Network Observational Methods and Quantitative Metrics: II

- Topics
  - Degree correlation
  - Exploring whether the sign of degree correlation can be predicted from network type or similarity to regular structures, or details about the network itself, or maybe nothing
  - Calculating degree correlation for simple regular structures like trees and grids

Summary Properties of Several Big Networks (Newman)

	Network	Type	n	m	z	l	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
Social	Film actors	Undirected	449, 913	25, 516, 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	Company directors	Undirected	7, 673	55, 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	Math coauthorship	Undirected	253, 339	496, 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	Physics coauthorship	Undirected	52, 909	245, 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	Biology coauthorship	Undirected	1, 520, 251	11, 803, 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	Telephone call graph	Undirected	47, 000, 000	80, 000, 000	3.16		2.1				8, 9
	Email messages	Directed	59, 912	86, 300	1.44	4.95	1.5/2.0		0.16		136
	Email address books	Directed	16, 881	57, 029	3.38	5.22	–	0.17	0.13	0.092	321
	Student relationships	Undirected	573	477	1.66	16.01	–	0.005	0.001	-0.029	45
	Sexual contacts	Undirected	2, 810				3.2				265, 266
Information	WWW.nd.edu	Directed	269, 504	1, 497, 135	5.55	11.27	2.1/2.4	0.11	0.29	-0.067	14, 34
	WWW.Altavista	Directed	203, 549, 046	2 ,130, 000, 000	10.46	16.18	2.1/2.7				74
	Citation network	Directed	783, 339	6, 716, 198	8.57		3.0/–				351
	Roget's thesaurus	Directed	1, 022	5, 103	4.99	4.87	–	0.13	0.15	0.157	244
	Word co-occurrence	Undirected	460, 902	17, 000, 000	70.13		2.7		0.44		119, 157
Technological	Internet	Undirected	10, 697	31, 992	5.98	3.31	2.5	0.035	0.39	-0.189	86, 148
	Power grid	Undirected	4, 941	6, 594	2.67	18.99	–	0.10	0.080	-0.003	416
	Train routes	Undirected	587	19, 603	66.79	2.16	–		0.69	-0.033	366
	Software packages	Directed	1, 439	1, 723	1.20	2.42	1.6/1.4	0.070	0.082	-0.016	318
	Software classes	Directed	1, 377	2, 213	1.61	1.51	–	0.033	0.012	-0.119	395
	Electronic circuits	Undirected	24, 097	53, 248	4.34	11.05	3.0	0.010	0.030	-0.154	155
	Peer-to-peer network	Undirected	880	1, 296	1.47	4.28	2.1	0.012	0.011	-0.366	6, 354
Biological	Metabolic network	Undirected	765	3, 686	9.64	2.56	2.2	0.090	0.67	-0.240	214
	Protein interactions	Undirected	2, 115	2, 240	2.12	6.80	2.4	0.072	0.071	-0.156	212
	Marine food web	Directed	135	598	4.43	2.05	–	0.16	0.23	-0.263	204
	Freshwater food web	Directed	92	997	10.84	1.90	–	0.20	0.087	-0.326	272
	neural network	Directed	307	2, 359	7.68	3.97	–	0.18	0.28	-0.226	416, 421

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 $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.

Table 2 in Mark E. J. Newman's. "The Structure and Function of Complex Networks." *SIAM Review* 45, no. 2 (2003): 167-256.

Community-finding and Pearson Coefficient r

- Newman says technological networks seem to have $r < 0$ while social networks seem to have $r > 0$
- Newman and Park sought an explanation in community structure and clustering: "Why Social Networks are Different From Other Kinds of Networks" *Phys Rev E*, **68**, 036122 (2003)
- Social networks can arise by people joining multiple groups and generating multiple connections
- Networks derived from these multiple connections have positive *r*
- Networks coauthconn and coauthwhole are from this paper
  - coauthconn is the connected portion with 147 nodes
  - coauthwhole has 42 clusters, smallest has 2 nodes, biggest has 5

