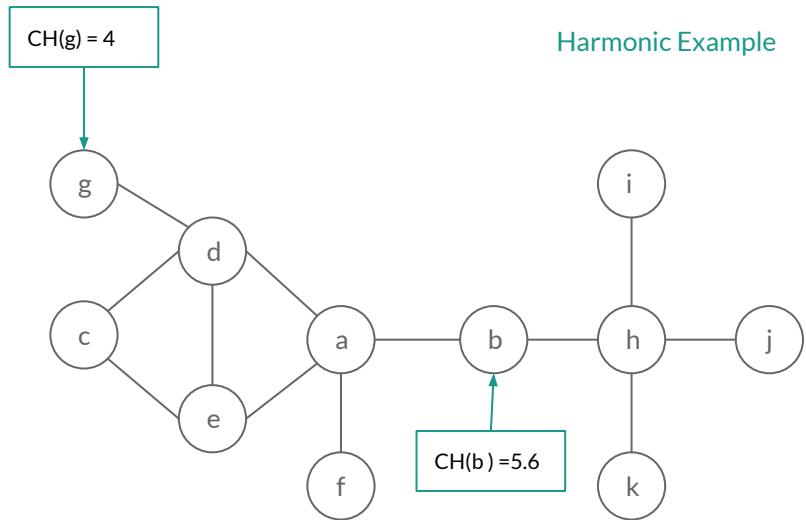


Closeness Formula

$$C_{cl}(i) = \frac{n - 1}{\sum_{d_{ij} < \infty} d_{ij}}$$

Harmonic Centrality

Harmonic mean of the geodesic
(shorted paths) distances from a given node
to all others.



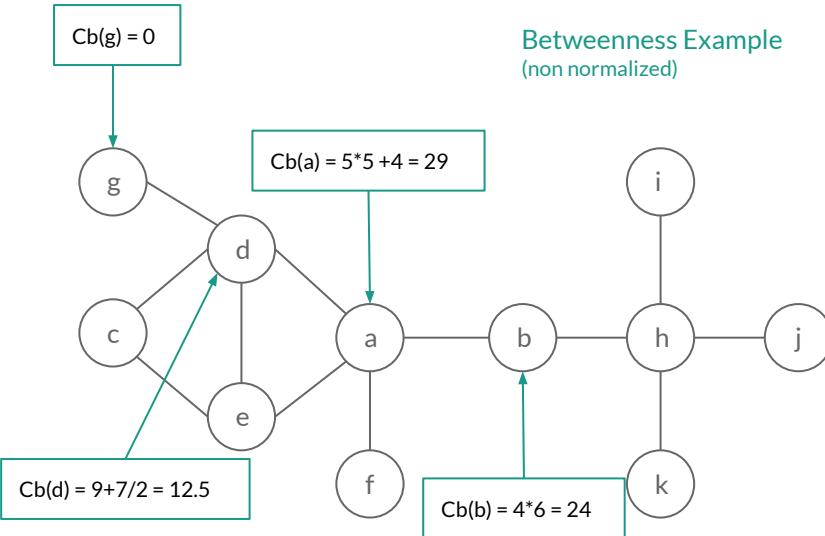
$$CH(i) = \frac{1}{n-1} \sum_{i \neq j} \frac{1}{d_{ij}}$$

- In case of no paths between two nodes i and j $d_{ij} = \infty$
- Well defined for disconnected graphs

Betweenness Centrality

Number of shortest paths that go through a node.

- **Assumption:** important vertices are bridges over which information flows
- **Practically:** if information spreads via shortest paths, important nodes are found on many shortest paths



$$\sigma_{jk}(i) = \text{number of geodesic path from } j \text{ to } k \text{ via } i: j \rightarrow \dots \rightarrow i \rightarrow \dots \rightarrow k$$
$$\sigma_{jk} = \text{number of geodesic path from } j \text{ to } k: j \rightarrow \dots \rightarrow k$$

Definition

$$C_b(i) = \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

Normalized def.

$$C_b(i) = \frac{1}{n^2} \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \quad \text{where} \quad C_b \in [0, 1]$$

00	Degree	<ul style="list-style-type: none"> • How many friends do you have?
01	Eigenvector	<ul style="list-style-type: none"> • Are you connected to important nodes?
02	PageRank	<ul style="list-style-type: none"> • How many important interactions do you have?
03	Katz	<ul style="list-style-type: none"> • What's your degree of influence?
04	Closeness	<ul style="list-style-type: none"> • What's your average distance w.r.t. the rest of the network?
05	Harmonic	<ul style="list-style-type: none"> • What's your harmonic average distance w.r.t. the rest of the network?
06	Betwenness	<ul style="list-style-type: none"> • How much do you help the network to stay connected?

Connectivity-based centralities Geometric centralities

Each centrality measure is a **proxy** of an underlying **network process**.

If such a process is **irrelevant** for the network than the centrality measure **makes no sense**

- E.g. If information does not spread through shortest paths, betweenness centrality is irrelevant

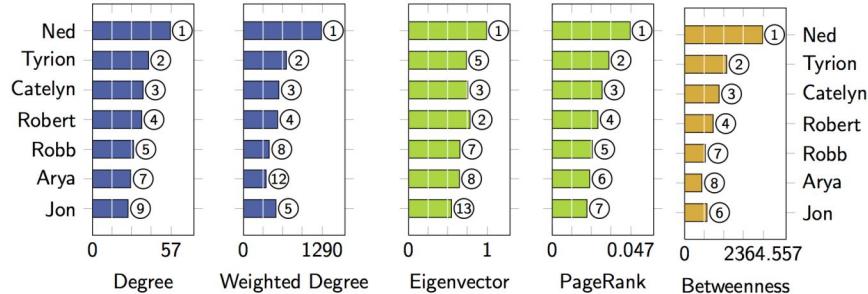
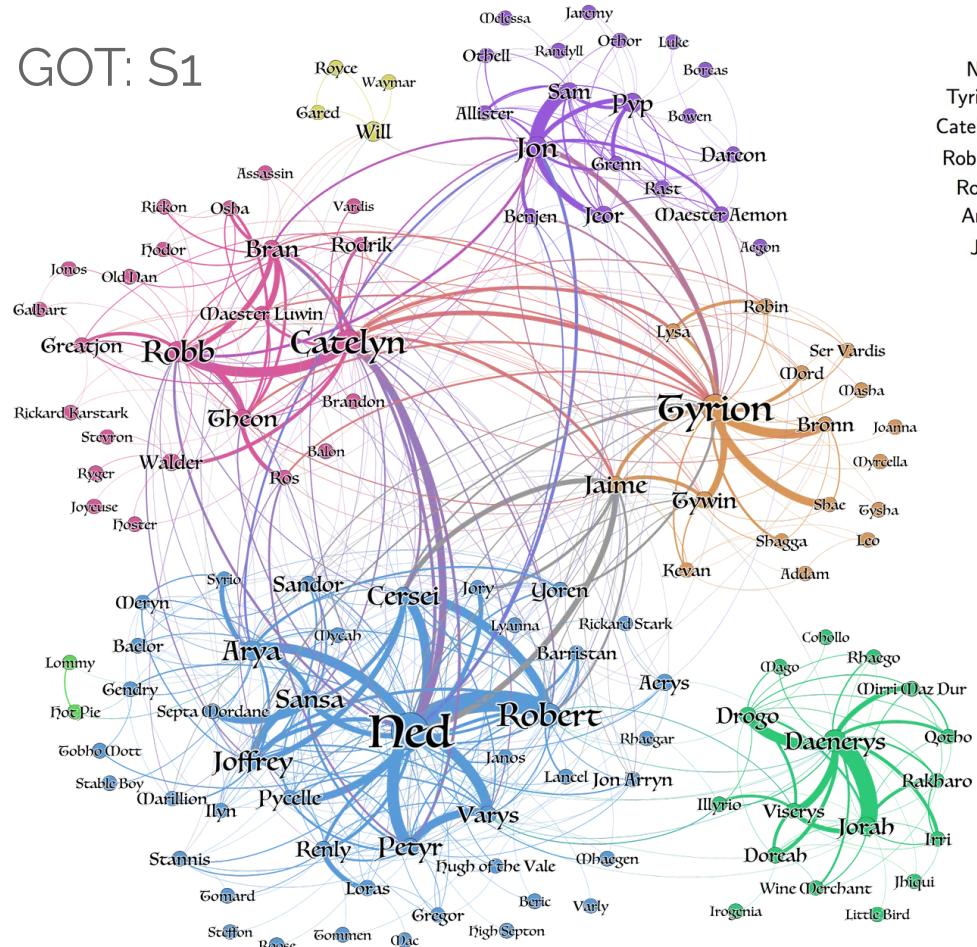
Centrality measures should be used with caution (a) for exploratory purposes and (b) for characterisation

