Network Observational Methods and Quantitative Metrics: II

- Whitney topics
  - Community structure (some done already in Constraints - I)
  - The Zachary Karate club story
  - Degree correlation
  - Calculating degree correlation for simple regular structures like trees and grids

Clustering or Grouping Metrics

- Community structure
  - Seek to find tightly connected subgroups within a larger network
- Clustering coefficient
  - Measure the extent to which nodes link to each other in triangles
  - Are your friends friends?
  - Clusters are often called “modules” by network researchers and are also associated by them with function
- Assortativity and disassortativity (AKA degree correlation)
  - Do highly linked nodes (“hubs”) link to each other (assortative) or do they link with weakly linked nodes (disassortative)
- Average (shortest) path length (AKA geodesic)
  - How far apart are nodes
  - Max geodesic is called network diameter

Community-finding and Pearson Coefficient r

- Technological networks seem to have $r < 0$
- Social networks seem to have $r > 0$
- Newman and Park sought an explanation in community structure and clustering
- Their algorithm for finding communities looks like a flow algorithm
- Zachary used a flow algorithm to find the communities in the Karate Club

Summary Properties of Several Big Networks (Newman)

Network	Type	n	m	z	l	α	$C^{(1)}$	$C^{(2)}$	r
SOCIAL									
Film actors	undirected	449913	25516482	113.43	3.48	2.3	0.20	0.78	0.208
Company directors	undirected	7673	55392	14.44	4.60	-	0.59	0.88	0.276
Math coauthorship	undirected	253339	496489	3.92	7.57	-	0.15	0.34	0.120
Physics coauthorship	undirected	52909	245300	9.27	6.19	-	0.45	0.56	0.363
Biology coauthorship	undirected	1520251	11803064	15.53	4.92	-	0.088	0.60	0.127
Telephone call graph	undirected	47000000	80000000	3.16		2.1			
E-mail messages	directed	59912	86300	1.44	4.95	1.5/2.0		0.16	
E-mail address books	directed	16881	57029	3.38	5.22	-	0.17	0.13	0.092
Student relationships	undirected	573	477	1.66	16.01	-	0.005	0.001	-0.029
Sexual contacts	undirected	2810				3.2			
INFORMATION									
WWW nd.edu	directed	269504	1497135	5.55	11.27	2.1/2.4	0.11	0.29	-0.067
WWW Altavista	directed	203549046	2130000000	10.46	16.18	2.1/2.7			
Citation network	directed	783339	6716198	8.57		3.0/-			
Roget's Thesaurus	directed	1022	5103	4.99	4.87	-	0.13	0.15	0.157
Word co-occurrence	undirected	460902	17000000	70.13		2.7		0.44	
TECHNOLOGICAL									
Internet	undirected	10697	31992	5.98	3.31	2.5	0.035	0.39	-0.189
Power grid	undirected	4941	6594	2.67	18.99	-	0.10	0.080	-0.003
Train routes	undirected	587	19603	66.79	2.16	-		0.69	-0.033
Software packages	directed	1439	1723	1.20	2.42	1.6/1.4	0.070	0.082	-0.016
Software classes	directed	1377	2213	1.61	1.51	-	0.033	0.012	-0.119
Electronic circuits	undirected	24097	53248	4.34	11.05	3.0	0.010	0.030	-0.154
Peer-to-peer network	undirected	880	1296	1.47	4.28	2.1	0.012	0.011	-0.366
BIOLOGICAL									
Metabolic network	undirected	765	3686	9.64	2.56	2.2	0.090	0.67	-0.240
Protein interactions	undirected	2115	2240	2.12	6.80	2.4	0.072	0.071	-0.156
Marine food web	directed	135	598	4.43	2.05	-	0.16	0.23	-0.263
Freshwater food web	directed	92	997	10.84	1.90	-	0.20	0.087	-0.326
Neural network	directed	307	2359	7.68	3.97	-	0.18	0.28	-0.226

Figure by MIT OCW.

Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex-vertex distance l; exponent α of degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$; clustering coefficient $C^{(2)}$; and degree correlation coefficient r. Blank entries indicate unavailable data.

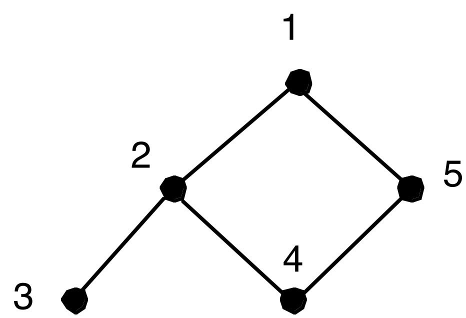
Calculating r

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

#	x	y
1	2	3
1	2	2
2	3	1
2	3	2
2	3	2
3	1	3
4	2	3
4	2	3
5	2	2
5	2	2

$\bar{x} = 2$

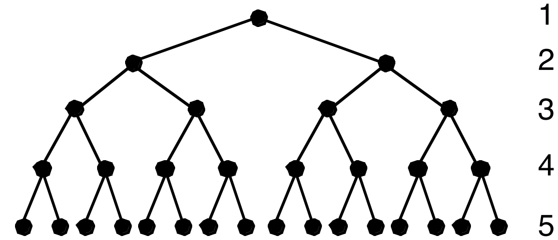
$\bar{y} = 2$



r = -0.676752968 using Pearson function in Excel

Note: if all nodes have the same k then r = 0/0

Census of Pairs for Pure Binary Tree



n=1

2 rows like this - ignore

n=2

6 rows like this - ignore

n=3

All other rows like this except last two sets

n=4

2^{n-2} rows of 3-3

$2*2^{n-2}$ rows of 3-1

3-3 means $(3-\bar{x})^2$

3-1 = 1-3 and means $(3-\bar{x})(1-\bar{x})$

n=5

2^{n-1} rows of 3-1

binary tree with n=5	
2	3
2	3
3	2
3	3
3	3
3	2
3	3
3	3
3	3
3	3
3	3
3	3
3	3
3	3
3	3
3	3
3	3
3	3
3	3
3	3
3	3
3	1
3	1
3	3
3	1
3	1
3	3
3	1
3	1
3	3
3	1
3	1
3	3
3	1
3	1
3	3
3	1
3	1
3	3
3	1
3	1
3	3
3	1
3	1
1	3
1	3
1	3
1	3
1	3
1	3
1	3
1	3
1	3
1	3
1	3
1	3
1	3
1	3
1	3
1	3

