Chapter 6

Centrality & Assortative Mixing

Summary

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

Reading

- Chapters 3 & 4 of Kleinberg's book

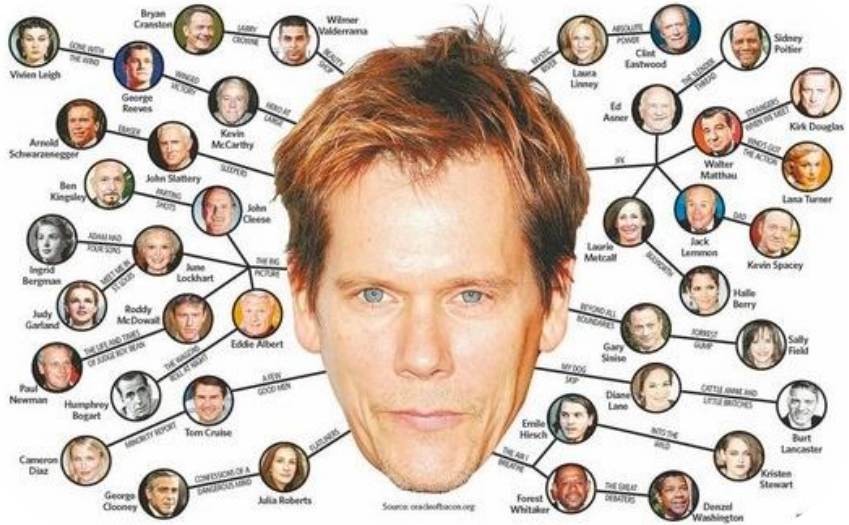
How important is a node in a network?

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/

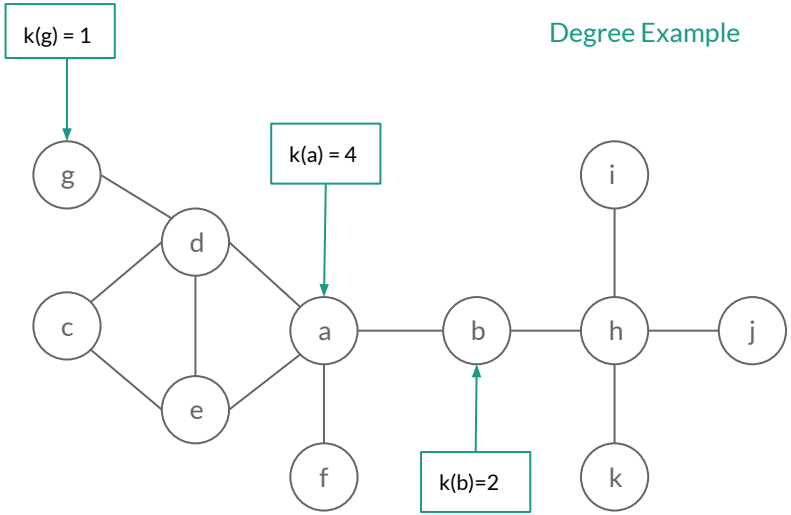
Degree Centrality

How many neighbors does a node have?

Often enough to find important nodes
(e.g., main characters of a 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

$A_{i,j}$ { 1 if i and j are connected, 0 otherwise

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

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)

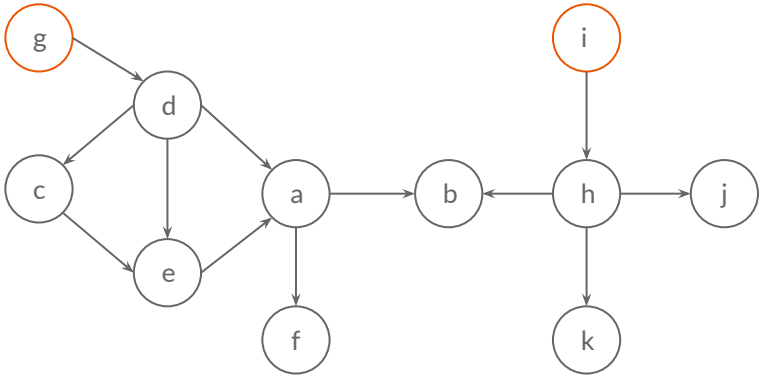
Eigenvector Centrality

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

Eigenvector Example



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
- ...

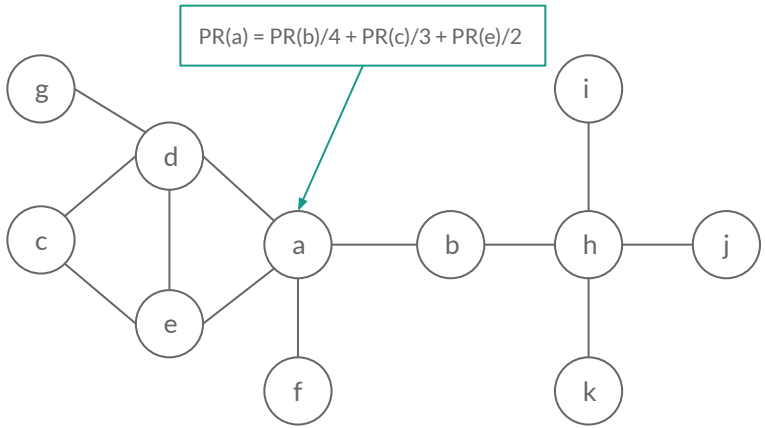
PageRank

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

PageRank Example



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

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-α it will jump to a random node

- α< 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 α=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