Understanding Centralities



Data and Viz @mathbeveridge
www.networkofthrones.wordpress.com

GOT: S1



Node Label: PageRank
Node Size: Betweenness Centrality

Edge Size: #interactions
Colors: Community (with Louvain)

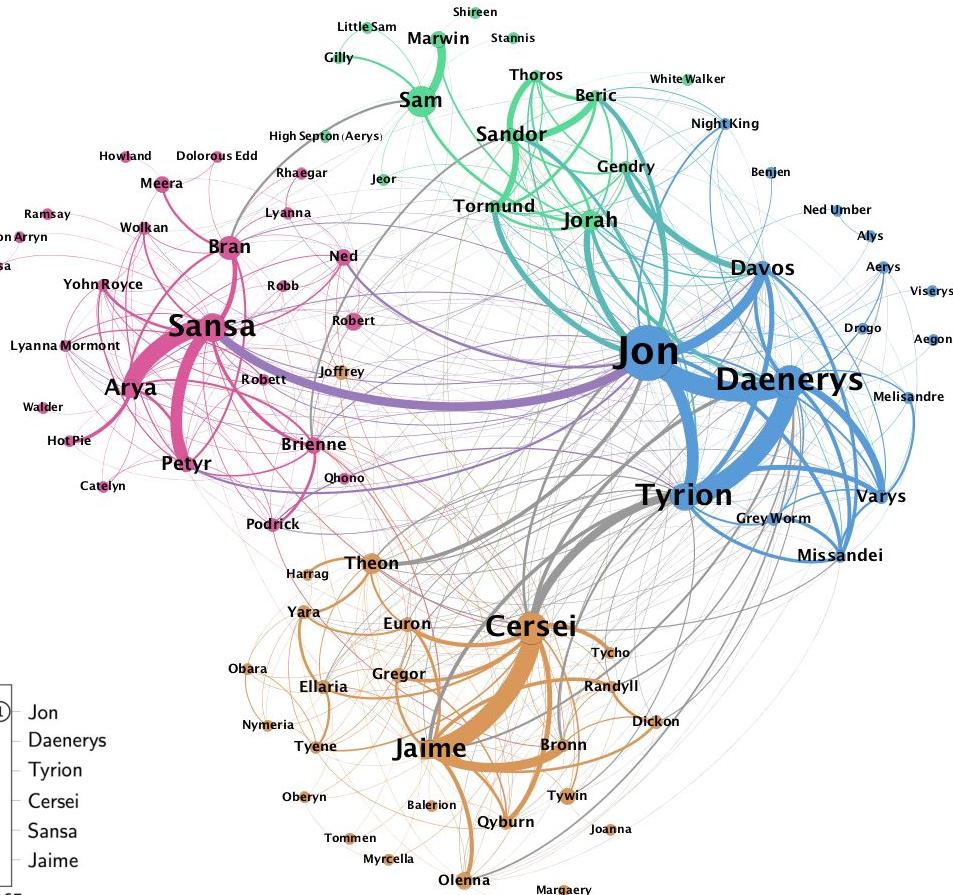
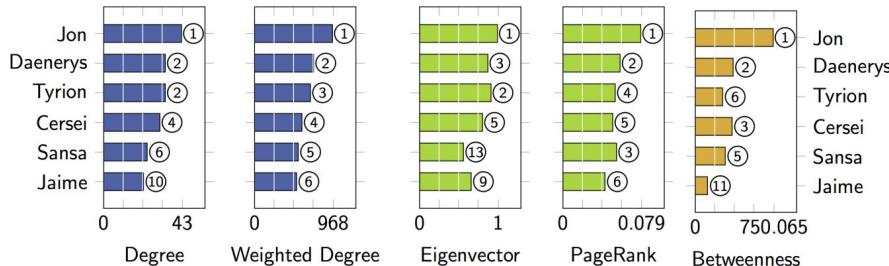
GOT: S7

Node Label: PageRank

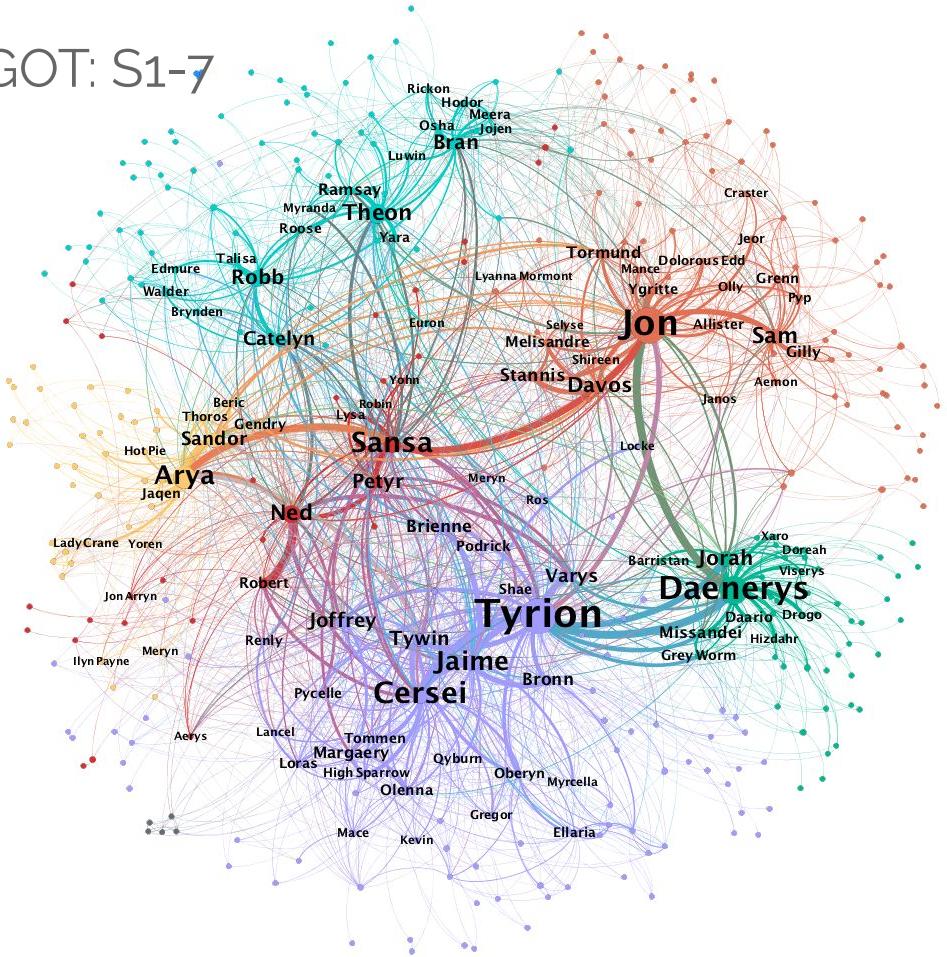
Node Size: Betweenness Centrality

Edge Size: #interactions

Colors: Community (with Louvain)



