Homophily (or Assortativity)

Prof. Ralucca Gera,
Applied Mathematics Dept.
Naval Postgraduate School
Monterey, California
rgera@nps.edu

Learning Outcomes

✓Understand how to measure that nodes with similar characteristics tend to cluster,

- Based on enumerative characteristics (nationality)
- Based on scalar characteristics (age, grade)
- Based on degree.

✓Analyze network using hommophily by identifying the assortativity values based on various caracteristics.

✓Evaluate: consider why behind the what is the assortativity values of your network.

Are hubs adjacent to hubs?

- Real networks usually show a non-zero *degree correlation* (defined later compared to random).
  - If it has a positive degree correlation, the network has assortatively mixed degrees (assort. based other attributes can also be considered).
  - If it is negative, it is disassortative.
- According to Newman, social networks tend to be assortatively mixed, while other kinds of networks are generally disassortatively mixed.

Homophily or assortativity

Sociologists have observed network partitioning based on the following characteristics:

- Friendships, acquaintances, business relationships
- Relationships based on certain characteristics:
  - Age
  - Nationality
  - Language
  - Education
  - Income level

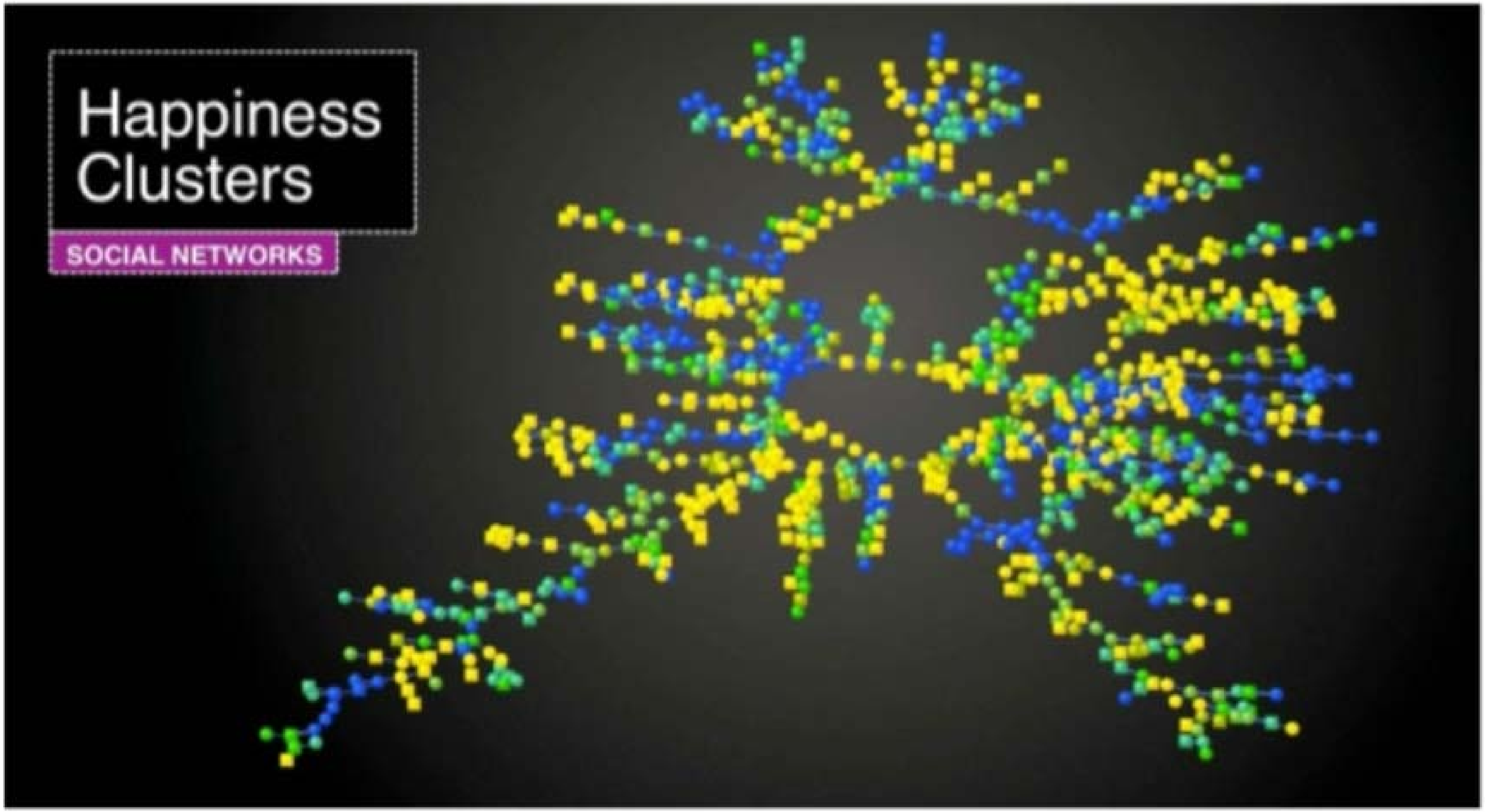
Homophily is the tendency of individuals to choose friends with similar characteristic. "Like links with like."

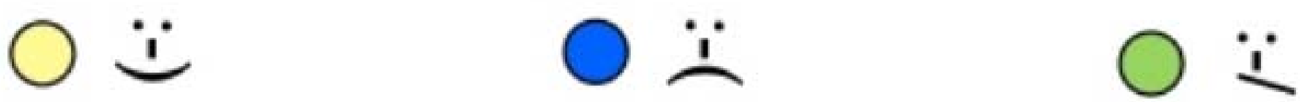
Homophily or assortativity

Homophily or assortativity is a common property of social networks (but not necessary):

- Papers in citation networks tend to cite papers in the same field
- Websites tend to point to websites in the same language
- Political views
- Race
- Obesity