Result of Census

Sum of row entries = $\sum k_i^{\,2} = 10 * 2^{n-1} - 14$

Total number of rows = $\sum k_i = 2^{n+1} + 4$ = ksum

$$\therefore \quad \bar{x} = \frac{\sum k^2}{\sum k} = 2.5 \text{ in the limit of large } n$$

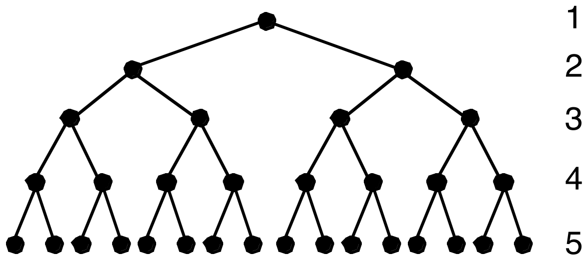
<k>=2

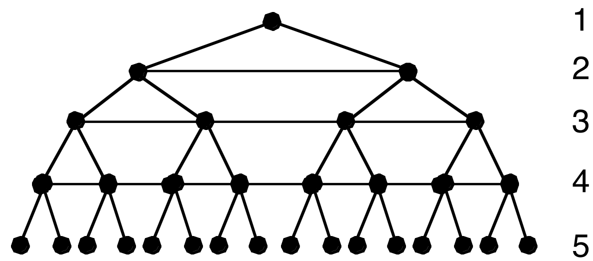
Total 2^n rows of 3-1

Approx (ksum - 2^n) rows of 3-3

r = - 0.4122

Closed Form Results

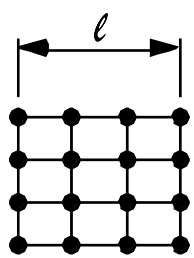




Property	Pure Binary Tree	Binary Tree with Cross-linking
$ksum$	$2^{n+1}-4$	$3*2^{n}-10$
$ksqsum$	$10*2^{n-1}-14$	$13*2^{n}-64$
$\bar{x}$	$\rightarrow 2.5$ as n becomes large (>~ 6)	$\rightarrow \frac{13}{3}$ as n becomes large (>~ 6)
Pearson numerator	$\sim 2^{n}(3-\bar{x})(1-\bar{x})+(ksum-2^{n})(3-\bar{x})^{2}$	$\sim 2^{n}(5-\bar{x})(1-\bar{x})+(ksum-2^{n})(5-\bar{x})^{2}$
Pearson denominator	$\sim 2^{n-1}(1-\bar{x})^{2}+(ksum-2^{n-1})(3-\bar{x})^{2}$	$\sim 2^{n-1}(1-\bar{x})^{2}+(ksum-2^{n-1})(5-\bar{x})^{2}$
r	$\rightarrow -\frac{1}{3}$ as n becomes large	$\rightarrow -\frac{1}{5}$ as n becomes large

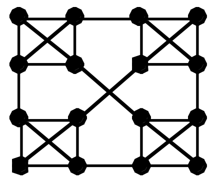
Note: Western Power Grid r = 0.0035

Bounded grid

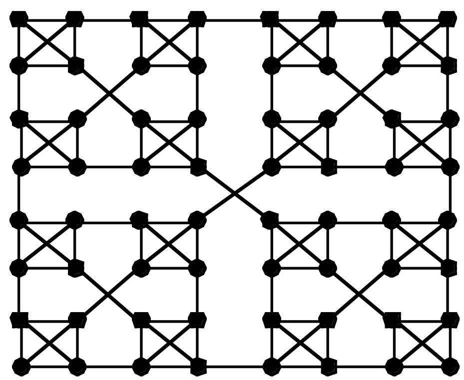


$$r=\frac{16(2-\bar{x})(3-\bar{x})+8(\ell-3)(3-\bar{x})^{2}}{2(2-\bar{x})^{2}+12(\ell-2)(3-\bar{x})^{2}} \rightarrow \frac{2}{3}$$