GOT: S1-7

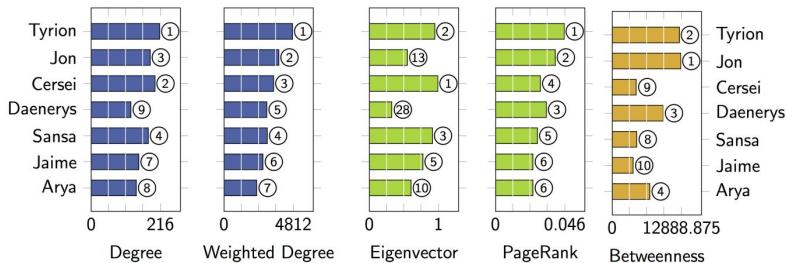


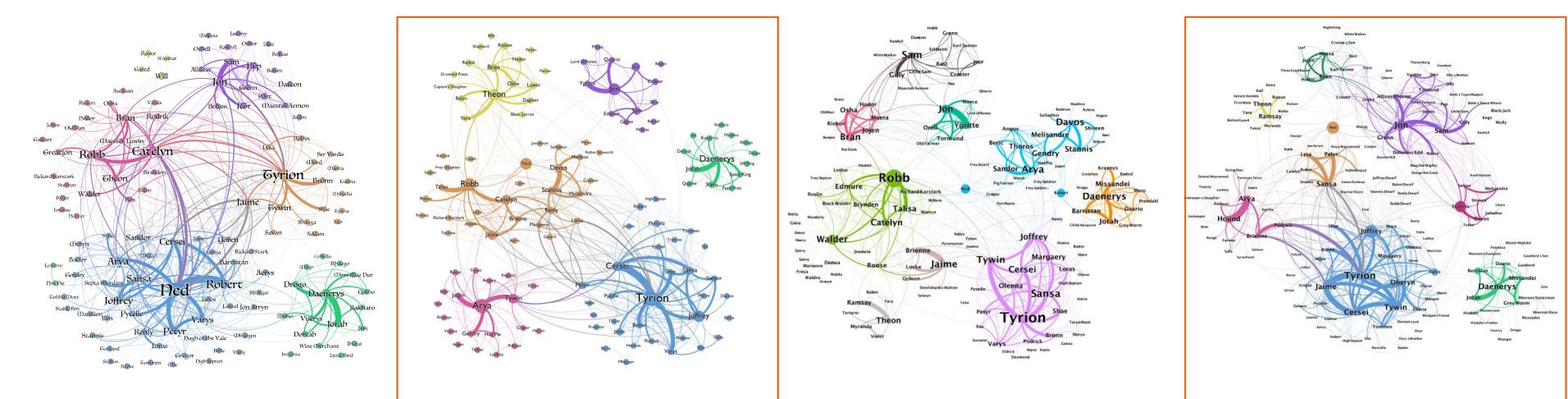
More on: www.networkofthrones.wordpress.com

All characters interactions
(up to the last season... data are coming)

Node Label: PageRank
Node Size: Betweenness Centrality

Edge Size: #interactions
Colors: Community (with Louvain)



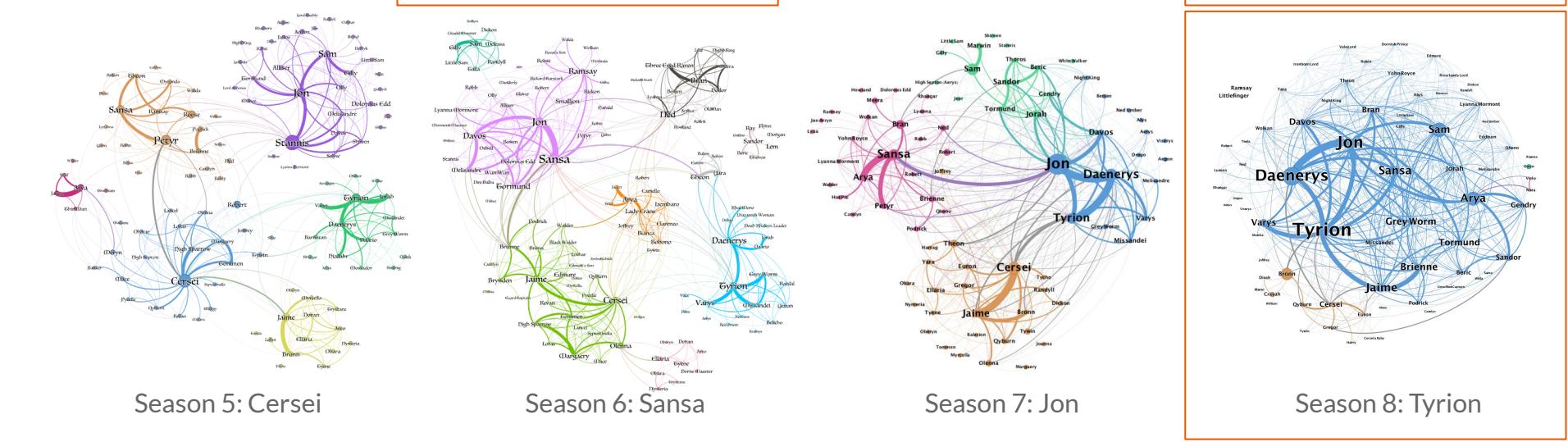


Season 1: Ned

Season 2: Tyrion

Season 3: Rob

Season 4: Tyrion



Season 5: Cersei

Season 6: Sansa

Season 7: Jon

Season 8: Tyrion

Do Birds of a Feather Flock Together?

Homophilic behaviors in complex networks



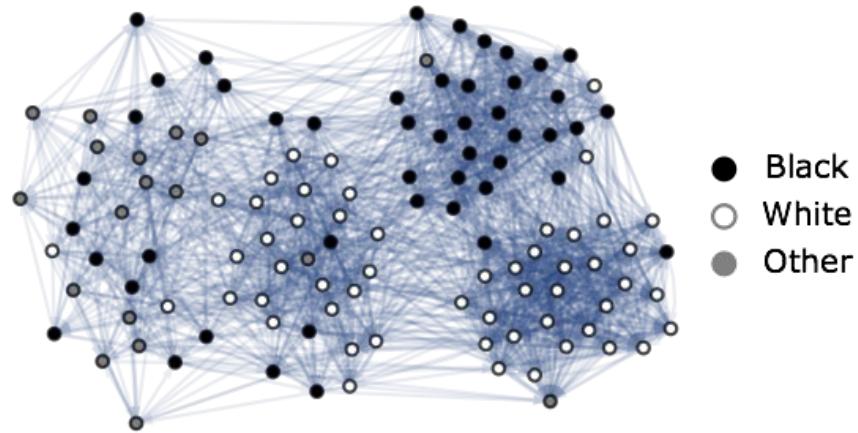
Homophily

Property of (social) networks that **nodes of the same attitude tends to be connected** with a higher probability than expected

- It appears as correlation between vertex properties of $x(i)$ and $x(j)$ if $(i,j) \in E$

Disassortative mixing:

Contrary of homophily: dissimilar nodes tend to be connected
(e.g., sexual networks, predator-prey)



Examples of Vertex properties

age, gender, nationality,
political beliefs, socioeconomic status,
obesity, ...