Example of homophily

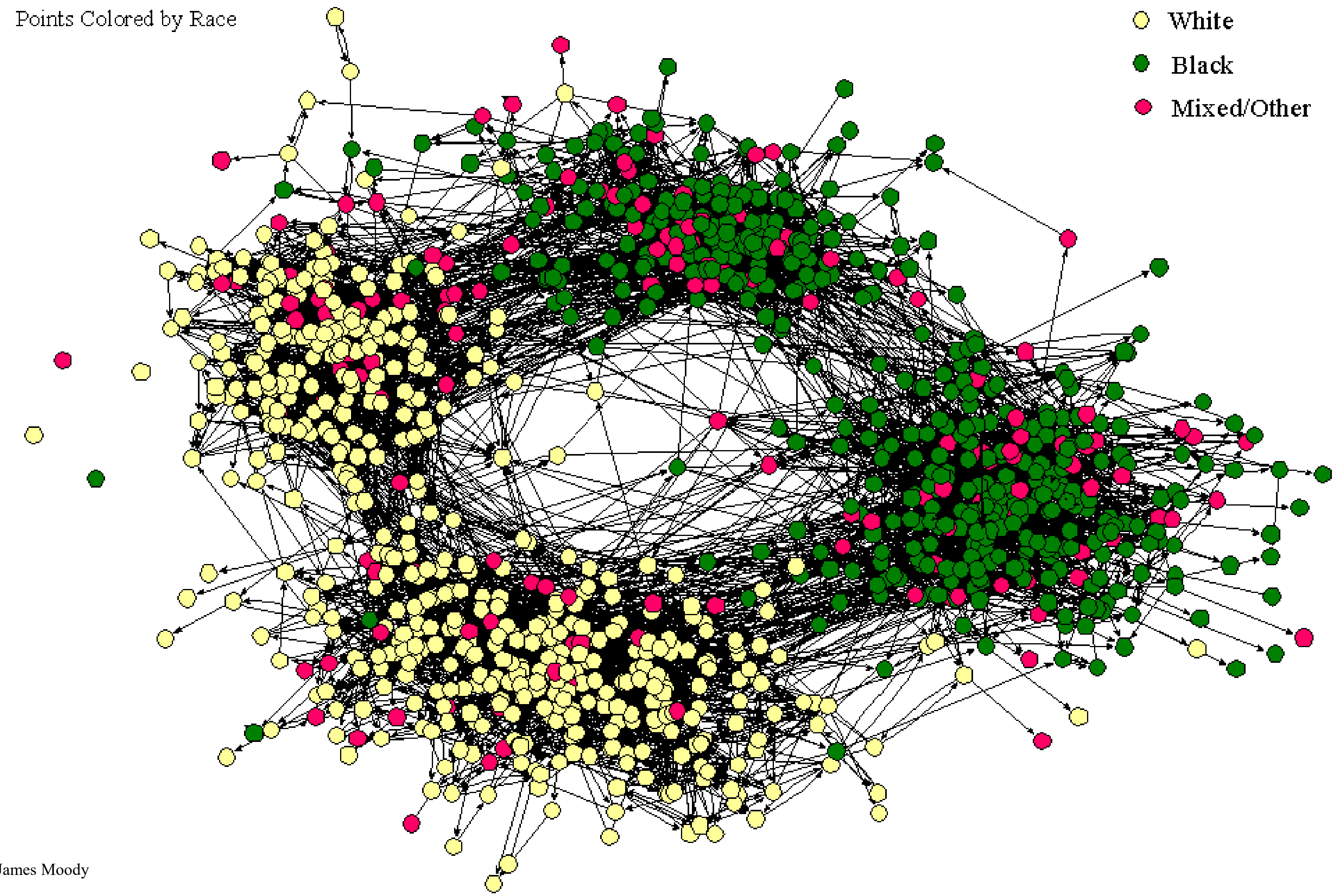




http://www.slideshare.net/NicolaBarbieri/homophily-and-influence-in-social-networks