Nested Self-Similar Networks



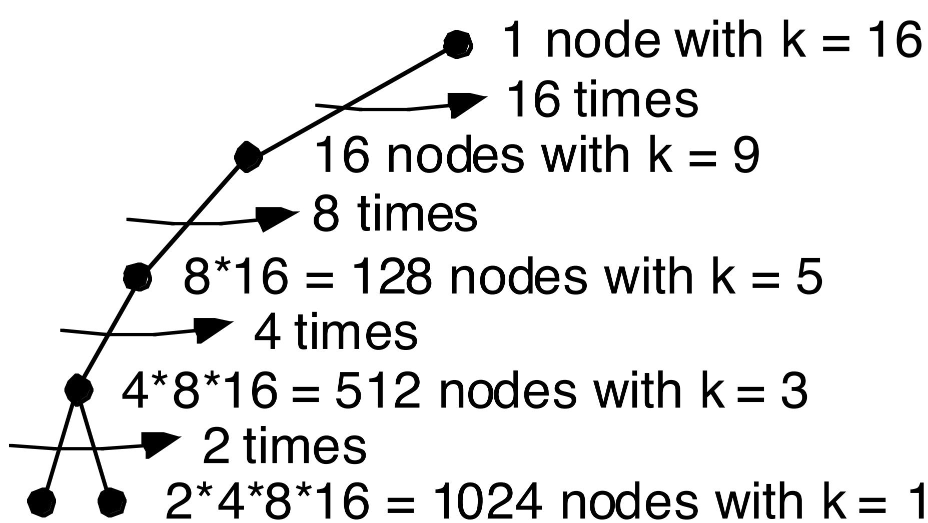
r = - 0.25, c = 0.625



r = - 0.0925, c = 0.5500

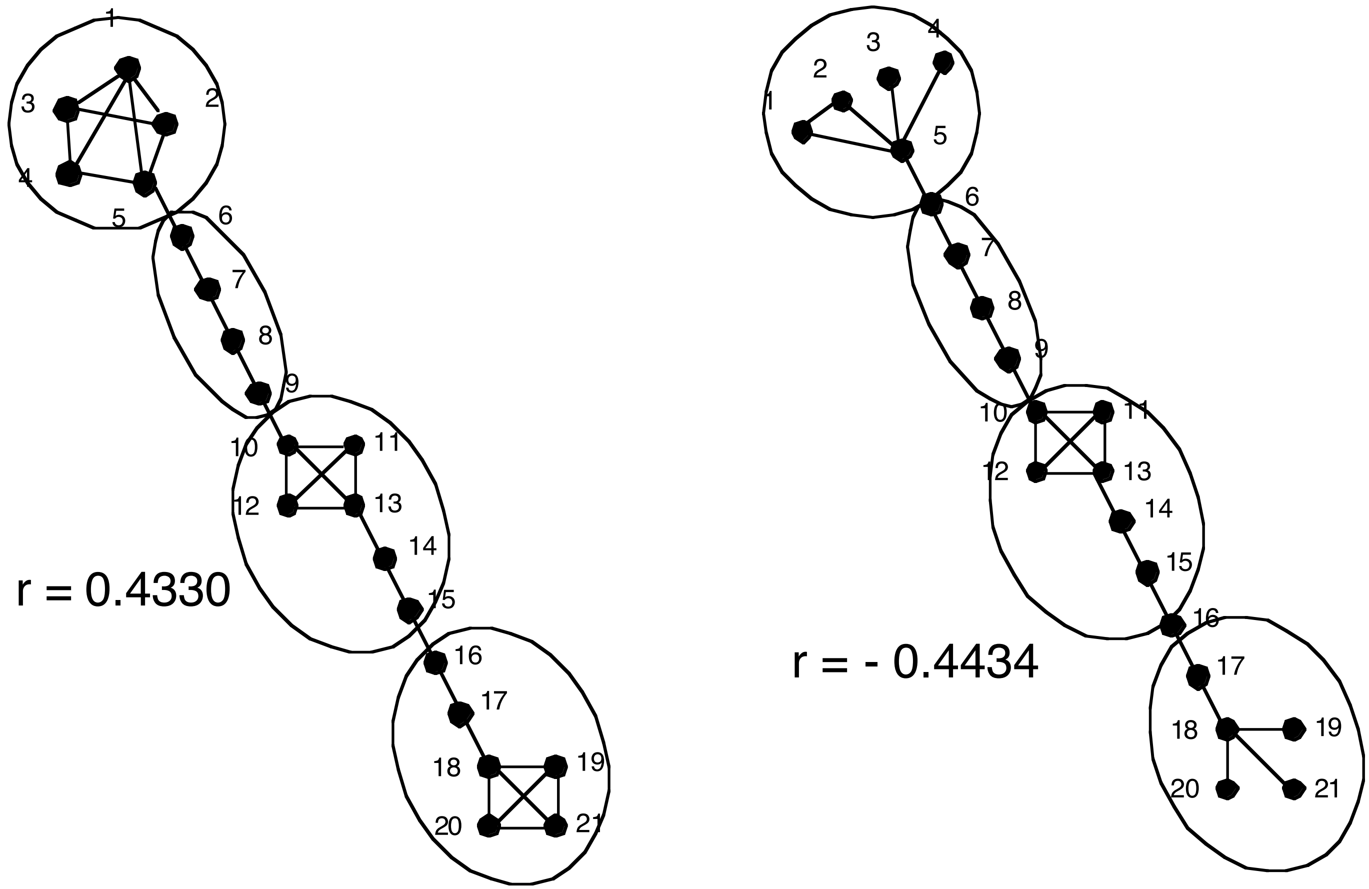
Probably, r = 0 in the limit as the network grows

Tree with Diminishing Branching Ratio



r = 0.38166

Toy Networks with Positive and Negative r



Toward Matlab for Pearson (symmetric)

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

Look at numerator, ignore xbar for the moment

$$\sum\left(x_i y_j\right) = x_i^{'} \delta_{ij} y_j = x^{'} A x$$

$$\delta_{ij} = 1 \;\; if \;\; i \text{ links to } j$$

$$\delta_{ij} = 0 \;\; if \;\; i \text{ does not link to } j$$

Essentially the calculation is a quadratic form.
My bias: control theory, where quadratic forms are common