Homophily can be a **link creation mechanism** or **consequence of social influence** (and it is difficult to distinguish)

Assortative Mixing

(Newman's assortativity)

Quantify homophily while **discrete (categoric)** node properties are involved

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

where:

- e_{ij} fraction of links connecting nodes of type i and j
- a_i fraction of out-links from nodes of type a
- b_i fraction of in-links for type b nodes

Interpretation

- $r=0$: no assortative mixing
- $r=1$: perfectly assortative
- $-1 < r < 0$: disassortative mixing

		women				a_i
		black	hispanic	white	other	
men	black	0.258	0.016	0.035	0.013	0.323
	hispanic	0.012	0.157	0.058	0.019	0.247
	white	0.013	0.023	0.306	0.035	0.377
	other	0.005	0.007	0.024	0.016	0.053
		b_i	0.289	0.204	0.423	0.084

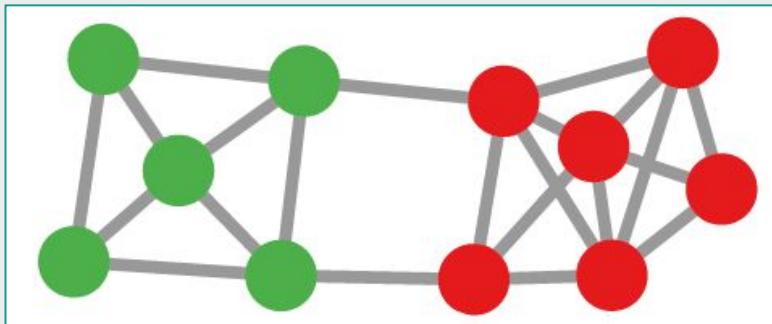
$$r = \frac{\sum_i e_{ii}}{1 - \sum_i a_i b_i} = \frac{(0.258 + 0.157 + 0.306 + 0.016) - ((0.323 \cdot 0.289) + (0.247 \cdot 0.204) + (0.377 \cdot 0.423) + (0.053 \cdot 0.083))}{(0.258 + 0.157 + 0.306 + 0.016) - ((0.323 \cdot 0.289) + (0.247 \cdot 0.204) + (0.377 \cdot 0.423) + (0.053 \cdot 0.083))} = 0.621$$

Newman, Mark EJ. "Mixing patterns in networks." Physical Review E 67.2 (2003): 026126.

Assortative Mixing

Newman's assortativity - Example

Compute Newman's assortativity given the network below:



In undirected network $a_i = b_i$

Assortativity Matrix

	Green	Red	a_i
Green	0.363	0.090	0.453
Red	0.090	0.545	0.635
b_i	0.453	0.635	

Assortativity Index

$$r = \frac{((8/22)+(12/22)) - ((10/22)^2 + (14/22)^2)}{1 - ((10/22)^2 + (14/22)^2)} = \sim 0.766$$

Newman, Mark EJ. "Mixing patterns in networks." Physical Review E 67.2 (2003): 026126.

Assortative Mixing

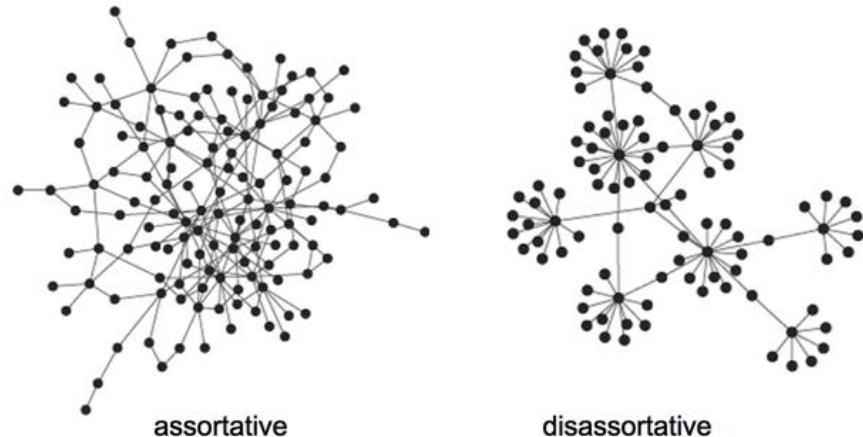
(Newman's assortativity)

Quantify homophily while **scalar** node properties are involved (e.g., *degree assortativity*)

$$R = \frac{\sum_{xy} xy(e_{xy} - a_x \cdot b_y)}{\sigma_a \sigma_b}$$

where:

- e_{xy} fraction of links connecting nodes with values x and y
- a_x fraction of out-links from nodes having value x
- b_y fraction of in-links for nodes having value y
- σ standard deviations



Degree assortative

Nodes tends to connect homogeneously w.r.t. their degree (e.g., hubs with hubs)

Degree disassortative

Nodes tends to connect in a star-like topology

Newman, Mark EJ. "Mixing patterns in networks." Physical Review E 67.2 (2003): 026126.

	network	type	size n	assortativity r
social	physics coauthorship	undirected	52 909	0.363
	biology coauthorship	undirected	1 520 251	0.127
	mathematics coauthorship	undirected	253 339	0.120
	film actor collaborations	undirected	449 913	0.208
	company directors	undirected	7 673	0.276
	student relationships	undirected	573	-0.029
	email address books	directed	16 881	0.092
technological	power grid	undirected	4 941	-0.003
	Internet	undirected	10 697	-0.189
	World-Wide Web	directed	269 504	-0.067
	software dependencies	directed	3 162	-0.016
biological	protein interactions	undirected	2 115	-0.156
	metabolic network	undirected	765	-0.240
	neural network	directed	307	-0.226
	marine food web	directed	134	-0.263
	freshwater food web	directed	92	-0.326

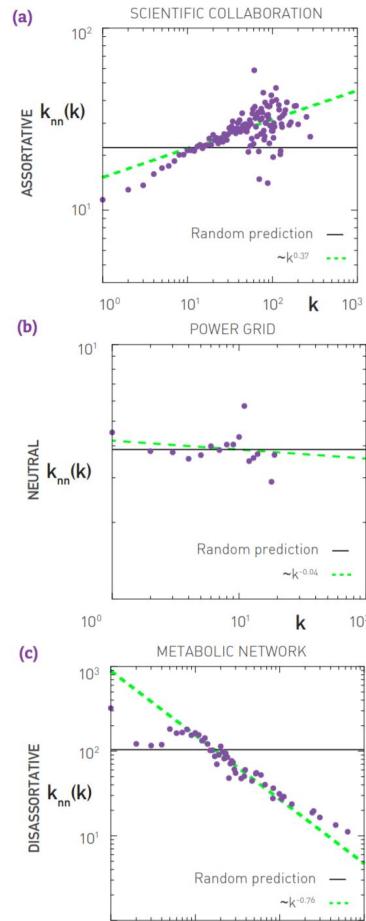
Assortativity by **degree** w.r.t. different network **types**

Degree correlation (insights)

- Using a single number: Newman's assortativity
- Plotting $k_{nn}(k)$ in function of k , where

$$k_{nn}(k) = \sum_{k'} k' P(k'|k),$$

is a **degree correlation function**, e.g., the average degree of the neighbors of all degree- k nodes.



Clumpiness

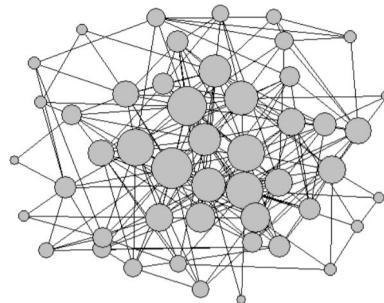
Similar (**scalar**) characteristics are more prominent along **short distances**
e.g., high-degree nodes are separated by few links

$$\Lambda(G, k, \alpha) = \sum_{i>j}^{n(n-1)/2} \frac{k_i k_j}{(d_{ij})^\alpha},$$

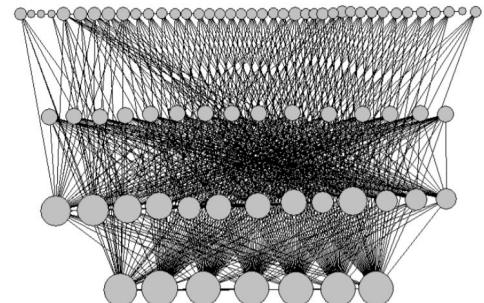
where:

- k_i the degree of node i
- $\alpha > 0$ a real parameter
- d_{ij} the distance between nodes i and j

a) Clumped Assortative



b) Clumped Disassortative



Clumped assortative

Λ increases with the increase of the degrees of the nodes in the network

Clumped disassortative

Λ decreases with the increase in the separation between the degrees of the nodes in the network