Assortativity by race

The Social Structure of "Countryside" School District

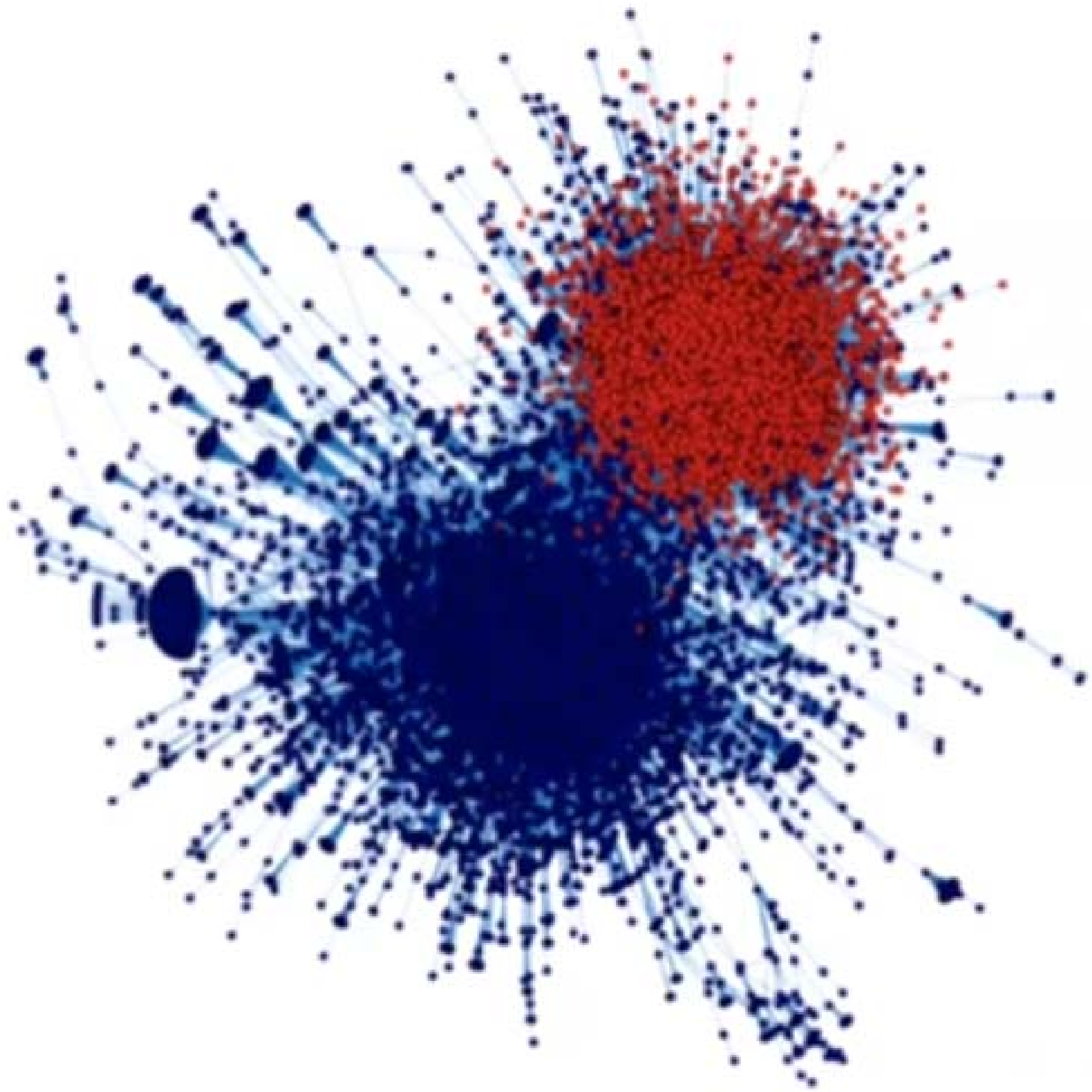


James Moody

Assortativity by political views

Titter data: political retweet network
Red = Republicans
Blue = Democrats

Note that they mostly
tweet and re-tweet
to each other



Conover et at., 2011

Disassortative

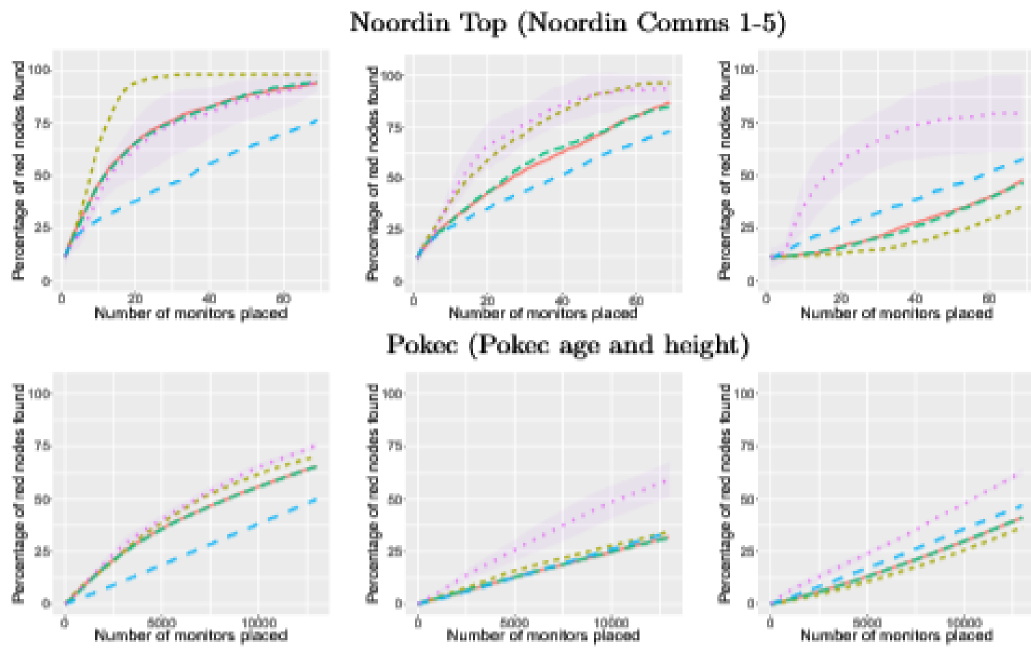
- Disassortative mixing: “like links with dislike”.
- Dissasortative networks are the ones in which adjacent nodes tend to be dissimilar:
  - – Dating network (females/males)
  - – Food web (predator/prey)
  - – Economic networks (producers/consumers)

Why?

NAVAL POSTGRADUATE SCHOOL

Why?

- Identifying people of interest could be easy if the network presents homophily

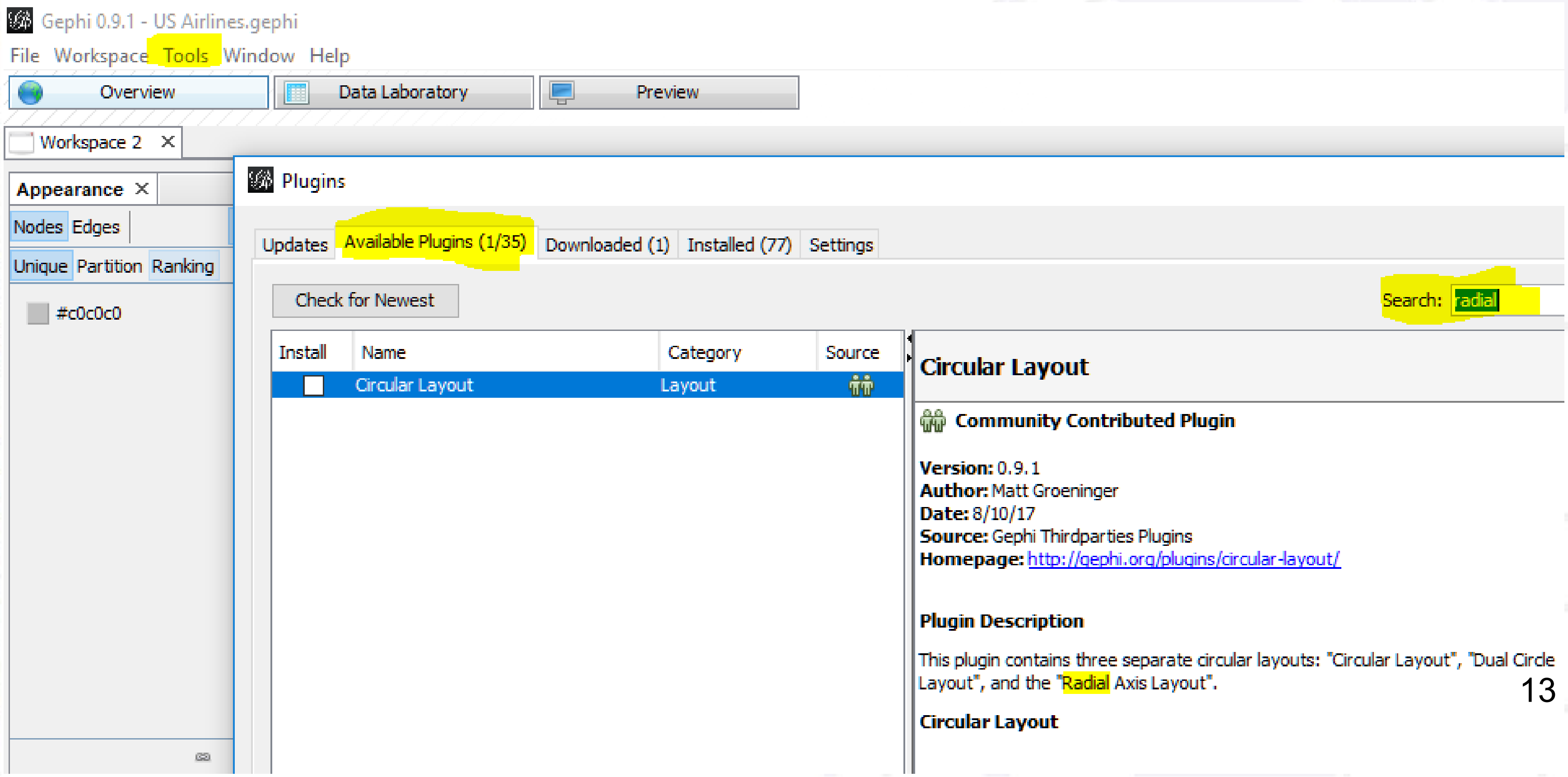


- When assortivity (homophily) is low lie Pokec, RedLearnRS (machine learning algorithm that depends on the count of POIs neighbors of nodes) outperforms all other strategies.
- When attributes show high homophily, RedLearnRS performs quite similar to the other algorithms.

How?

Gephi: Install Circular Layout

The Radial Axis Layout groups nodes and draws the groups in axes :
•Group nodes by degree, in degree, out degree, etc.
•Group nodes by attribute sort (based on data type of attribute).
•Draw axes/spars in ascending or descending order.
•Allows top, middle or bottom "knockdown" of axes/spars, along with ability to specify number of spars resulting after knockdown.