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

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

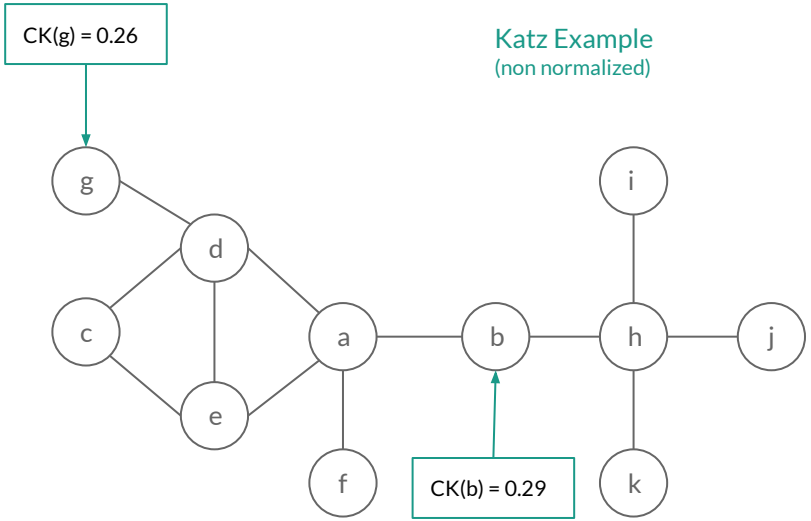
$\ell(p_i,p_j)$ is the ratio between number of links outbound from page j to page i to the total number of outbound links of page j

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**→**j** of length **k**
(multiplied by an attenuation factor α)



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

NB: Attenuation factor α must be smaller than 1/|λ| , 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"

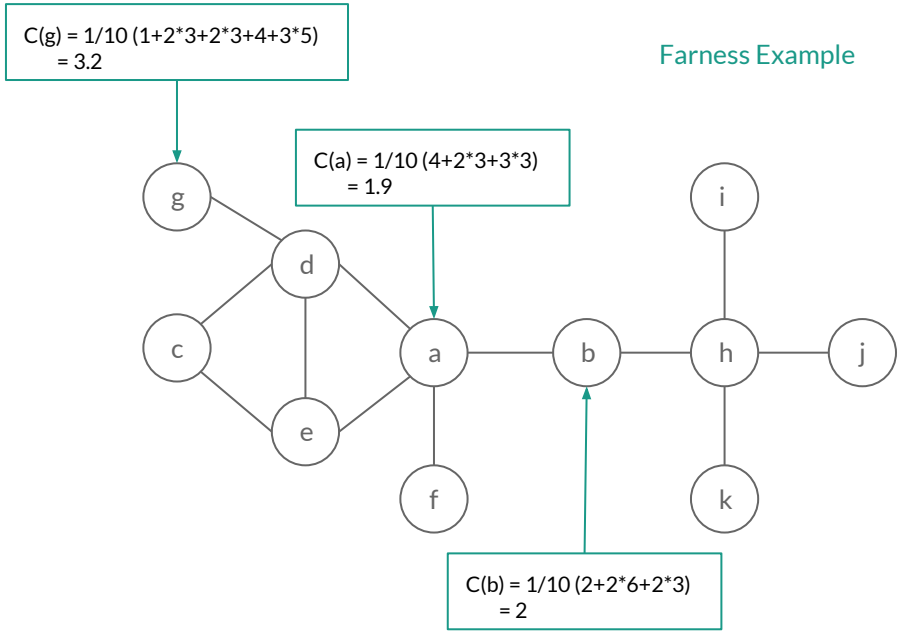
Closeness Centrality

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

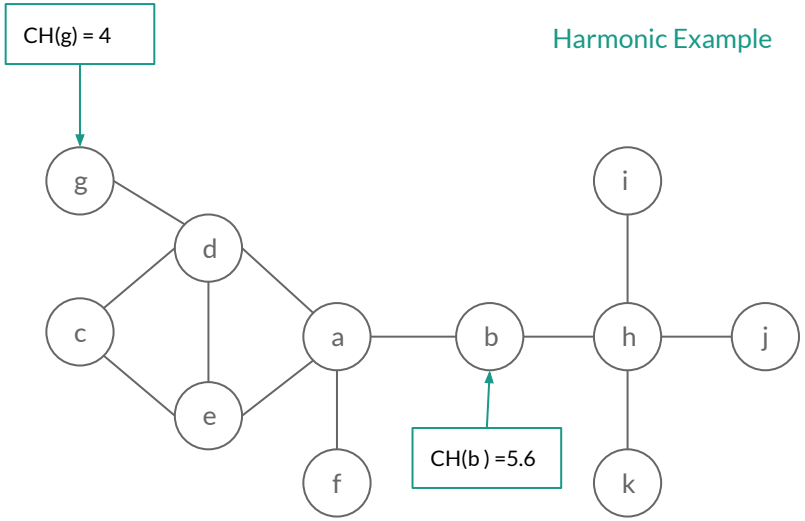


Closeness Formula

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

Harmonic Centrality

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



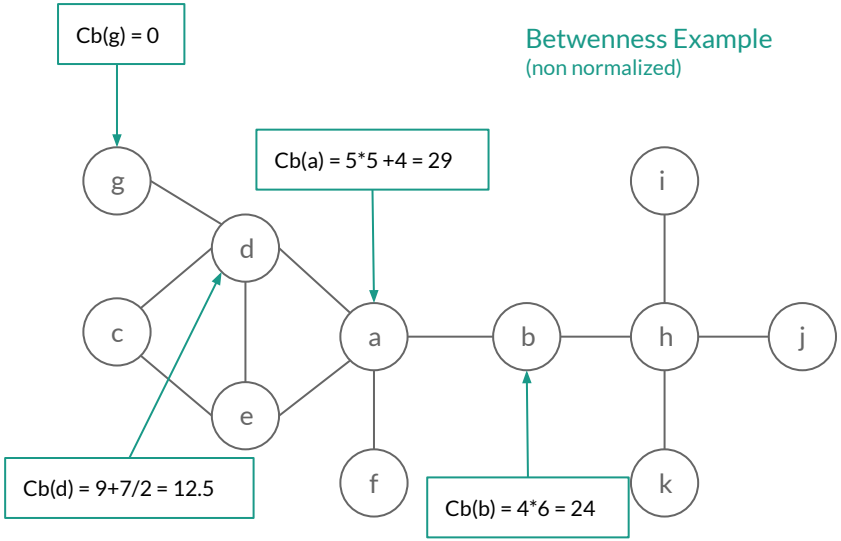
$$CH(i) = \frac{1}{n-1} \sum_{i \neq j} \frac{1}{d_{ij}}$$

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

Betwenness 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



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

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

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

Summarizing...