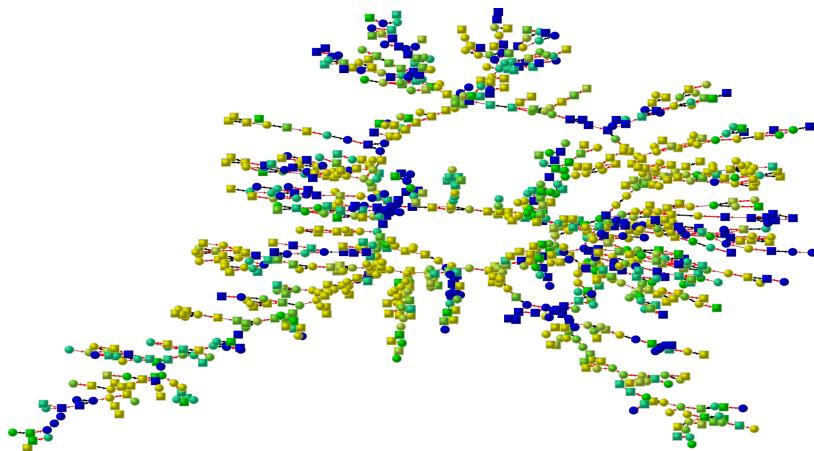
Estrada, N. Hatano, A. Gutierrez, "Clumpiness mixing in complex networks", Journal of Statistical Mechanics: Theory and Experiment.

Examples and Case Study

—



Case study: Happiness



James H. Fowler, Nicholas A. Christakis.
*Dynamic Spread of Happiness in a Large Social Network:
Longitudinal Analysis Over 20 Years in the Framingham Heart Study*
British Medical Journal 337 (4 December 2008)

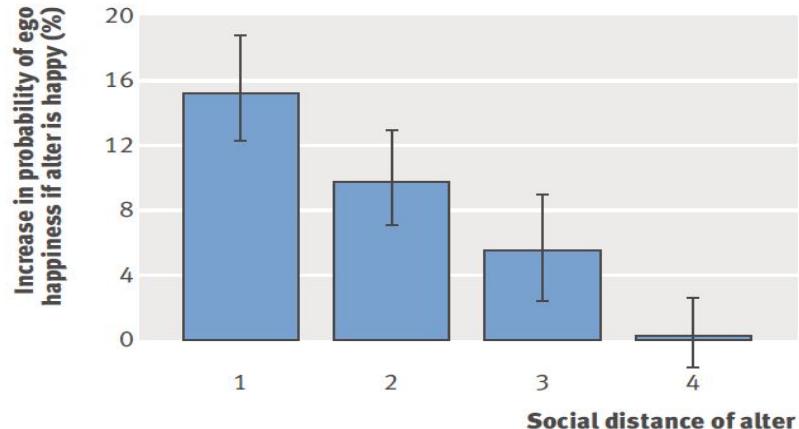
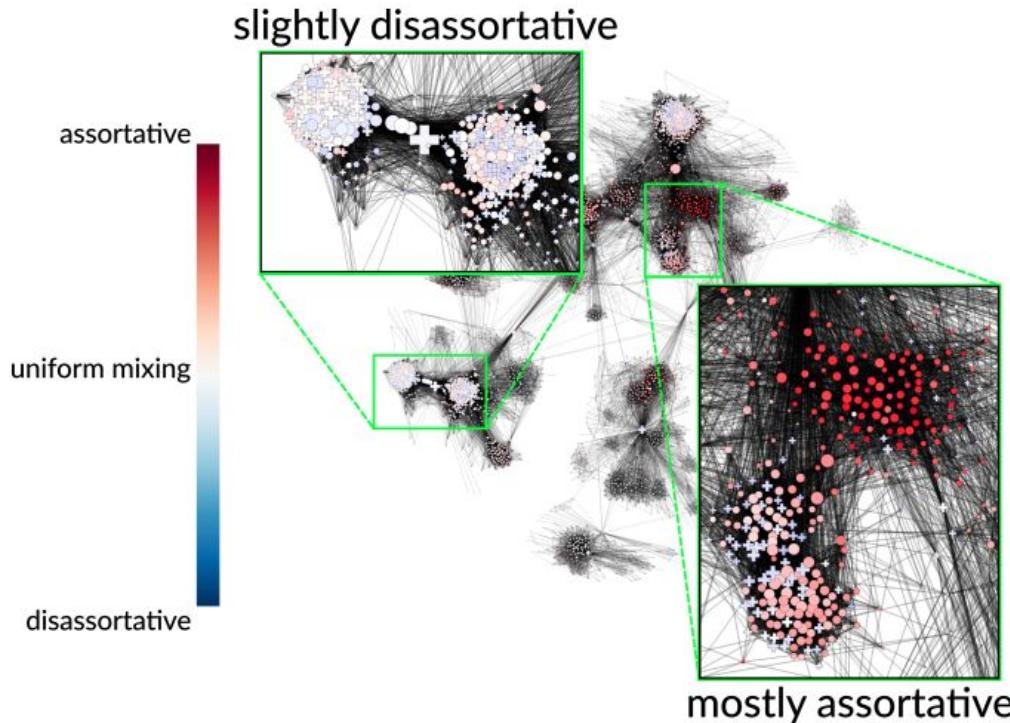


Fig 2 | Social distance and happiness in the Framingham social network. Percentage increase in likelihood an ego is happy if friend or family member at certain social distance is happy (instead of unhappy). The relationship is strongest between individuals who are directly connected but remains significantly >0 at social distances up to three degrees of separation, meaning that a person's happiness is associated with happiness of people up to three degrees removed from them in the network

Is a Global Measure enough?

"Sure I can work with the means, but I'd rather party with the outliers..."

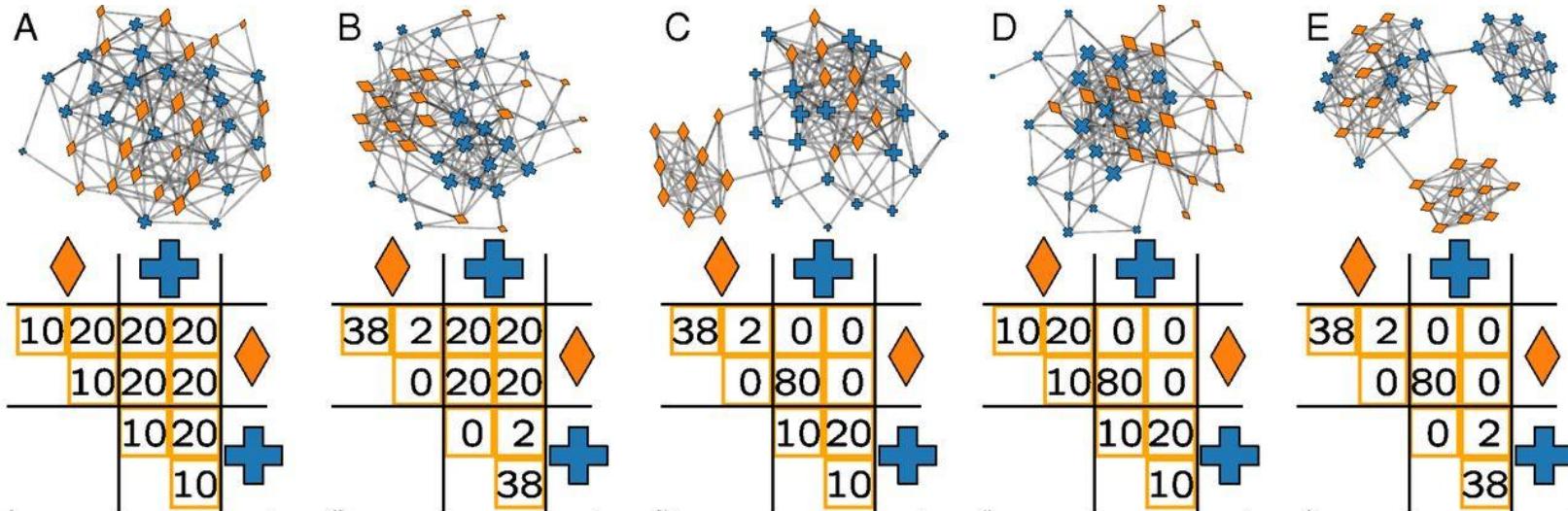




Local assortativity of gender in a sample of Facebook friendships (McAuley and Leskovec 2012).