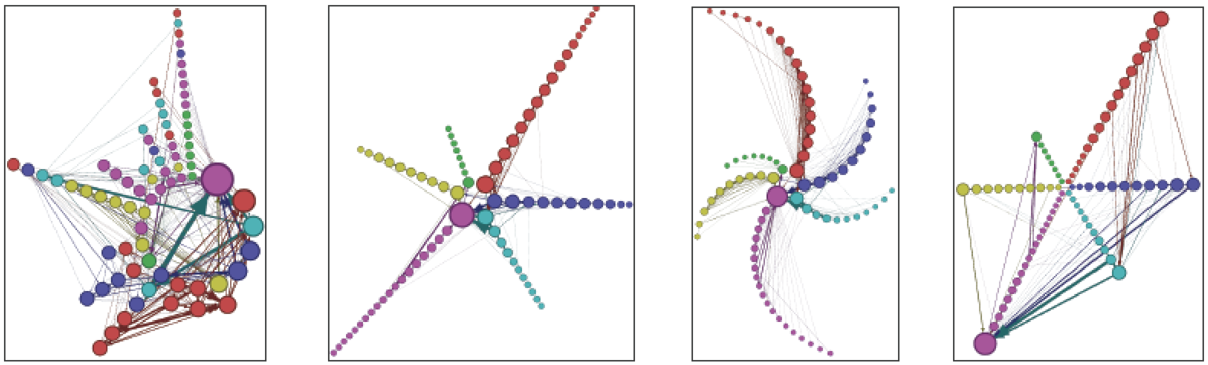


Homophily in Gephi

Run Radial Axis Layout [here](https://gephi.org/users/tutorial-layouts/)

Run the layout by applying the following settings step by step:

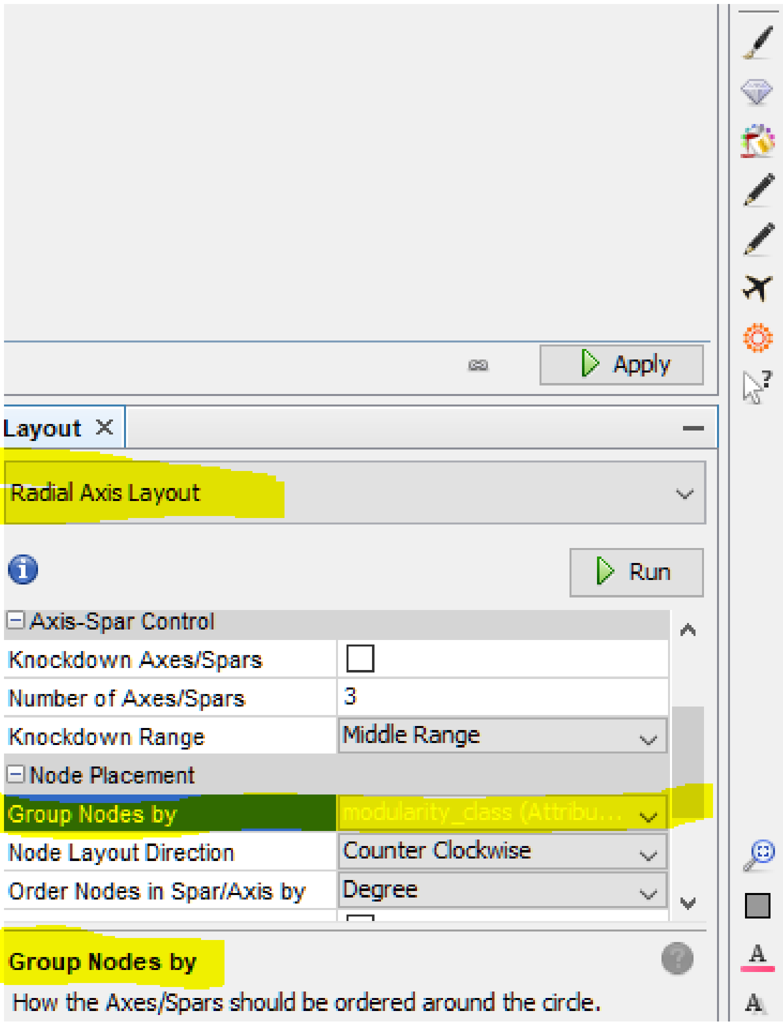
• Group nodes by = "Degree"	Run	Homophily by degree?
• Group nodes by = "Modularity Class" • Order nodes by = "Degree"	Run	Distribution of nodes by degree inside each community.
• Draw spar/axis as spiral = checked	Run	Better show links inside communities
• Draw spar/axis as spiral = unchecked • Ascending order = checked	Run	Better show links between communities



https://gephi.org/users/tutorial-layouts/

NAVAL POSTGRADUATE SCHOOL

An example: ordered by communities



Homophily in Python

- To check an attribute's assortativity:

assortivity_val=nx.attribute_assortativity_coefficient(G, "color")

The attribute "color" can be replaced by other attributes that your data was tagged with.

- If the attribute is "degree" then we obtain degree assortativity:

r = nx.degree_assortativity_coefficient(G)

- If the attribute is "communities" then we obtain modularity:

https://stackoverflow.com/questions/29897243/graph-modularity-in-python-networkx

What?

Assortative mixing (homophily)

We will study two types of assortative mixing:

1. Based on enumerative characteristics (the characteristics don’t fall in any particular order):
    1. Nationality
    2. Race
    3. Gender
    4. Communities
2. Based on scalar characteristics, such as:
    1. Age
    2. Income
    3. By degree: high degree connect to high degree

Based on enumerative characteristics (characteristics that don't fall in any particular order), such as:

- **Nationality**
- **Race**
- **Gender**
- **Or just communities**

Possible defn assortativity

A network is assortative if there is a significant fraction of edges between same-type vertices

- How to quantify the assortativity, r, of a network?

Method 1:

- Define c_i to be the class of vertex I, and tag the nodes to belong to each class c_i →

$$r_i = \frac{\#\ edges\ within\ C_i}{all\ possible\ edges} \rightarrow \quad r = \sum_i r_i$$

Then: What is the assortativity if we consider c_i =V(G) as the only class? Does it make sense?

Alternative definitions

Method 2 (used): compare the assortativity of the current network to the one of a random graph:

- Compute the fraction of edges in c_i in the given network,
- Compute the fraction of edges in c_i in a random graph,
- r is their difference.

This is the same process used for similarity, rather counting edges between nodes instead of neighbors of pairs of nodes.