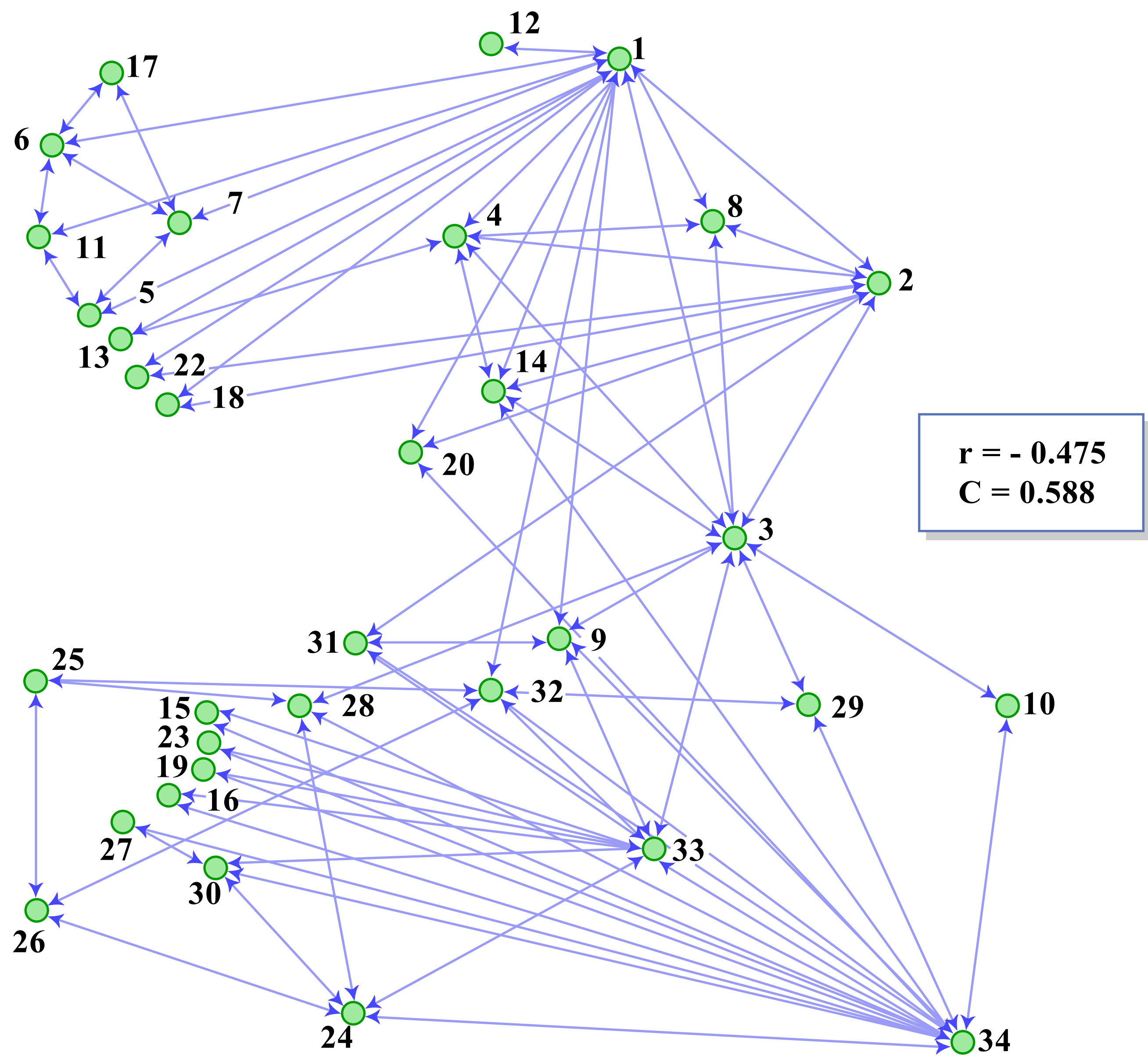
Matlab Implementation

```
function prs = pearson(A)
%calculates pearson degree correlation of A
[rows,colms]=size(A);
won=ones(rows,1);
k=won'*A;
ksum=won'*k';
ksqsum=k*k';
xbar=ksqsum/ksum;
num=(k-won'*xbar)*A*(k'-xbar*won);
kkk=(k'-xbar*won).*(k'.^.5);
denom=kkk'*kkk;
prs=num/denom;
```


Newman-Girvan Algorithm

- Seeks edges along which a lot of traffic flows between nodes, revealed by high edge betweenness
  - Edge betweenness rises with number of shortest paths between all node pairs that pass along that edge
- Removing this edge and repeating the process reveals clusters that roughly conform to Modularity 1 (?)

Zachary's Karate Club: A Social Network with $r < 0$ (from UCINET)



There is no link between 23 and 34. Every later scholar has this error.

Figure by MIT OCW.

Zachary's Karate Club - Most Studied by Community-Finding Researchers

- Zachary studied a karate club that had an internal fight and split into two
- Based on data he took about relationships between club members, he "predicted" how the group would split
- His algorithm correctly assigned all but one person to the groups they actually joined after the split
- "An Information Flow Model for Conflict and Fission in Small Groups," *J Anth Res* v 33, 1977, pp 452-473

The Reason for the Split

- The karate instructor "Mr Hi" wanted more money
- The club president "John A" felt the club administrators should set his salary
- Many angry club meetings occurred over this conflict
- When John A fired Mr Hi, the group split
- Half formed a new club around Mr Hi
- The other half found another instructor or gave up karate

The Dynamics

- Different club members took different sides
- Club meetings (different from karate lessons) were fights based on votes, and the faction with the most votes prevailed at any given meeting
- "Political" activity occurred outside the club as the sides' activists recruited others to attend meetings and vote their way