00	Degree	How many friends do you have?
01	Eigenvector	Are you connected to important nodes?
02	PageRank	How many important interactions do you have?
03	Katz	What's your degree of influence?
04	Closeness	What's your average distance w.r.t. the rest of the network?
05	Harmonic	What's your harmonic average distance w.r.t. the rest of the network?
06	Betwenness	How much do you help the network to stay connected?

Connectivity-based centralities

Geometric centralities

Each centrality measures is a proxy of an underlying network process.

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

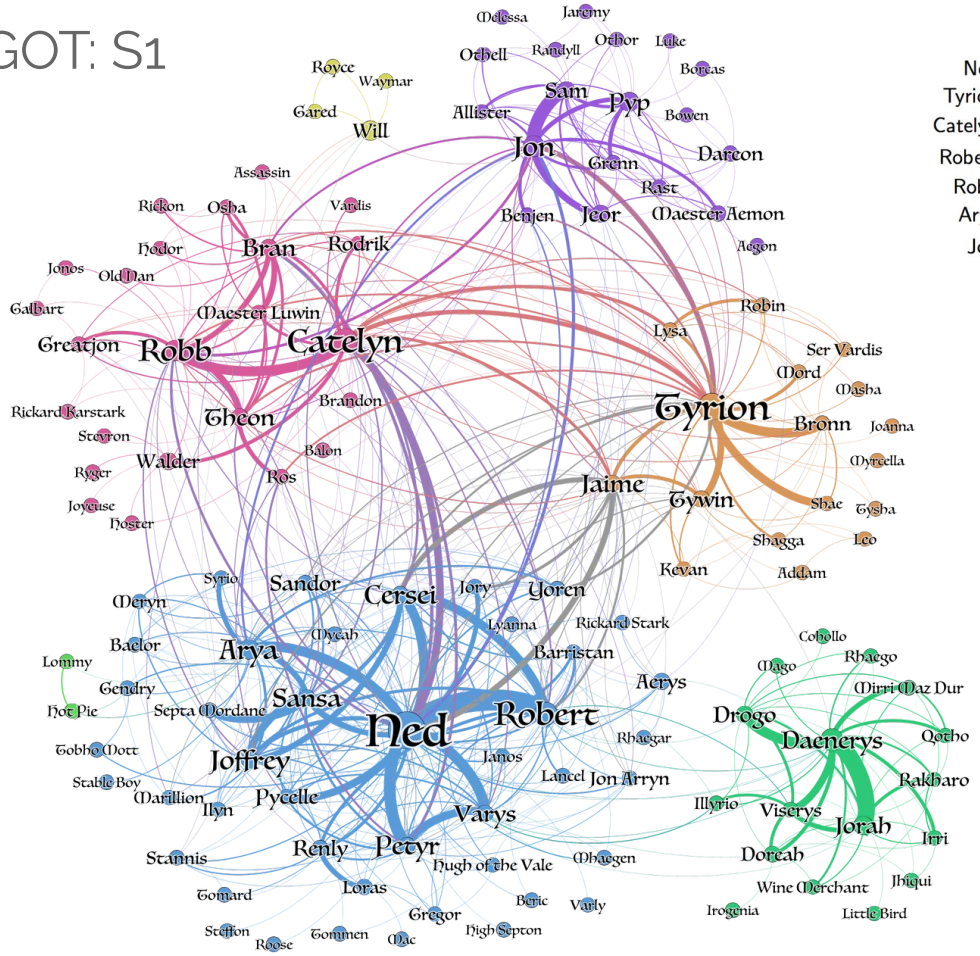
- E.g. If information does not spread through shortest paths, betweenness centrality is irrelevant

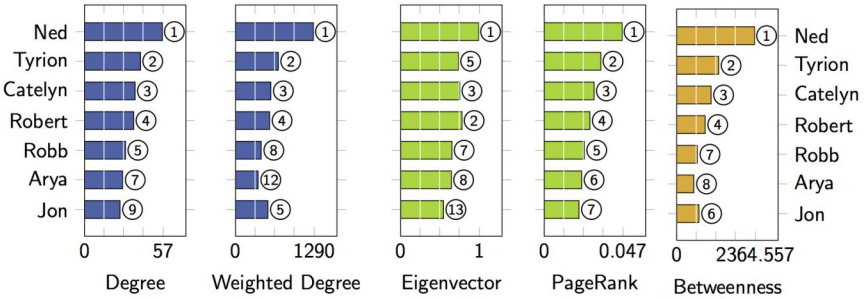
Centrality measures should be used with caution for (a) for exploratory purposes and (b) for characterisation

Understanding Centralities

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

GOT: S1





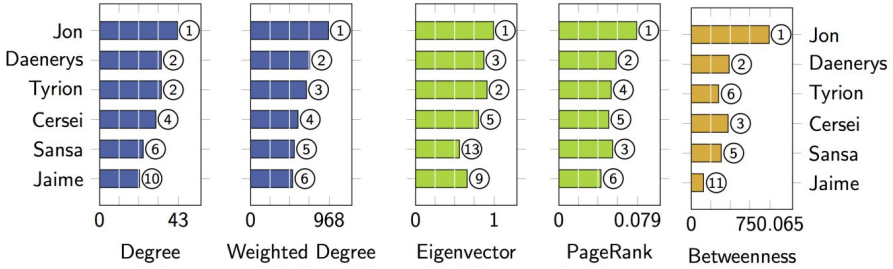
Node Label: PageRank
Node Size: Betweenness Centrality