Different regions of the graph exhibit strikingly different patterns, suggesting that a single variable, e.g. **global assortativity (Newman's)**, would provide a **poor description** of the system.

Limits of a **global** assortativity score



Five networks (top) of $n=40$ nodes and $m=160$ edges with the same global assortativity $r=0$

Moving toward a **multiscale** approach to measure assortativity

Multiscale Mixing Patterns

Idea:

A local measure that captures the mixing patterns within the local neighbourhood of a given node.

Trivial solution:

Consider only the node's neighbors

- issue with sample size
(what about low degree nodes?)

Better approximation:

Considering the **stable state of a RW**
(probability to reach a given node)
to weight the edges

Issue:

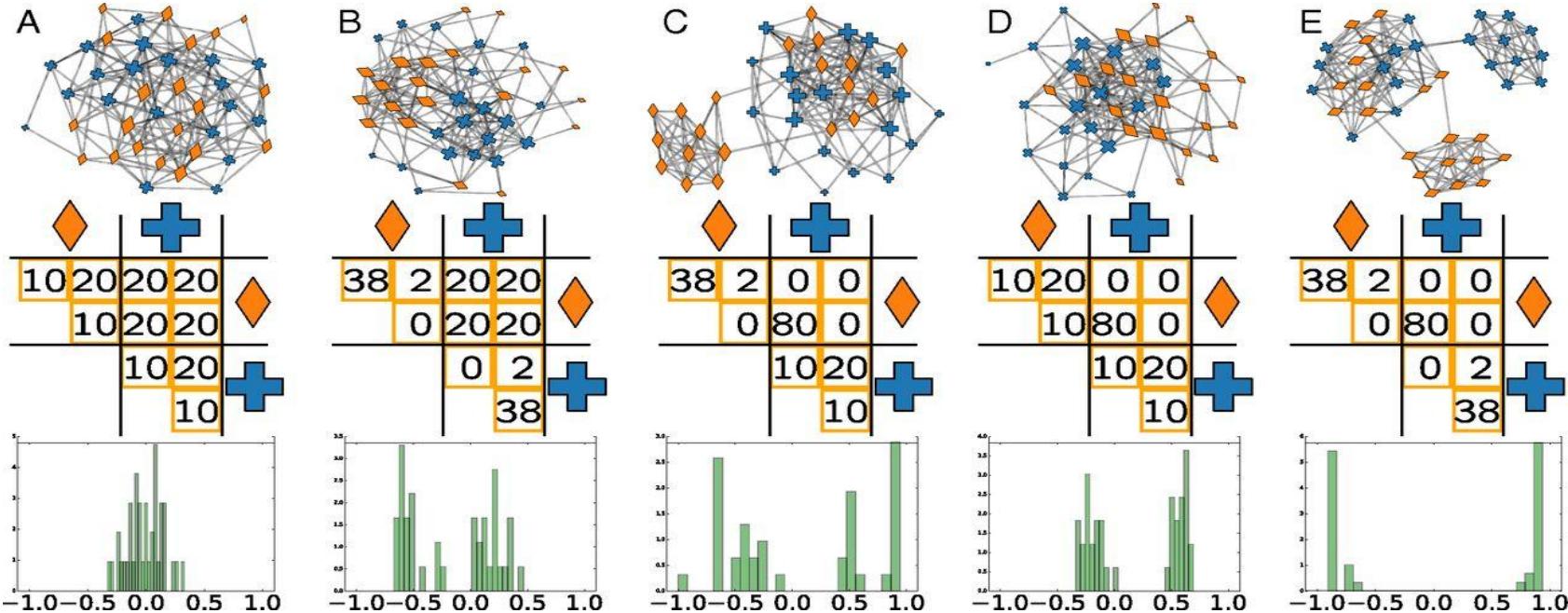
Need to fix the value of α

- $\alpha=0$ the RW stays put,
- $\alpha=1$ the RW never restarts

Solution:

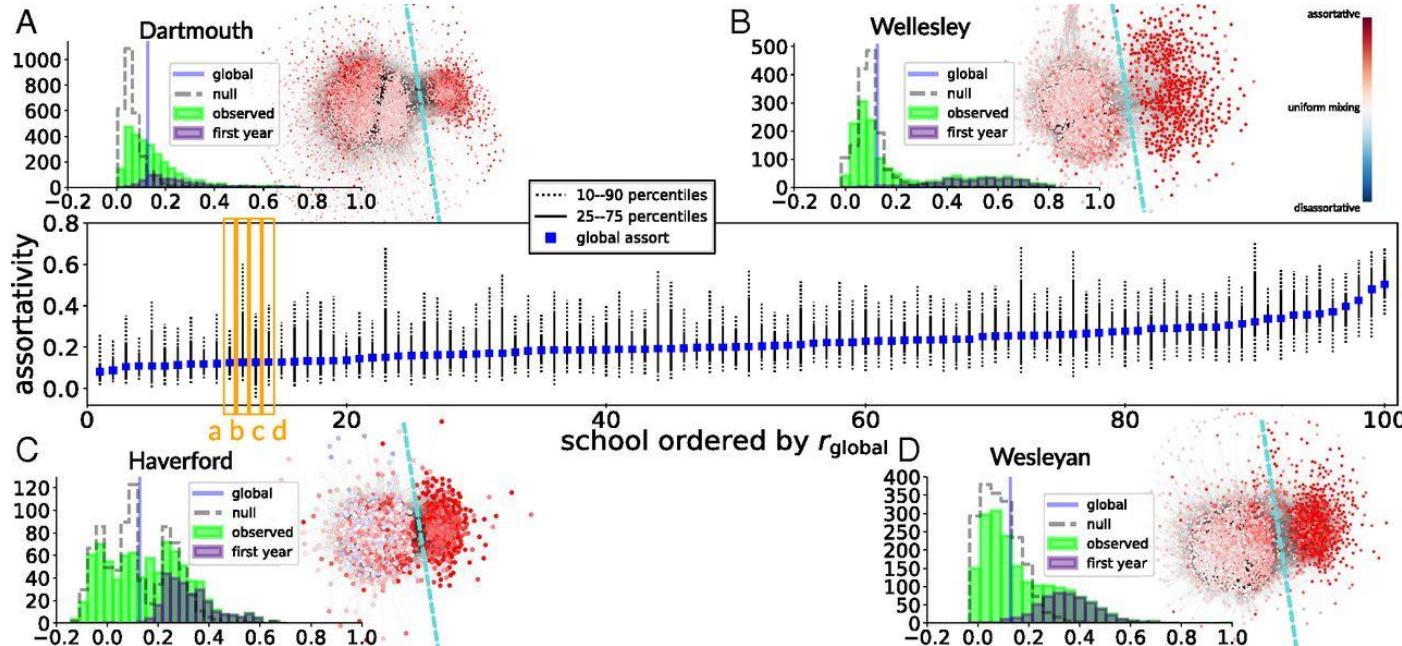
Integrate over all possible α (multiscale approach)