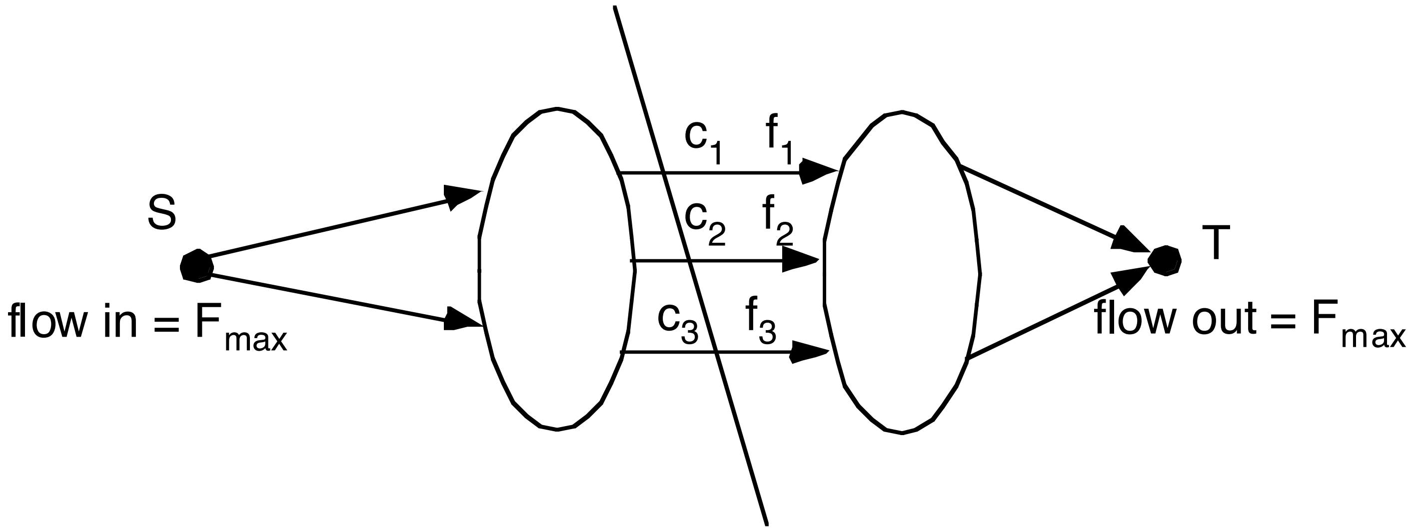
Zachary's Model

- Nodes are club members, plus Mr Hi
- Mr Hi is node #1, John A is node #34
- There is a line between two nodes if those people meet in some venue outside of the club
  - Venues include local campus pub, Mr Hi's private karate school, common classes, outside karate tournaments, etc
- Each edge has a weight = the number of outside venues that the two people have in common
  - Based on the idea that communication, including recruiting people to come to club meetings, happens in the outside venues, and that more venues in common means stronger communication, represented by stronger edge weight

Zachary's Algorithm

- Zachary assumed that each side tried to recruit its adherents and keep the other side from learning about a meeting
- So communication flow was important, and the group would likely split at "chokepoints" of communication between the groups
- He adopted the Ford-Fulkerson capacitated flow algorithm - max flow/min cut - from "source" Mr Hi to "sink" John A: the cut closest to Mr Hi that cuts saturated edges divides the network into the two factions
- He correctly predicted every member's decision except #9
- His algorithm depended on knowing "who was who" and "what was what"

Max-flow Min-Cut Theorem



The cut divides the network in two

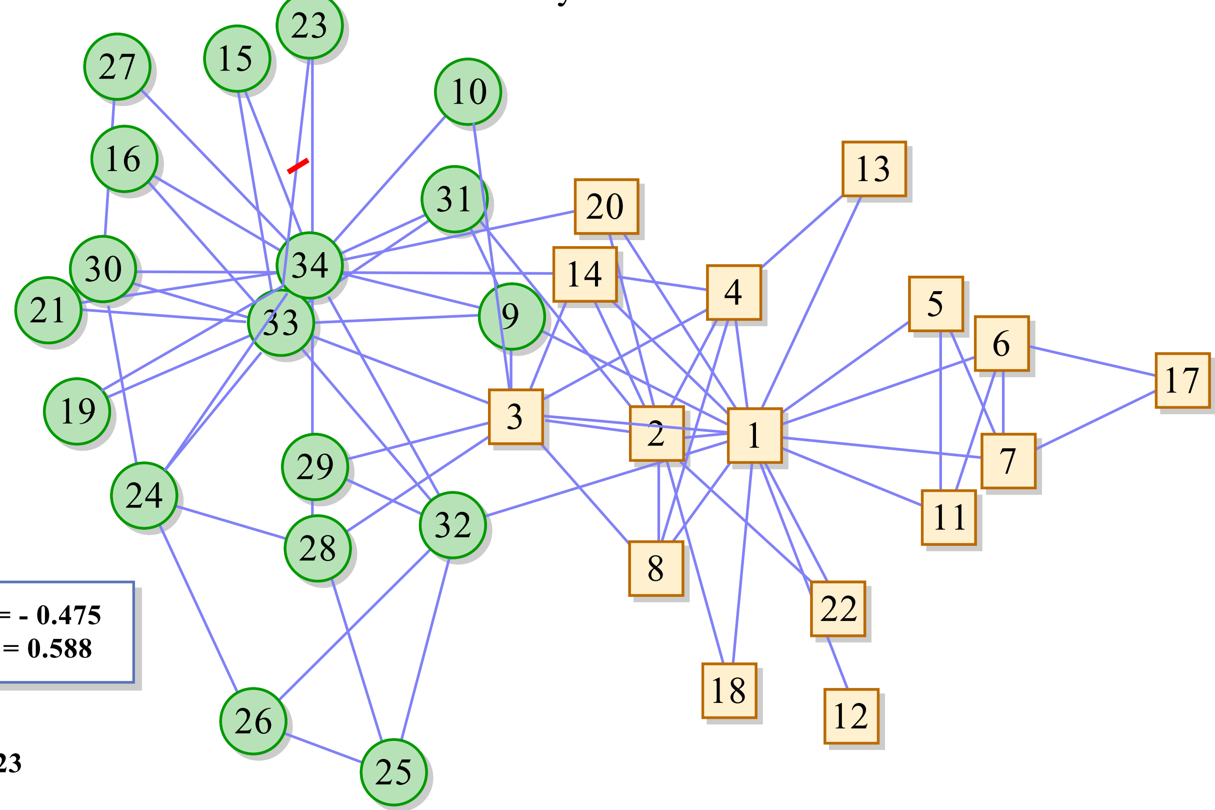
Its capacity = $c_1+c_2+c_3$

Its flow = $f_1+f_2+f_3$

"There is a cut such that $c_1+c_2+c_3 = F_{max}$ "
No other cut can have less capacity
or else the total flow will be less than F_{max}
Other cuts can have more capacity
but that makes no difference.