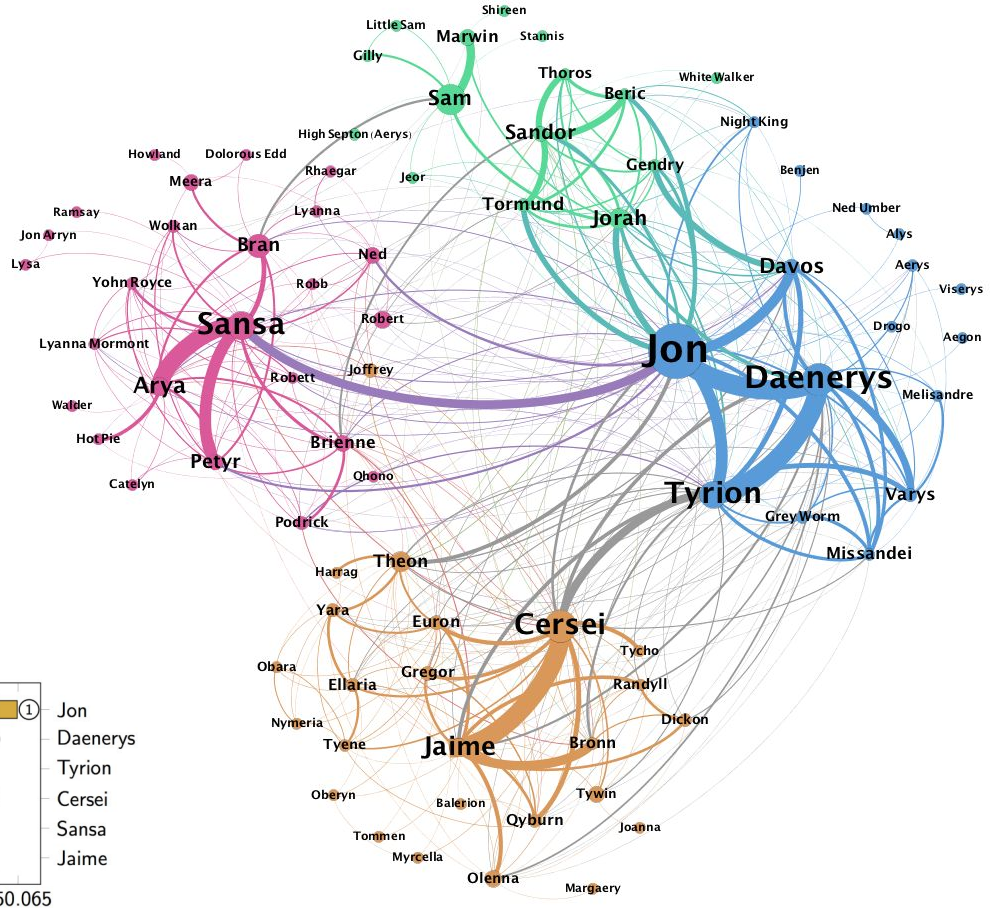
Edge Size: #interactions
Colors: Community (with Louvain)

GOT: S7

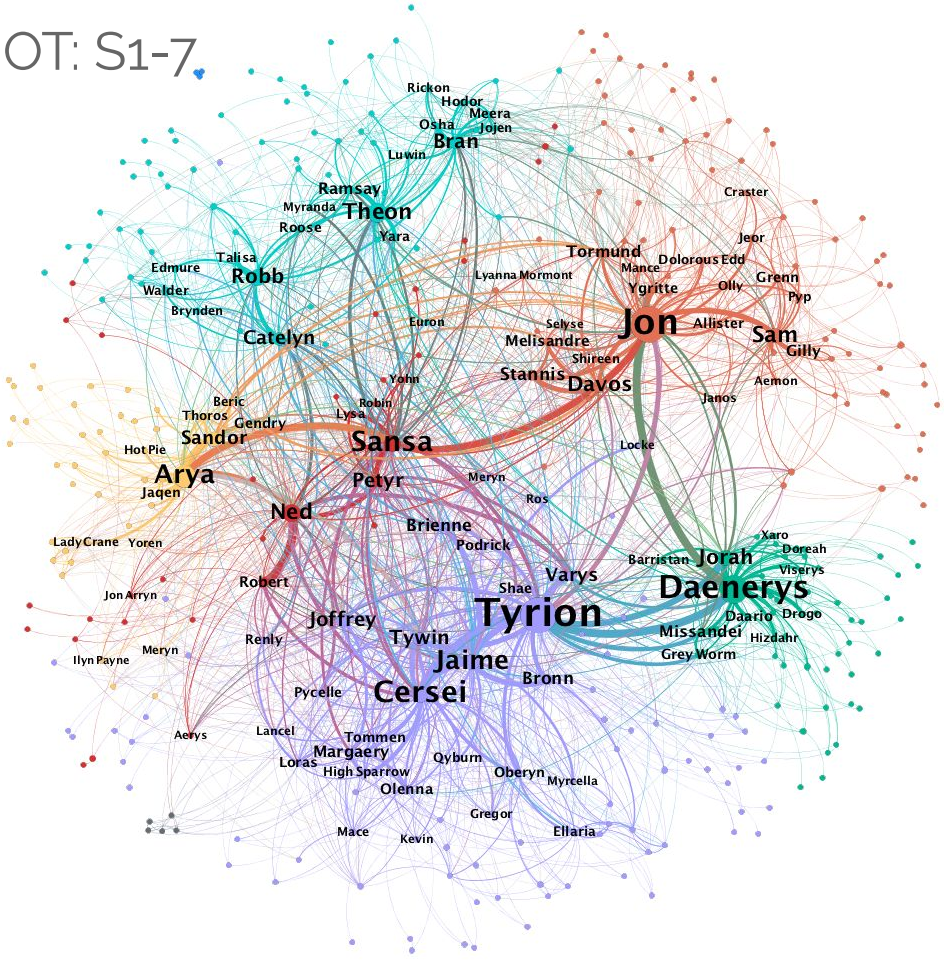
Node Label: PageRank
Node Size: Betweenness Centrality