The fraction of edges in c_i in the given network

- Let c_i be the class of vertex i.
- Let n_c be the total number of classes
- Let $\delta(c_i, c_j) = \begin{cases} 1, & if \;\; i = j \\ 0, & if \; i \neq j \end{cases}$ be the Kronecker δ that accounts for vertices in the same class.
- Then the number of edges of the same type is:

$$r = \sum_{ij \in E(G)} \delta\left(c_i, c_j\right) = \frac{1}{2} \sum_{ij} a_{ij}\, \delta(c_i, c_j)$$

Checks if vertices are in the same class

Checks for adjacent nodes

Compute the fraction of edges in c_i in the random network

- Construct a random graph with the same degree distrib.
- Let c_i be the class of vertex *i*
- Let n_c be the total number of classes

Checks if vertices are in the same class

- Let $\delta(c_i, c_j) = \begin{cases} 1, & if \;\; i = j \\ 0, & if \; i \neq j \end{cases}$ be the Kronecker δ
- Pick an arbitrary edge in the random graph:
  - pick a vertex $i \rightarrow$ there are $\deg i$ edges incident with it, so $\deg i$ choices for *i* to be the 1st end vertex of our arbitrary edge
  - and then there are $\deg j$ choices for j to be the other end-vertex.
- If $\mathrm{m} = |E(G)|$ edges are placed at random, the expected number of edges between *i* and *j* is $\frac{\deg i \; \deg j}{2m}$

Compute the fraction of edges in c_i in the random network

- Construct a random graph with the same degree distrib.
- Let c_i be the class of vertex i
- Let n_c be the total number of classes

Checks if vertices are in the same class

- Let $\delta(c_i, c_j) = \begin{cases} 1, & if \;\; i = j \\ 0, & if \; i \neq j \end{cases}$ be the Kronecker δ
- If $|E(G)|$ edges are placed at random, the expected number of edges between i and j is $\frac{\deg i \;\deg j}{2|E(G)|}$
- Then the number of edges between same class nodes is:

$$r = \frac{1}{2} \sum_{ij \in E(G)} \frac{\deg i \;\deg j}{2|E(G)|} \delta(c_i, c_j)$$

duplications: as you choose vertex j above, the edge ji will be counted after edge ij was counted

NAVAL POSTGRADUATE SCHOOL

Consider their difference

$$r = \frac{1}{2}\sum_{ij} A_{ij}\,\delta(c_i, c_j) - \frac{1}{2}\sum_{ij\in E(G)} \frac{\deg i \ \deg j}{2|E(G)|}\delta(c_i, c_j)$$

Checks if vertices are in the same class

$$r = \frac{1}{2}\sum_{ij}[A_{ij} - \frac{\deg i \deg j}{2|E(G)|}]\delta(c_i, c_j)$$

And now normalize by m = |E(G)|:

Modularity = $\frac{1}{2m}\sum_{ij}[A_{ij} - \frac{\deg i \deg j}{2m}] \cdot \delta(c_i, c_j)$

The same defn as the modularity in community detection, since it measures assortativity based on predefined communities.

Modularity

- $Q = \frac{1}{2|E(G)|}\sum_{ij}[A_{ij} - \frac{\deg i \deg j}{2|E(G)|}] \cdot \delta(c_i, c_j)$ Checks if vertices are in the same class
- Measure used to quantify the like vertices being connected to like vertices
- -1 < Q < 0 means there are fewer edges between like vertices in a class compared to a random network

  i.e. disassortative network
- 0 < Q < 1 means there are more edges between like vertices in a class compared to a random network i.e. assortative network
- Q = 0 means it behaves like a random network.

Enumerative characteristics

- Normalizing the modularity value Q, by the maximum value that it can get is realistic
  - Perfect mixing is when all edges fall between vertices of the same type

$\sum A_{ij}$

$Q_{max} \neq 1$

$$Q_{\max} = \frac{1}{2m}\left(2m - \sum_{ij} \frac{k_i k_j}{2m} \delta(c_i, c_j)\right)$$

- Then, the assortativity coefficient, $r = \frac{Q}{Q_{max}}$, is:

$$-1 \leq \frac{Q}{Q_{\max}} = \frac{\sum_{ij}(A_{ij} - k_i k_j / 2m)\delta(c_i, c_j)}{2m - \sum_{ij}(k_i k_j / 2m)\delta(c_i, c_j)} \leq 1$$

Based on scalar characteristics, such as:

- **Age**
- **Income**

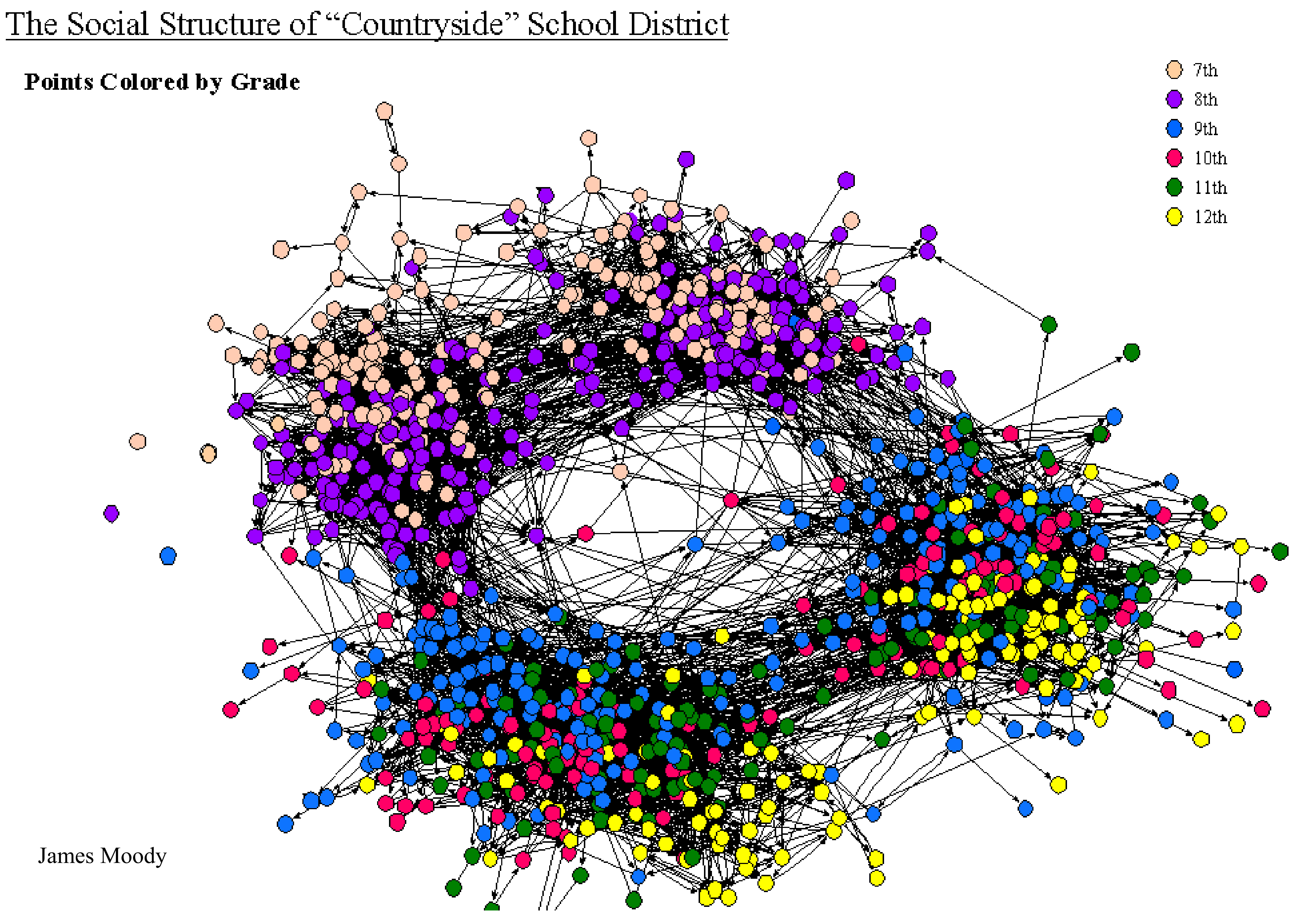
Scalar characteristics

- Scalar characteristics: enumerative characteristics taking numerical values, such as age, income
  - For example using age: two people are similar if:
    - they are born the same day or
    - within a year or within *x* years,
    - They are in the same class
    - Same generation

    different granularity based on the data and questions asked.
- If people are friends with others of the same age, we consider the network assortatively mixed by age (or stratified by age)