The Relationship Graph

This version of the Karate club
appears in several papers
by Newman or Girvan and Newman



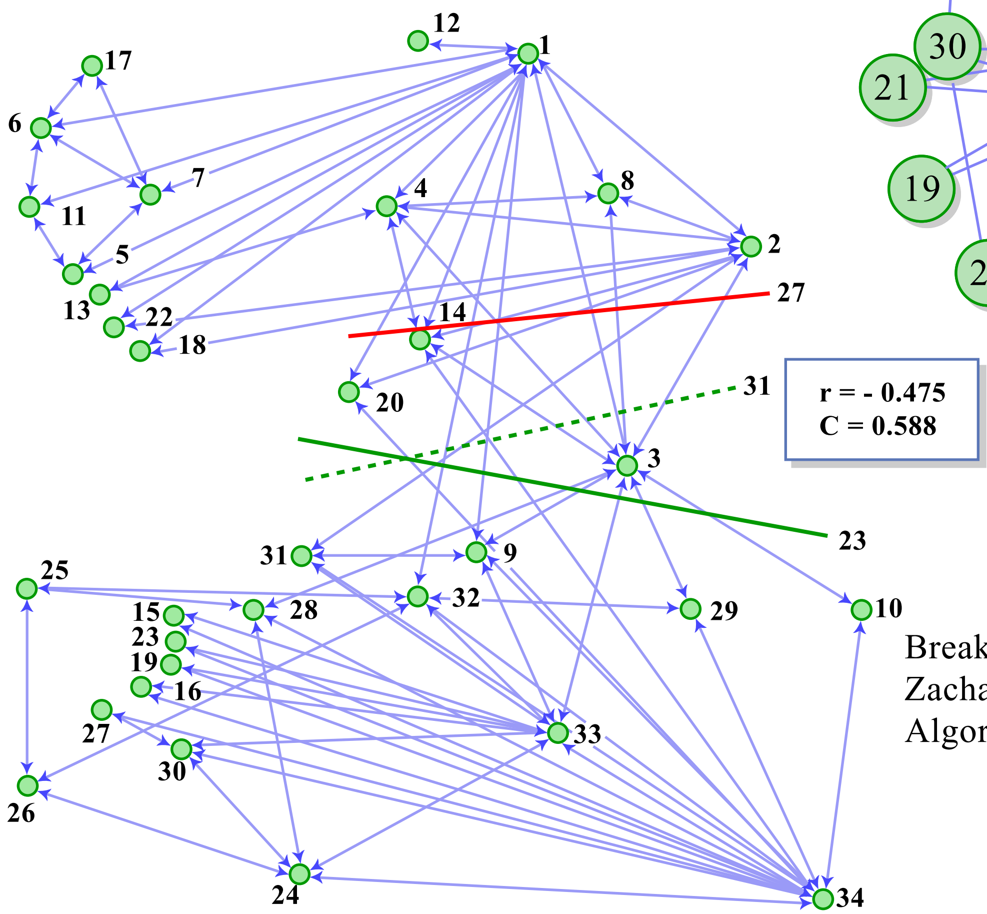


Figure by MIT OCW.

Breakdown of the Karate Club according to Newman. Every algorithm from Zachary to Newman puts 9 in 34's group - but 9 actually went with 1, not 34. Algorithms have trouble with #3, not #9.

Figure by MIT OCW.

Different Approaches and Lessons

- Zachary's method depends on knowing facts about both nodes and edges, and uses a weighted graph
- Edges describe relationships outside the club
- #9 chose Mr Hi's group for an inside reason, something no one else did
- Later scholars used no info about nodes and edges and used an unweighted graph, but get the same answer and make the same mistake with #9
- Newman uses geodesics between all pairs of nodes while Zachary uses only paths between 1 and 34.
- How come later scholars get the same answer?

Possible Explanation

- The uncommitted members were the only bridges between two committed groups
- There were only a few such people and they shared few venues with members of both factions
- Thus the break can practically be seen on the unweighted graph with the naked eye
- So possibly later scholars have simply been lucky
- The goal of abstraction is to learn as much as you can while knowing *a priori* as little as possible