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





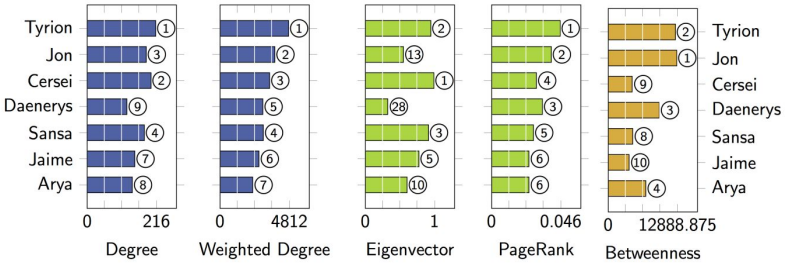
GOT: S1-7



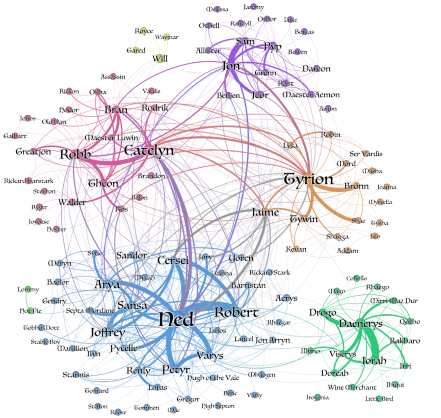
All characters interactions
(up to the last season... data are coming)

Node Label: PageRank
Node Size: Betweenness Centrality

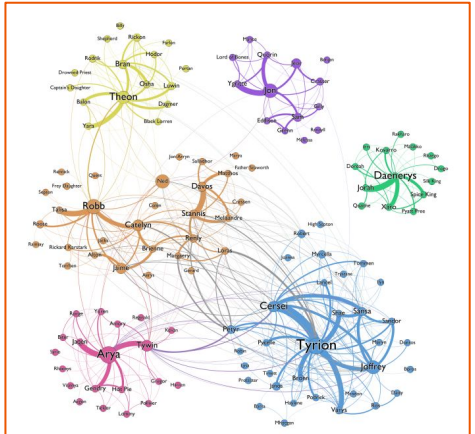
Edge Size: #interactions
Colors: Community (with Louvain)



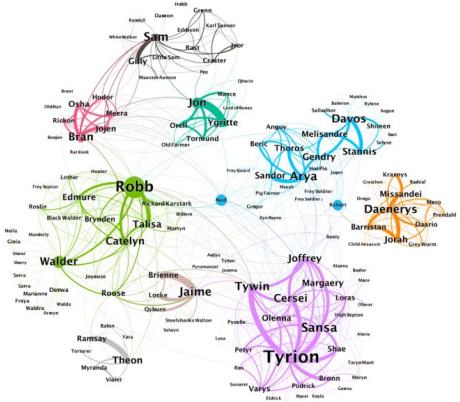
More on: www.networkofthrones.wordpress.com



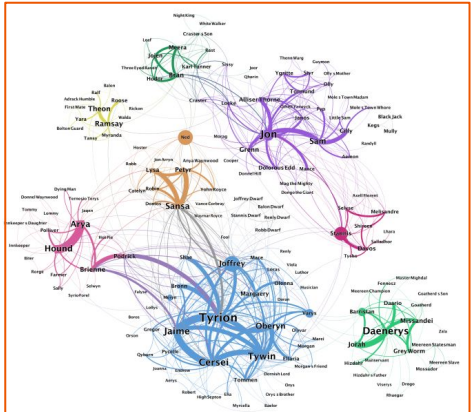
Season 1: Ned



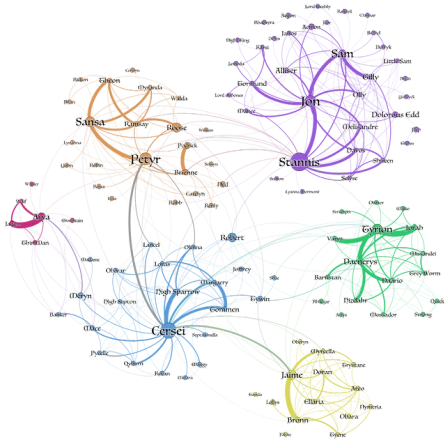
Season 2: Tyrion



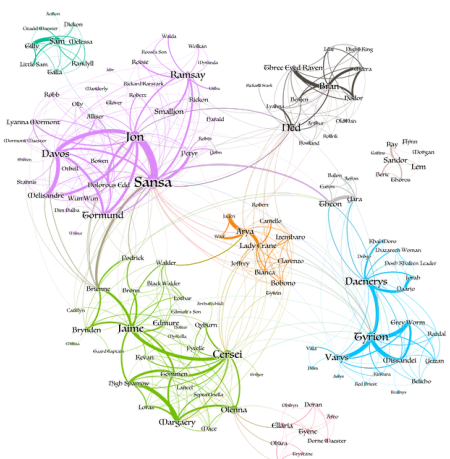
Season 3: Rob



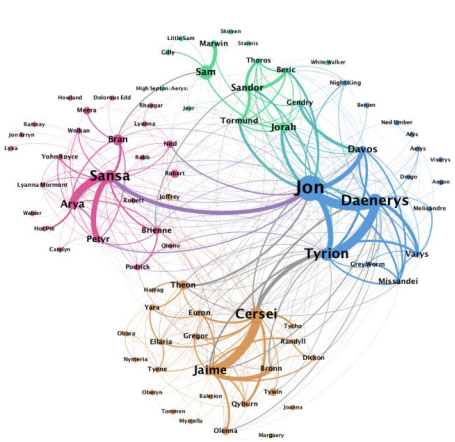
Season 4: Tyrion



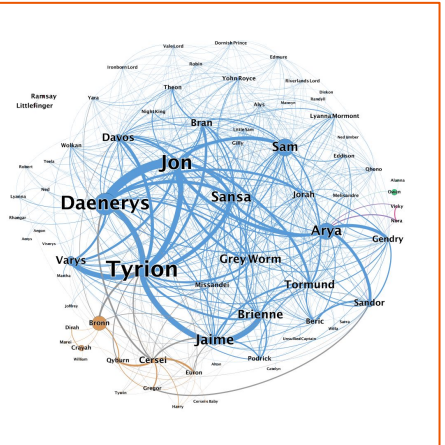
Season 5: Cersei



Season 6: Sansa



Season 7: Jon



Season 8: Tyrion

Do Birds of a Feather Flock Together?