NAVAL POSTGRADUATE SCHOOL

Assortativity by grade/age

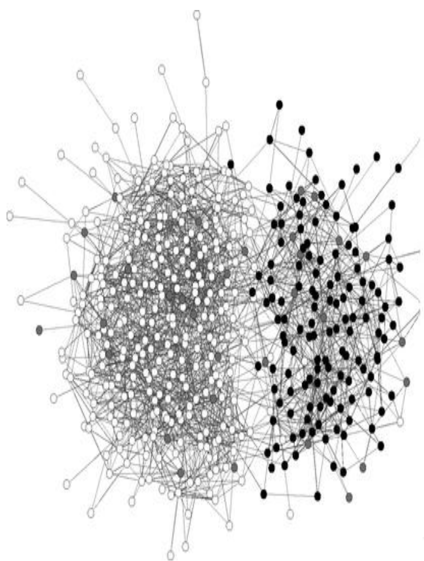


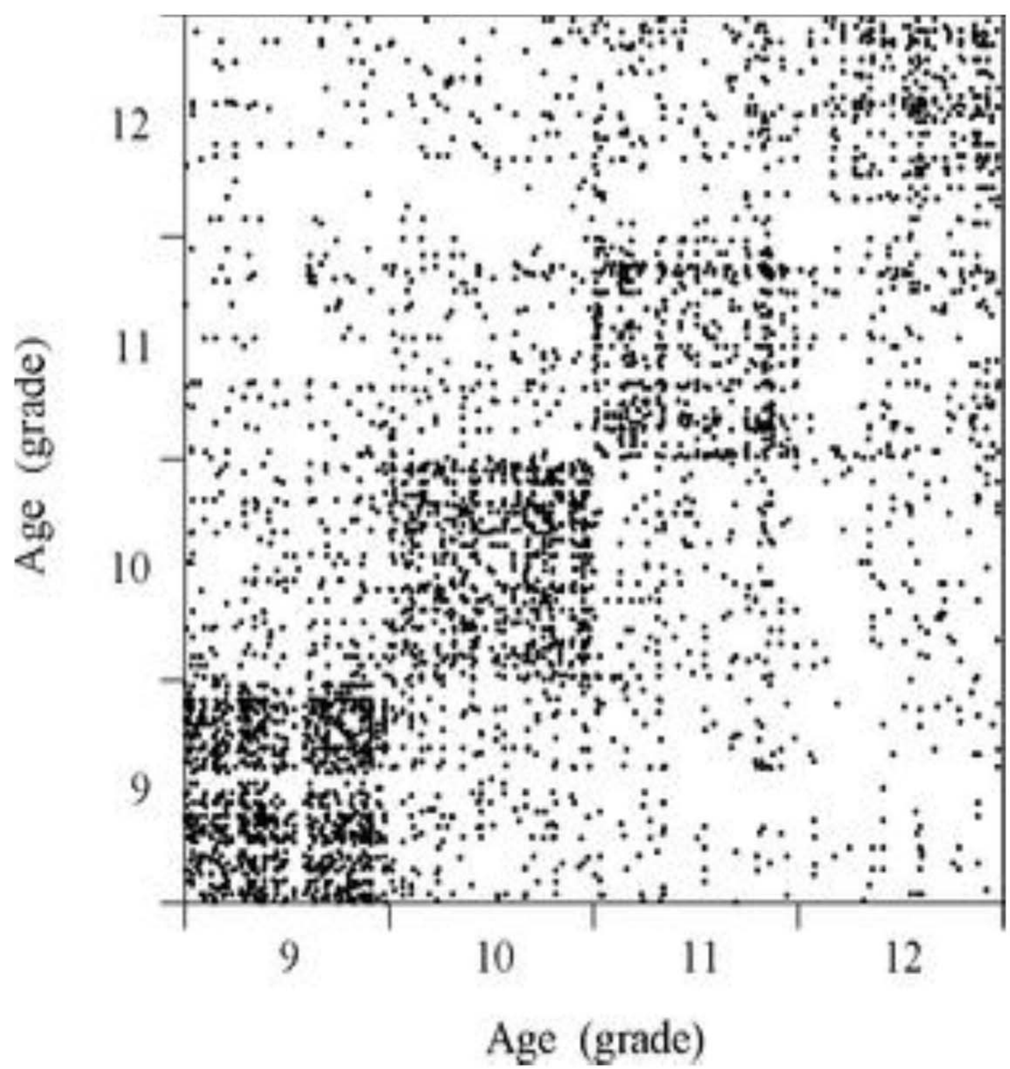
Scalar characteristics

- When we consider scalar characteristics we basically have an approximate notion of similarity between adjacent vertices (i.e. how far/close the values are)
  - There is no approximate similarity that can be measured this way when we talk about enumerative characteristics; rather present/absent

Assortativity matrix based on Scalar characteristics

Friendships at the same US high school: each dot represents a friendship (an edge from the network)