Network Comparisons

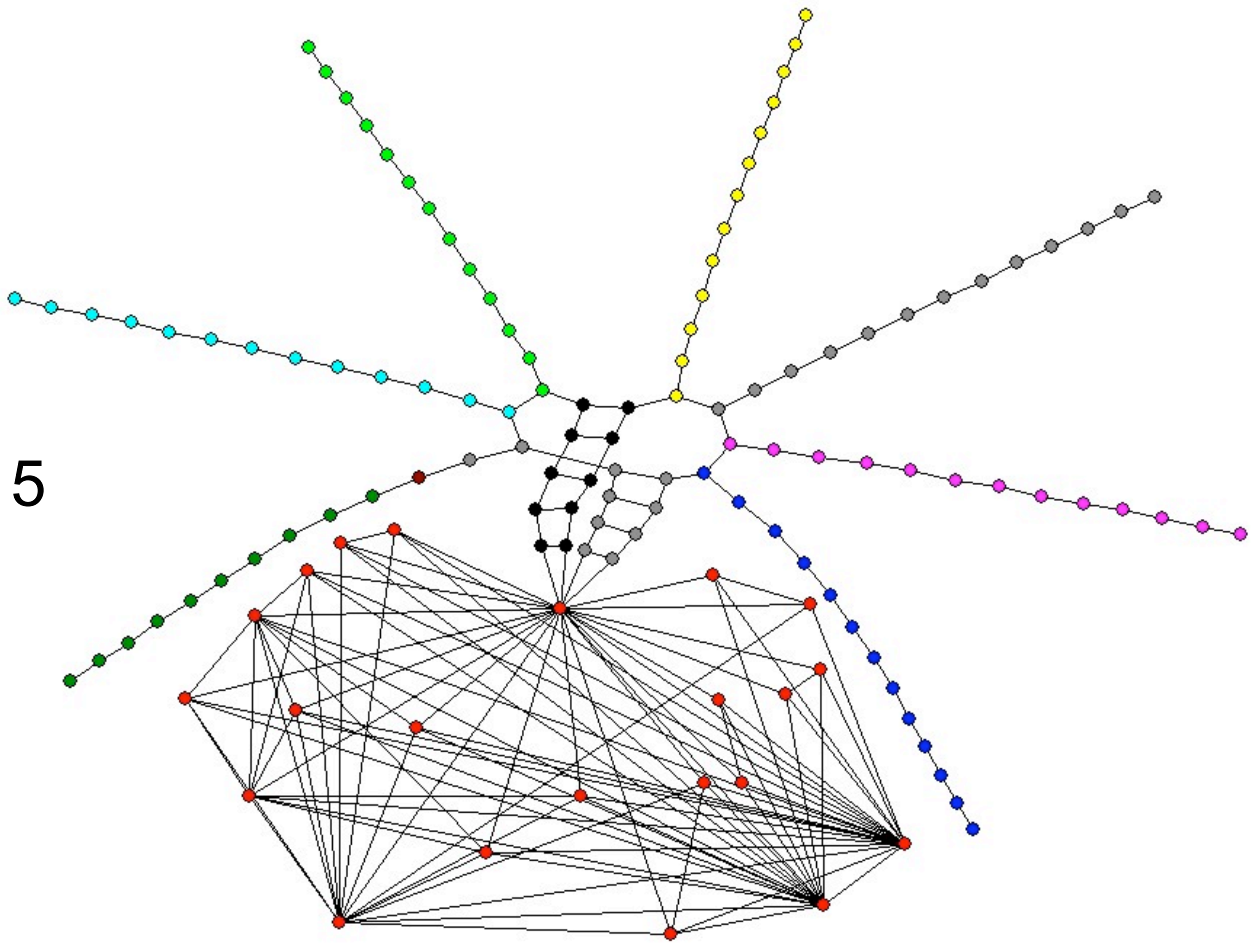
Network Type	**Clustering Coefficient**	**Path Length**	**Pearson Degree Coeff**
Random (Erdös & Rényi)	Small	Small	Zero
Regular Grid	Large	Large compared to random	Positive
Regular Grid Randomly Rewired (Watts and Strogatz)	Large	Small, similar to random	?
Trees	Small	Small	Negative
“Sociological”	Large compared to random	Small	Positive?
“Technological”	Large	?	Negative?

Conventional Wisdom Regarding *r*

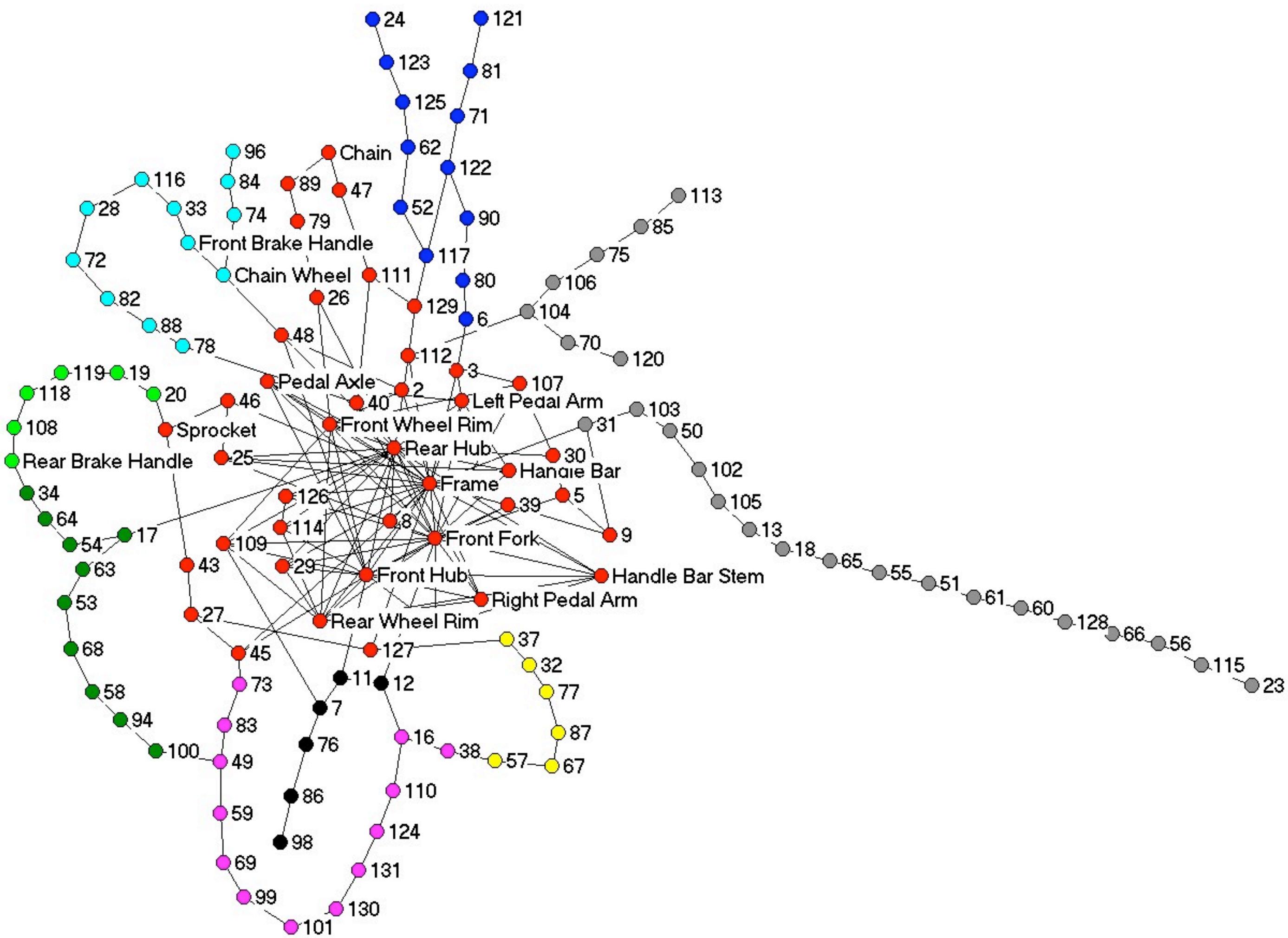
- Positive *r* means hubs connect to hubs
- Positive *r* means that high degree nodes tend to connect to each other and so do low degree nodes
- If self-loops and multiple edges between nodes are not allowed, then hubs have no choice but to connect to low-degree nodes, so r will be < 0 (“hubs repel each other”)
- These explanations do not work reliably, although the converses work sometimes
  – If high k link to high k and low k to low k then $r > 0$
- Note: a random graph has $r = 0$ (-0.0105 in MATLAB)
- Also, small networks can have big values of *r*