Homophilic behaviors in complex networks

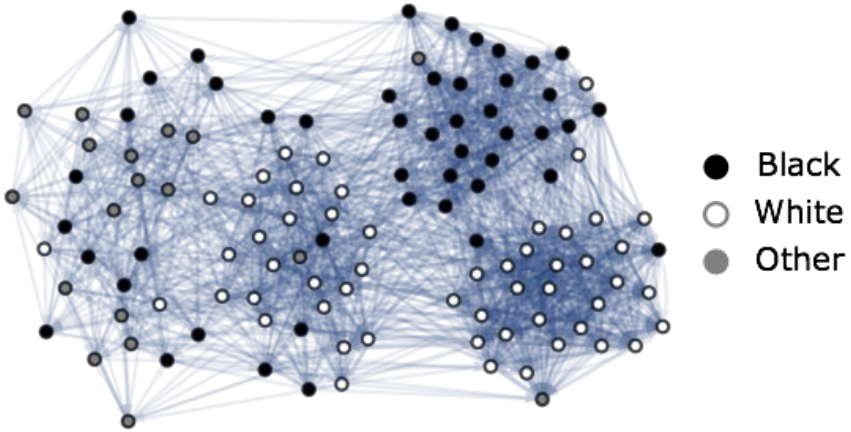
Homophily

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) ∈ 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 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				
		black	hispanic	white	other	a_i
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 = 0.621

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

Assortative Mixing

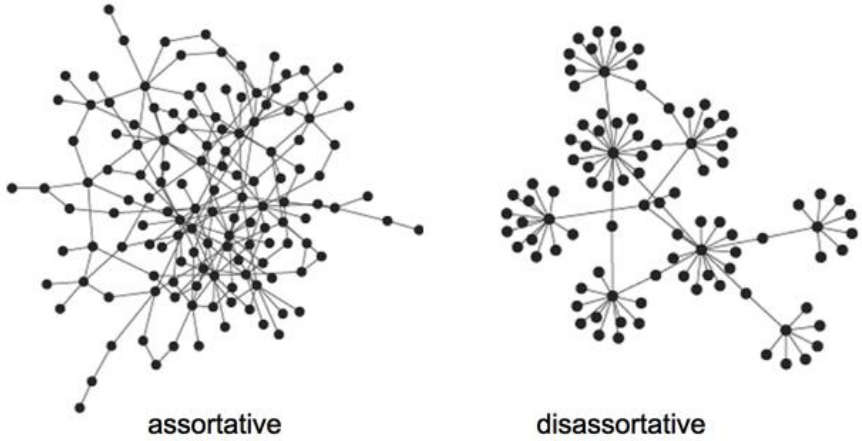
(Newman's assortativity)

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

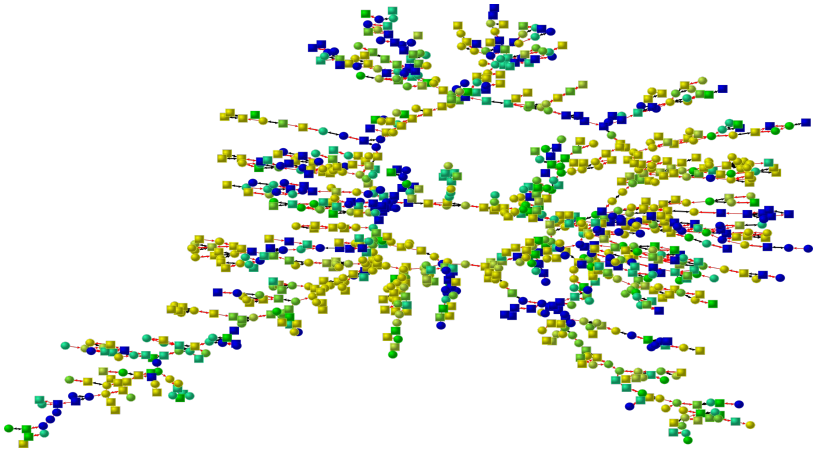
Newman, Mark EJ. "Mixing patterns in networks." *Physical Review E* 67.2 (2003): 026126.

Examples and Case Study

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)