$$w_{\text{multi}}(i; \ell) = \int_0^1 w_\alpha(i; \ell) d\alpha$$



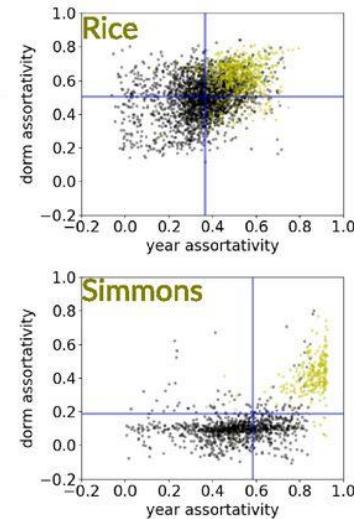
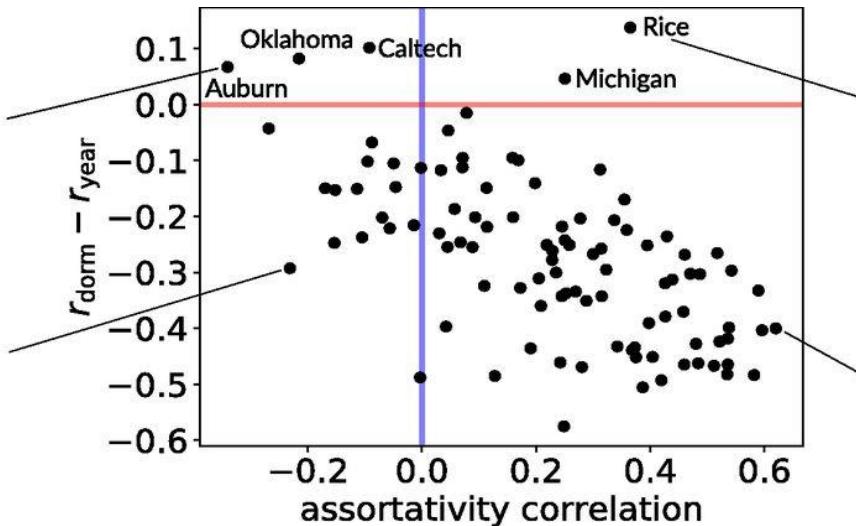
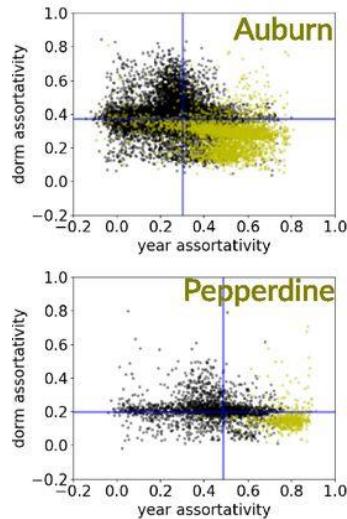
Peel's Quintet: Only A has a random-like assortment distribution

Evaluation on real data



Facebook100
Distribution of local assortativity for the “dorm” node feature

Evaluation on real data (cont'd)



Facebook100

Correlation of local assortativities by dorm and matriculation year (x axis)
and proportion of nodes which are more assortative by dorm than by year (y axis).

Distance matter: Conformity

Idea:

Random walkers flatten reachability distances:
can we weight local assortativity considering
shortest paths?

Solution:

Clumpiness like modeling:

- the closer similar nodes are, the higher their local "assortativity"

Strengths:

- Higher informative power w.r.t. Multiscale-Mixing
- Solves the "range" problem (the full interval [-1,1] is covered)
- Tunable distance dumping factor
- Allows for multi label profiles

Weaknesses:

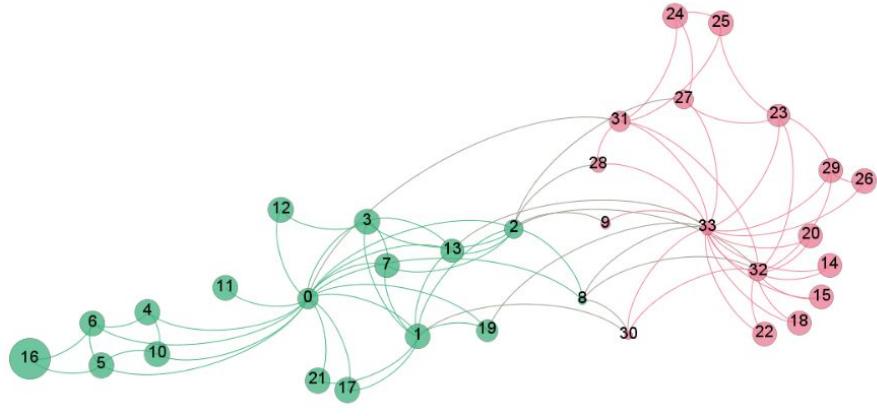
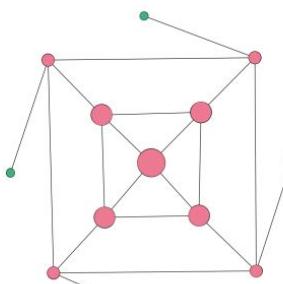
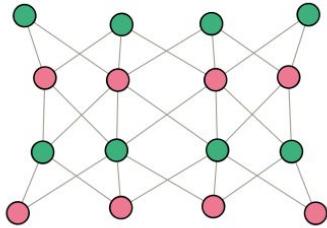
Costly computation

- requires the extraction of all shortest paths

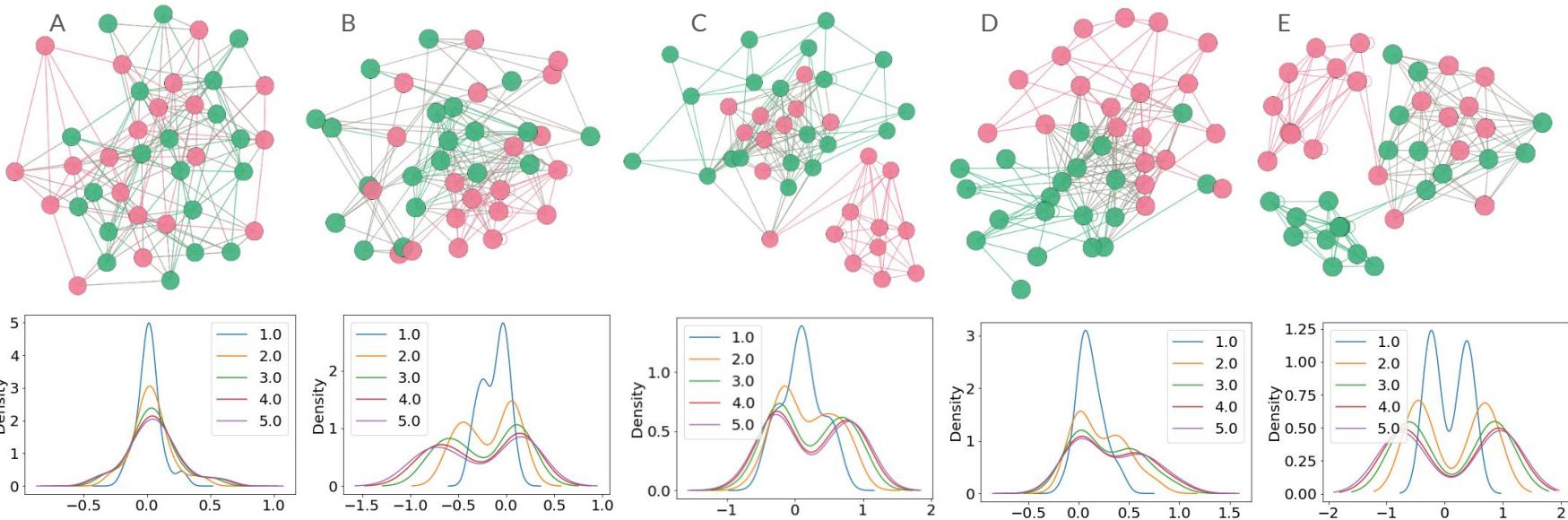
$$\psi(u, \alpha) = \frac{\sum_{d \in D} \frac{\sum_{v \in N_{u,d}} I_{u,v} f_{v,t_v}}{|N_{u,d}| d^\alpha}}{\sum_{d \in D} d^{-\alpha}},$$