Bike Rewired to Have Max r

r = 0.1815



Another Bike with r = 0.1448



Li-Alderson "Toy" Internet Client-Server Networks All Have Same Degree Sequence and $r \sim -0.17$

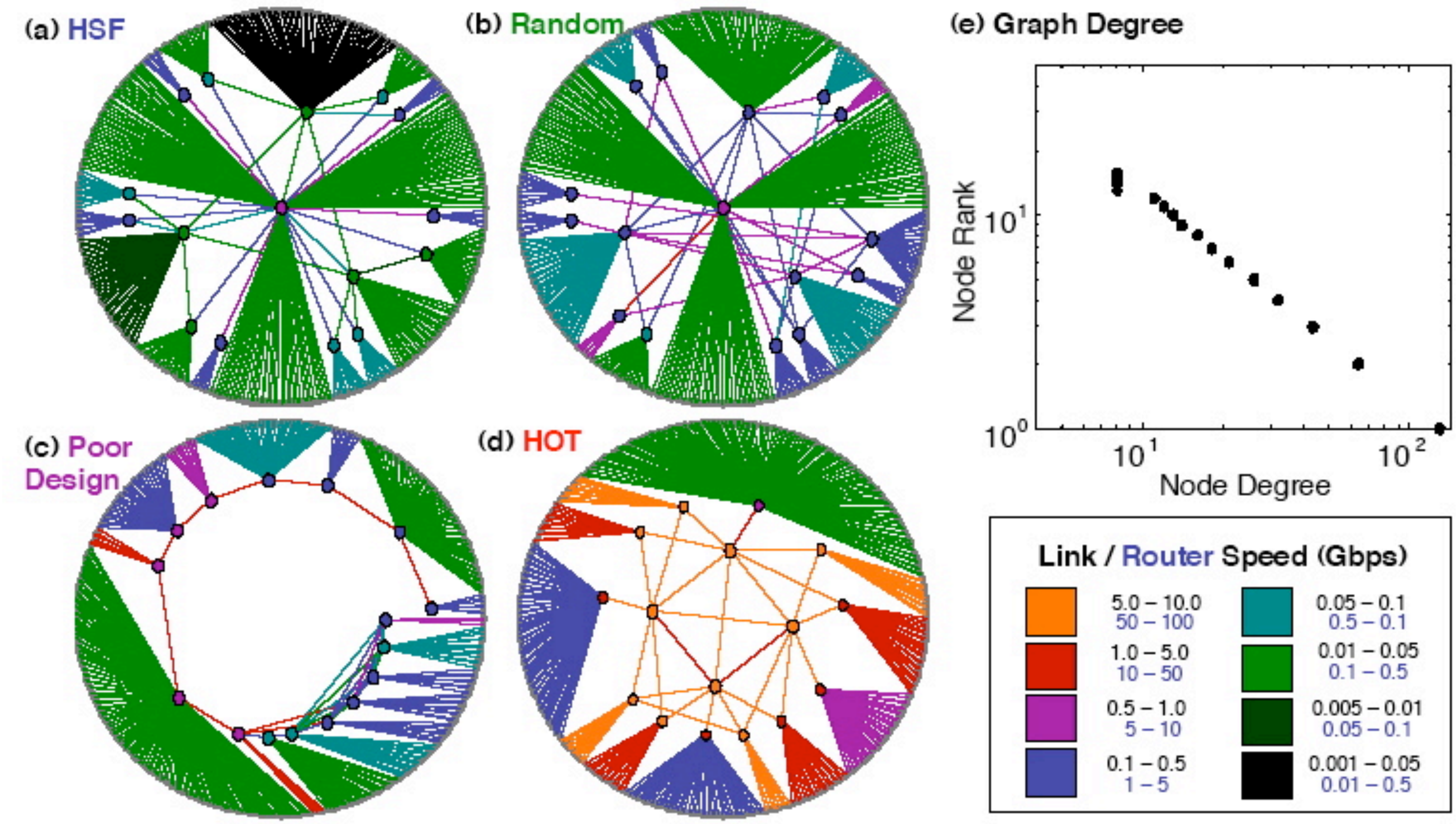


Figure 5 in Li, Lun, David Alderson, John C. Doyle, and Walter Willinger. "Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications." *Internet Mathematics* 2, no. 4 (2006): 431-523. Reproduced courtesy of A K Peters, Ltd. and David Alderson. Used with permission.