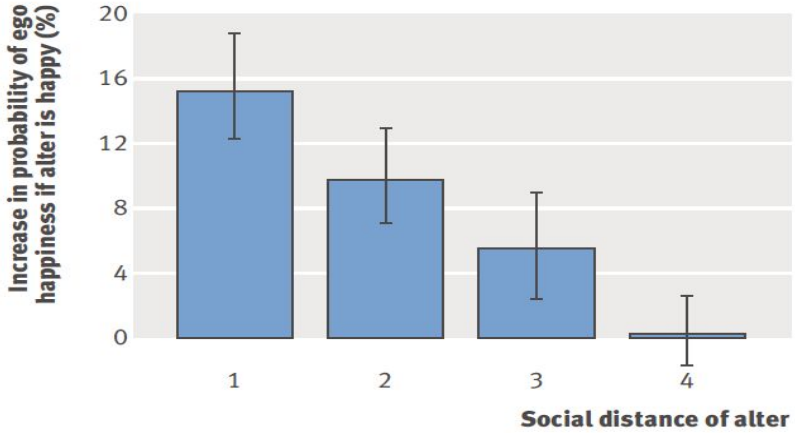
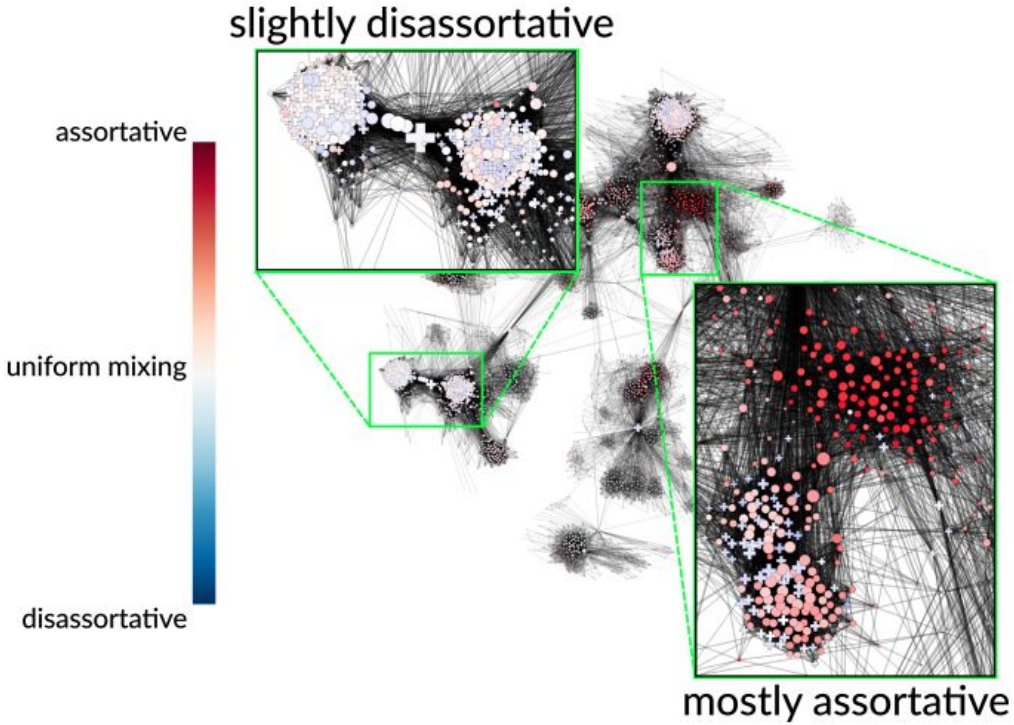


Fig 2 | Social distance and happiness in the Framingham social network. Percentage increase in likelihood an ego is happy if friend or family member at certain social distance is happy (instead of unhappy). The relationship is strongest between individuals who are directly connected but remains significantly >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

Case study: Happiness

Is a Global Measure enough?

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

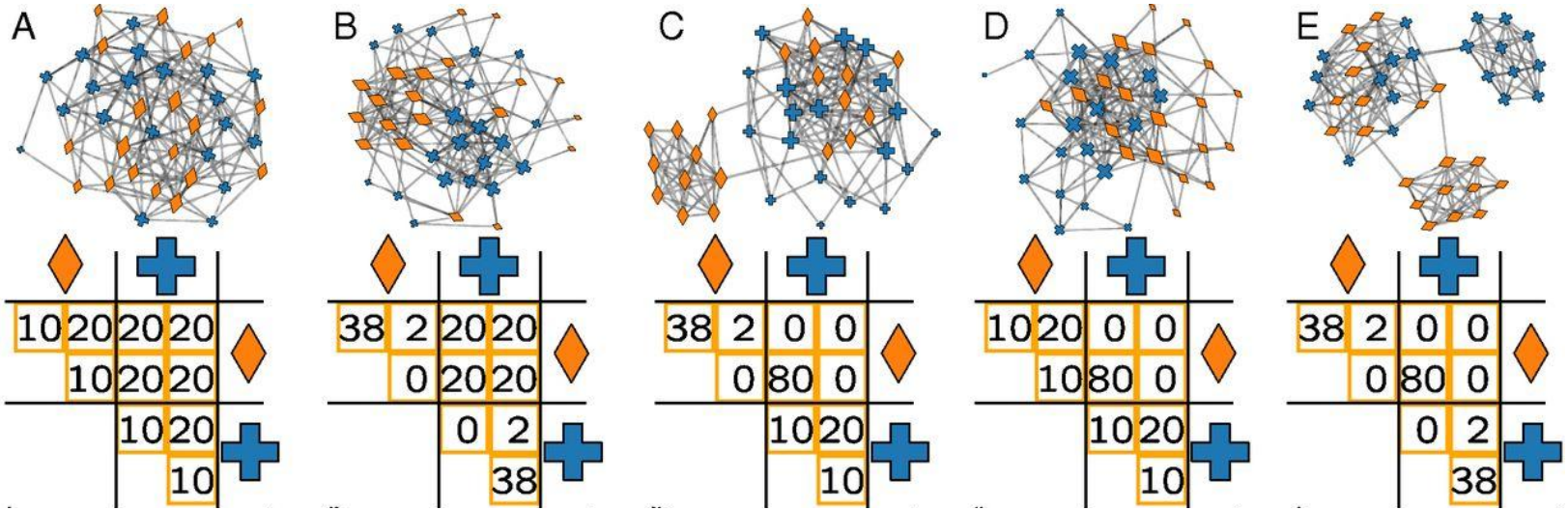


Local assortativity of gender in a sample of Facebook friendships (McAuley and Leskovec 2012).

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

Peel, Leto, et al. "Multiscale mixing patterns in networks." *PNAS* 115.16 (2018): 4057-4062.



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?)