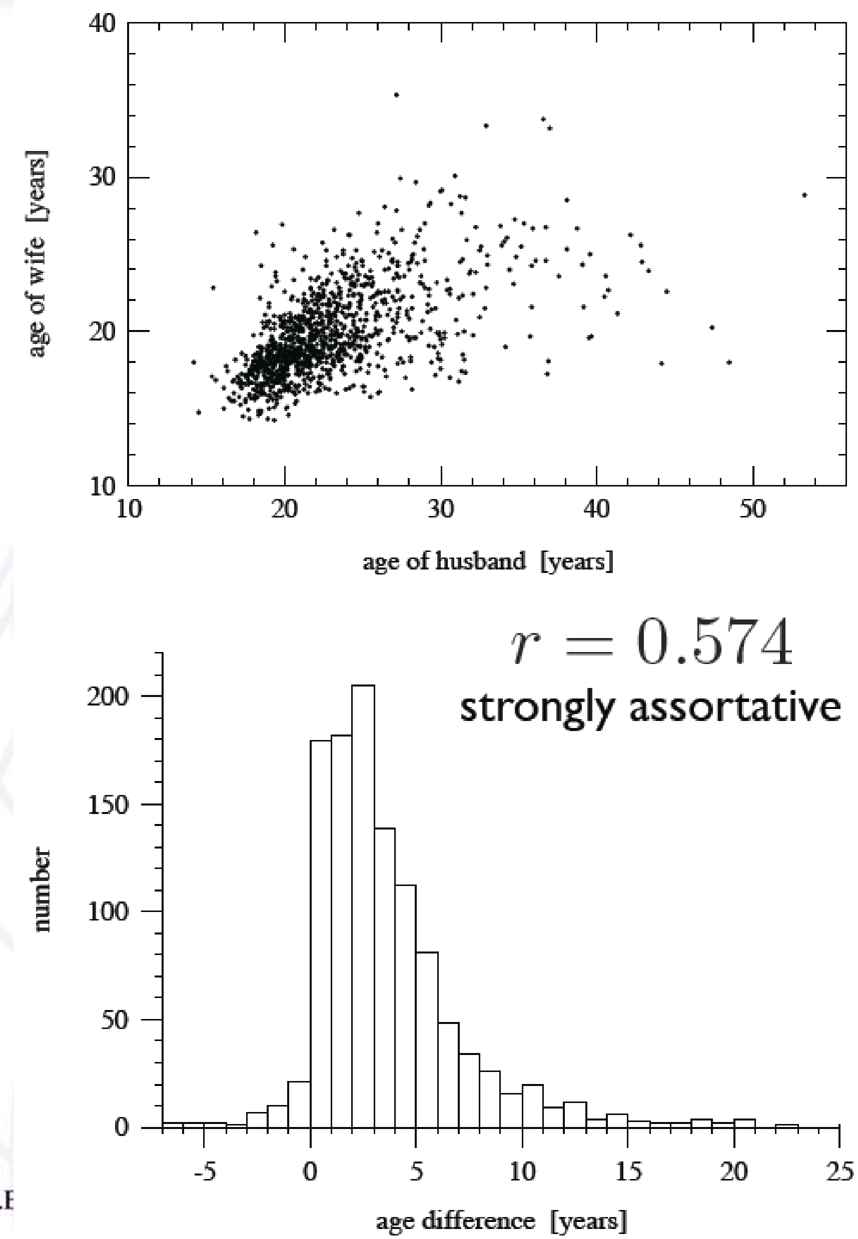
Denser along the y = x line (because of the way data is displayed)

Sparser as the difference in grades increases

Strongly assortative

Data: 1995 US National Survey of Family Growth

- Top figure:
  A scatter plot of 1141 married couples
- Bottom figure:
  The same data showing a histogram of the age difference



Scalar characteristics

- How do we measure scalar assortative mixing?
- Would the idea we use for the enumartive assortative mixing work?
- That is to place vertices in bins based on scalar values:
  - Treat vertices that fall in the same bin (such as age) as "like vertices" or "identical"
  - Apply modularity metric for enumerative characteristics

Scalar characteristics

Then the assortativity coefficient $r = \frac{Q}{Q_{max}}$ is defined again as:

$$r = \frac{\sum_{ij}(A_{ij} - k_i k_j / 2m)x_i x_j}{\sum_{ij}(k_i \delta_{ij} - k_i k_j / 2m)x_i x_j}$$

Similar to the enumerative one again

Either 0 or 1

r=1 → Perfectly assortative network

r=-1 → Perfectly disassortative network

r=0 → no correlation

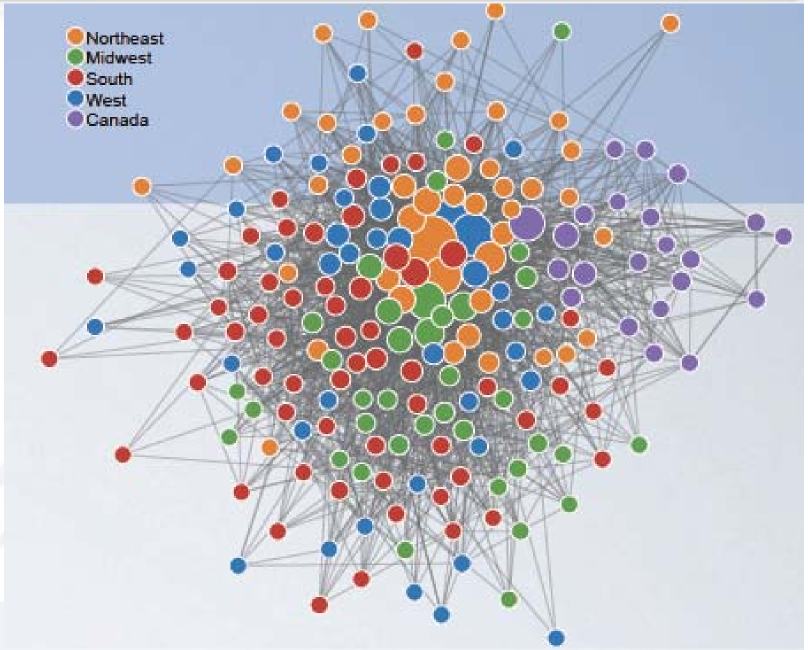
Same as Modularity or Pearson correlation coeff.

Computer Science faculty

88 Computer Science faculty:

- vertices are PhD granting institutions in North America
- Edge (i,j) means that PhD student at *i*, now faculty at *j*

labels are US census regions + Canada



	Northeast	Midwest	South	West	Canada	a_i
Northeast	**0.119**	0.053	0.074	0.055	0.022	0.322
Midwest	0.031	**0.067**	0.061	0.026	0.011	0.196
South	0.025	0.027	**0.083**	0.024	0.006	0.166
West	0.049	0.033	0.043	**0.073**	0.011	0.209
Canada	0.006	0.005	0.005	0.005	**0.085**	0.107
b_i	0.229	0.185	0.267	0.184	0.135	

$$r = 0.264$$

moderately assortative

By degree:
high degree nodes connect to high degree nodes

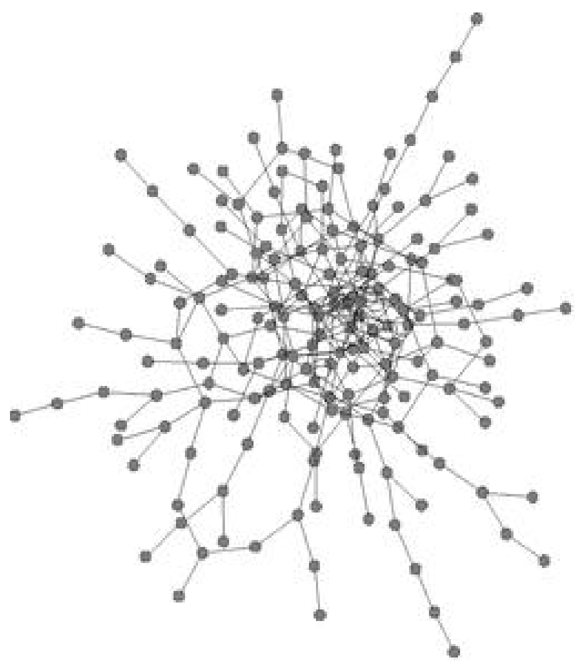
Assortative mixing by degree

A special case is when the characteristic of interest is the degree of the node

- Commonly used in social networks (the most used one of the scalar characteristics)
- More interesting since degree is a topological property of the network (not just a value like age or grade)
- This now reduces to Pearson Correlation Coefficient