similarity
distance dumping

Conformity: Intuition

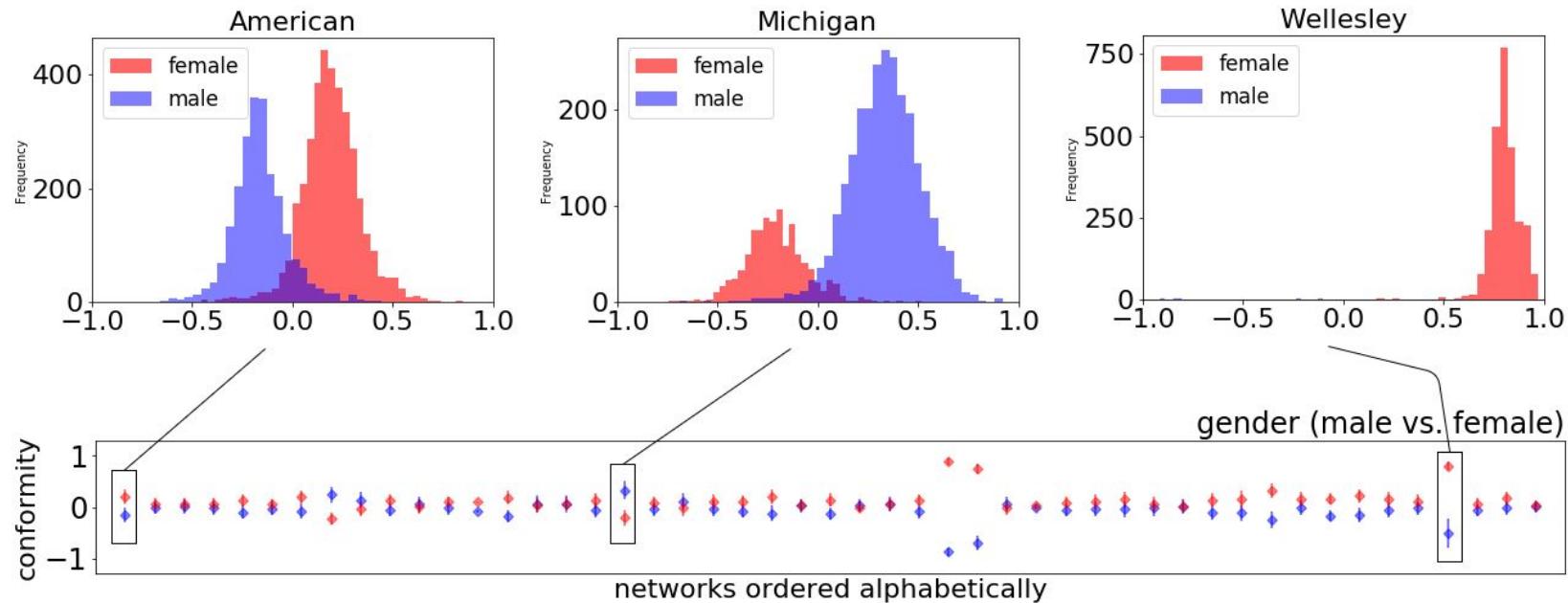


Node colors map categorical attribute values, while node sizes encode the respective Conformity scores (the smaller the size, the lower the score)



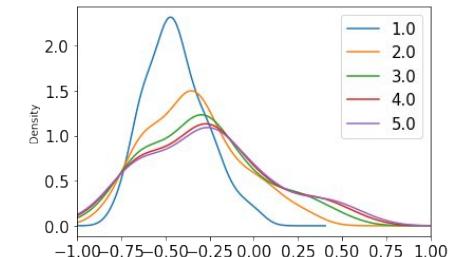
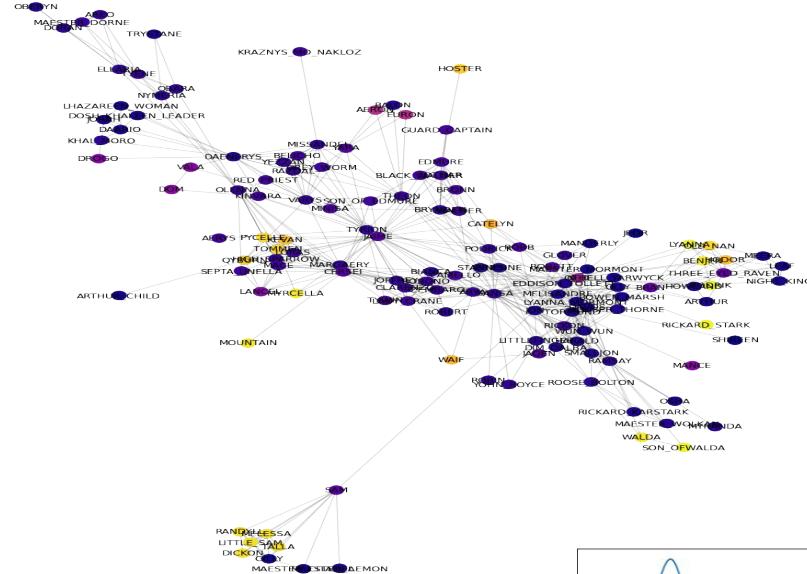
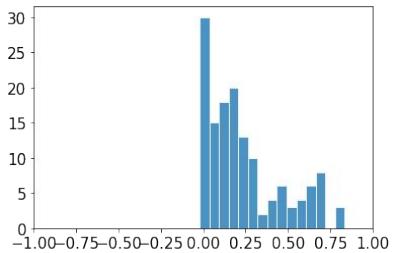
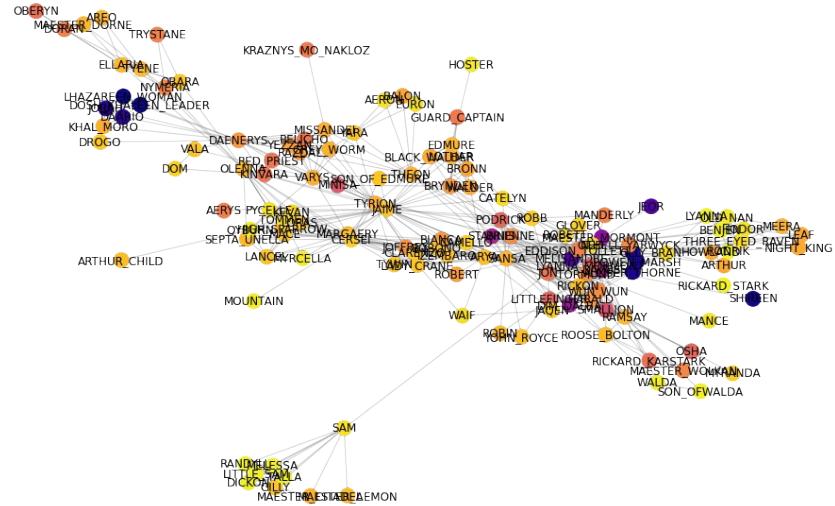
Peel's Quintet: Only A has a random-like assortment distribution

Evaluation on real data



Facebook100
Distribution of [local assortativity](#) for the “gender” node feature

Multiscale Mixing Vs. Conformity



GoT Season 6
Node attributes: Houses. Global attribute assortativity: 0.2
Multiscale Mixing (left), Conformity (right)

Chapter 6

Conclusion

Take Away Messages

1. Nodes positions play an important role in network topology
2. Different centralities allows to capture, valuable, information
3. In social contexts individuals tend to cluster following homophilic patterns

Suggested Readings

- Chapters 3 & 4 of Kleinberg's book

What's Next

Chapter 7:
Tie Strength & Resilience

Notebook

Chapter 6: Centrality & Assortativity
https://github.com/sna-unipi/SNA_lectures_notebooks