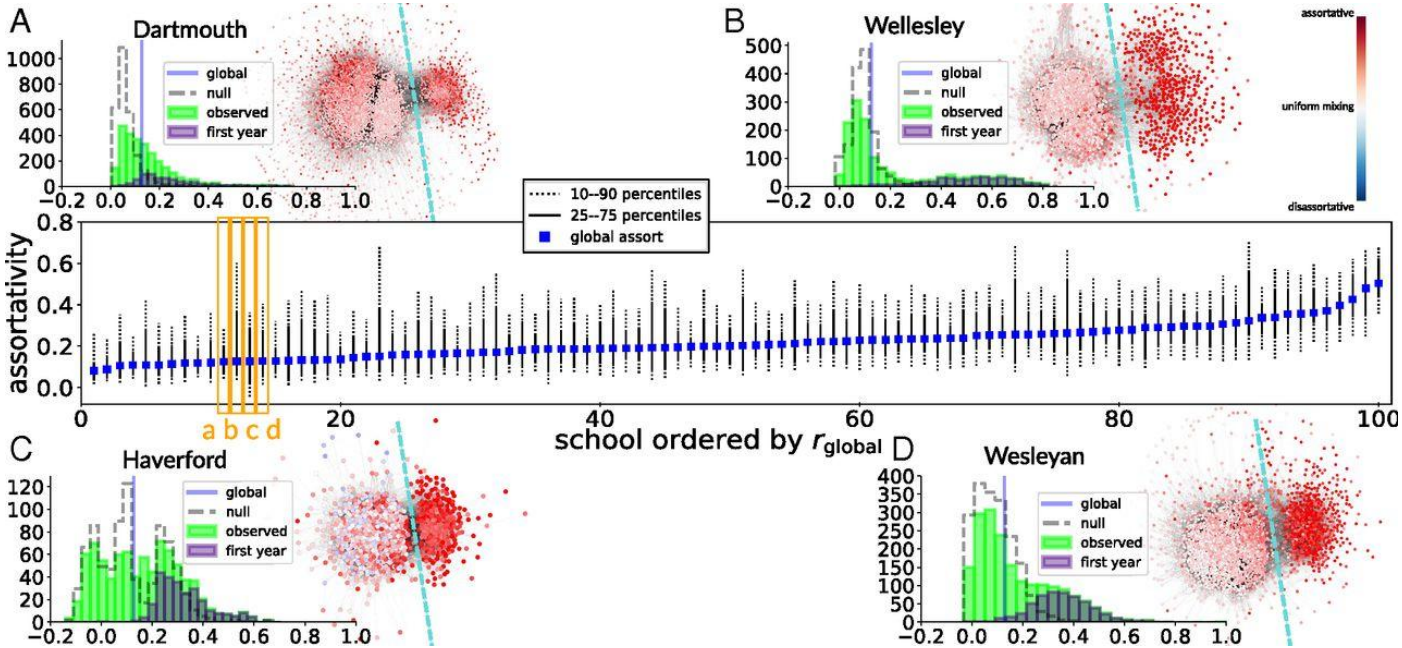
Better approximation:
Considering the stable state of a RW
(probability to reach a given node)
to weight the edges

Issue:
Need to fix the value of α
- α=0 the RW stays put,
- α=1 the RW never restarts

Solution:
Integrate over all possible α (multiscale approach)

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

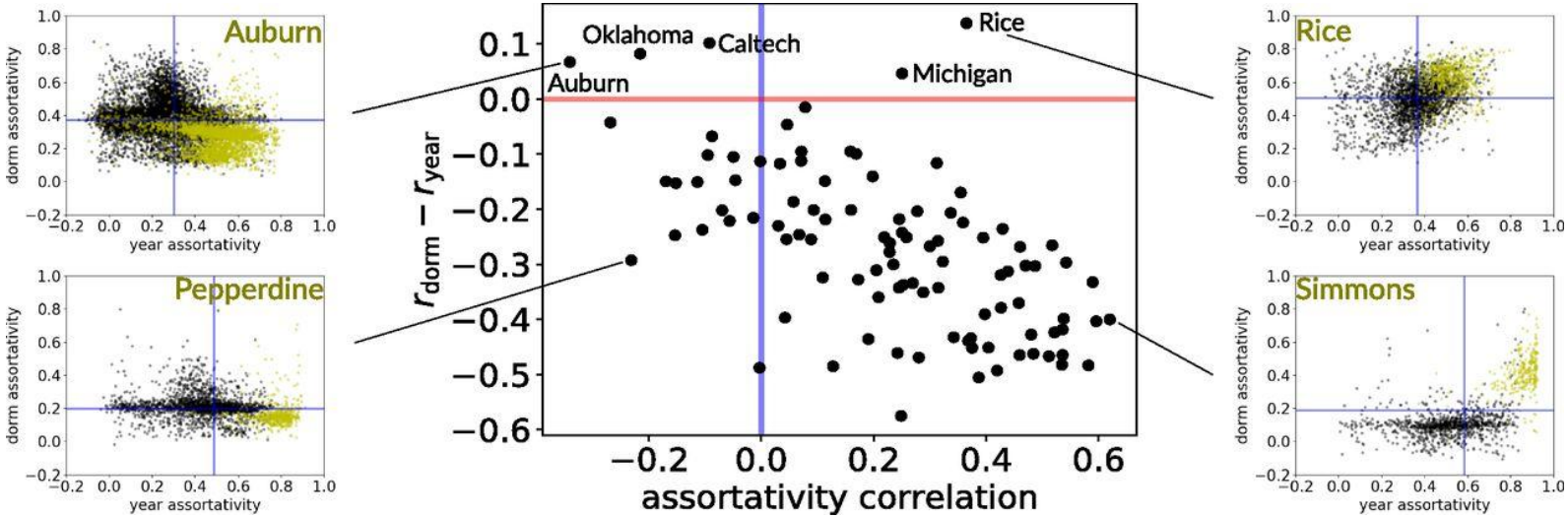
Evaluation on real data



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).

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