Assortative mixing by degree

- Assortative network by degree → core of high degrees and a periphery of low degrees (Figure (a) below)
- Disassortative network by degree → uniform: low degree adjacent to high degree (Figure (b) and (c) below)



(a)

(b)

(c)

Newman's book (2003)

r = assortativity coefficient

	Network	Type	n	m	c	S	ℓ	α	C	Cws	r	Ref(s).
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980	3.48	2.3	0.20	0.78	0.208	16, 323
	Company directors	Undirected	7 673	55 392	14.44	0.876	4.60	–	0.59	0.88	0.276	88, 253
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822	7.57	–	0.15	0.34	0.120	89, 146
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838	6.19	–	0.45	0.56	0.363	234, 236
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918	4.92	–	0.088	0.60	0.127	234, 236
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16			2.1				9, 10
	Email messages	Directed	59 812	86 300	1.44	0.952	4.95	1.5/2.0		0.16		103
	Email address books	Directed	16 881	57 029	3.38	0.590	5.22	–	0.17	0.13	0.092	248
	Student dating	Undirected	573	477	1.66	0.503	16.01	–	0.005	0.001	−0.029	34
	Sexual contacts	Undirected	2 810					3.2				197, 198
Information	WWW nd.edu	Directed	269 504	1 497 135	5.55	1.000	11.27	2.1/2.4	0.11	0.29	−0.067	13, 28
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914	16.18	2.1/2.7				56
	Citation network	Directed	783 339	6 716 198	8.57			3.0/–				280
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977	4.87	–	0.13	0.15	0.157	184
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000		2.7		0.44		97, 116
Technological	Internet	Undirected	10 697	31 992	5.98	1.000	3.31	2.5	0.035	0.39	−0.189	66, 111
	Power grid	Undirected	4 941	6 594	2.67	1.000	18.99	–	0.10	0.080	−0.003	323
	Train routes	Undirected	587	19 603	66.79	1.000	2.16	–		0.69	−0.033	294
	Software packages	Directed	1 439	1 723	1.20	0.998	2.42	1.6/1.4	0.070	0.082	−0.016	239
	Software classes	Directed	1 376	2 213	1.61	1.000	5.40	–	0.033	0.012	−0.119	315
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000	11.05	3.0	0.010	0.030	−0.154	115
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805	4.28	2.1	0.012	0.011	−0.366	6, 282
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2	0.090	0.67	−0.240	166
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4	0.072	0.071	−0.156	164
	Marine food web	Directed	134	598	4.46	1.000	2.05	–	0.16	0.23	−0.263	160
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–	0.20	0.087	−0.326	209
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–	0.18	0.28	−0.226	323, 328

Newman, M. E. J. "MEJ Newman, SIAM Rev. 45, 167 (2003)." *SIAM Rev.* 45 (2003): 167.

Examples (published in 2003)

Same formula:

$$\text{deg}corr\_coeff = \frac{\sum_{ij}(A_{ij} - k_i k_j / 2m) k_i k_j}{\sum_{ij}(k_i \delta_{ij} - k_i k_j / 2m) k_i k_j}$$

	network	type	size n	degree assortativity r	error σ_r
social	physics coauthorship	undirected	52 909	0.363	0.002
	biology coauthorship	undirected	1 520 251	0.127	0.0004
	mathematics coauthorship	undirected	253 339	0.120	0.002
	film actor collaborations	undirected	449 913	0.208	0.0002
	company directors	undirected	7 673	0.276	0.004
	student relationships	undirected	573	−0.029	0.037
	email address books	directed	16 881	0.092	0.004
technological	power grid	undirected	4 941	−0.003	0.013
	Internet	undirected	10 697	−0.189	0.002
	World-Wide Web	directed	269 504	−0.067	0.0002
	software dependencies	directed	3 162	−0.016	0.020
biological	protein interactions	undirected	2 115	−0.156	0.010
	metabolic network	undirected	765	−0.240	0.007
	neural network	directed	307	−0.226	0.016
	marine food web	directed	134	−0.263	0.037
	freshwater food web	directed	92	−0.326	0.031

Newman, Phys. Rev. E. 67 , 0626126 (2003)

Range of the value r for real networks

Some statistics about real networks published in 2011

Network	N	z	ℓ	ℓ_1	ℓ_1^B	C	$\widetilde{C}$	r
Power grid	4941	2.67	18.99	8.61	7.85	0.08	0.10	0.0035
PGP network	10680	4.55	7.49	5.40	2.66	0.27	0.38	0.23
AS Internet	28311	4.00	3.88	3.67	2.56	0.21	0.0071	−0.20
RL Internet	190914	6.34	6.98	5.25	3.17	0.16	0.061	0.025
Coauthorships	39577	8.88	5.50	4.45	2.93	0.65	0.25	0.19
Airports 500	500	11.92	2.99	2.76	1.62	0.62	0.35	−0.278
Interacting proteins	4713	6.30	4.22	4.05	2.96	0.09	0.062	−0.136
C. Elegans metabolic	453	8.94	2.66	2.55	1.93	0.65	0.12	−0.226
C. Elegans neural	297	14.46	2.46	2.33	1.84	0.29	0.18	−0.163
Facebook Caltech	762	43.70	2.34	2.26	1.55	0.41	0.29	−0.066
Facebook Georgetown	9388	90.67	2.76	2.55	1.79	0.22	0.15	0.075
Facebook Oklahoma	17420	102.47	2.77	2.66	1.79	0.23	0.16	0.074
Facebook UNC	18158	84.46	2.80	2.68	1.87	0.20	0.12	7×10^{-5}

https://www.semanticscholar.org/paper/The-unreasonable-effectiveness-of-tree-based-theor-Melnik-Hackett/0ef76143b83257592c4155a648286ba7a0cff474

References

- Newman, M. E. J. "MEJ Newman, SIAM Rev. 45, 167 (2003)." *SIAM Rev*. 45 (2003): 167.
- Newman, Mark EJ. "Mixing patterns in networks." *Physical Review E* 67.2 (2003): 026126.

Extra slides

Simpler reformulation for Corr. Coeff.

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} = \frac{\mathrm{Tr}\,\mathbf{e} - \|\,\mathbf{e}^2\,\|}{1 - \|\,\mathbf{e}^2\,\|}$$ = .621 for the network below (strongly assort.)

e_{ij} is the fraction of edges in a network that connect a vertex of type i to one of type j:

$$\sum e_{ij} = 1, \qquad \sum e_{ij} = a_i, \qquad \sum e_{ij} = b_j,$$

		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	

TABLE I: The mixing matrix e_{ij} and the values of a_i and b_i for sexual partnerships in the study of Catania *et al.* [23]. After Morris [24].

Ref: Newman, Phys Rev E. 67, 026126 (2003)