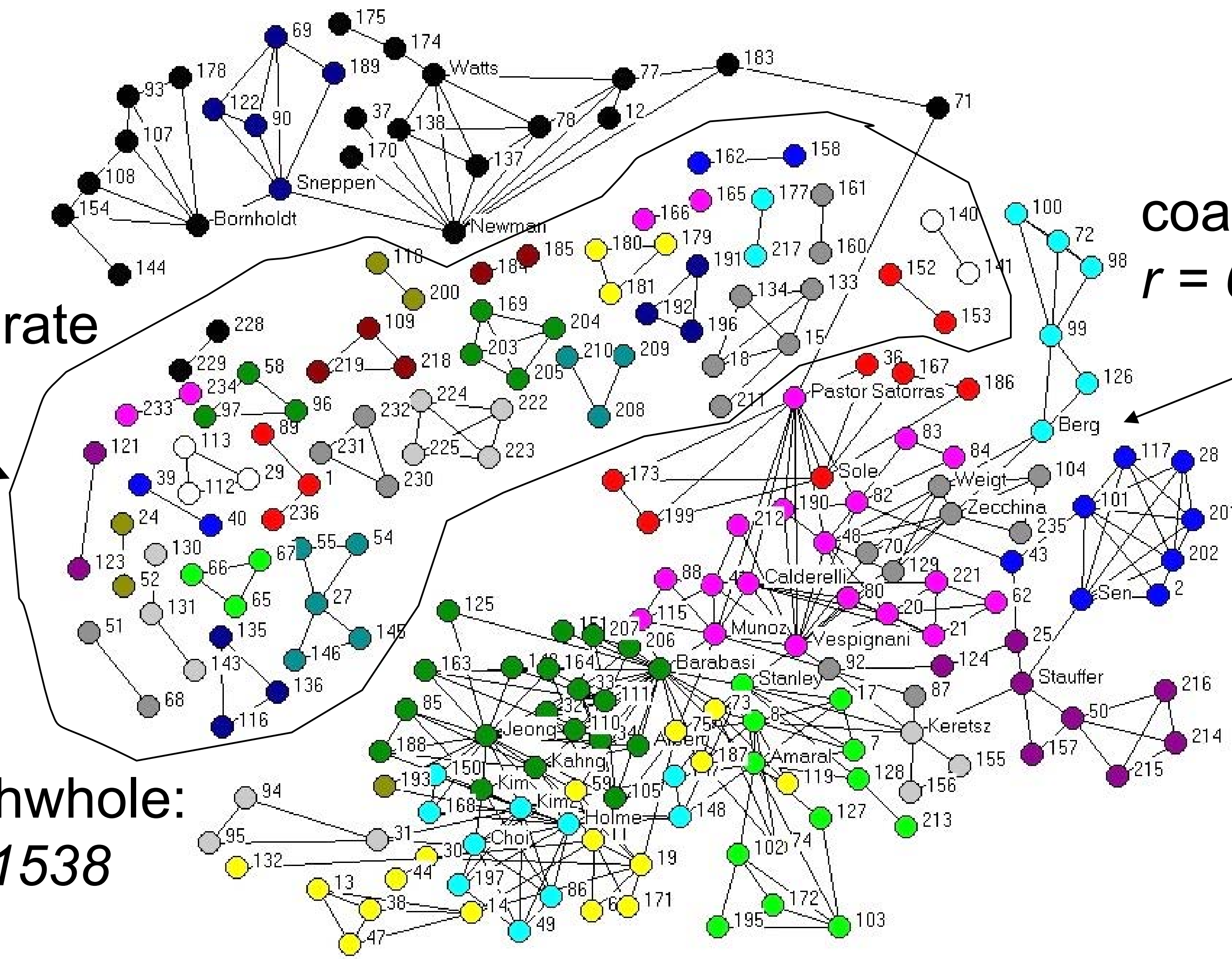
“Why Social Networks are Different”

- “Left to their own devices, we conjecture, networks normally have negative values of *r*. In order to show a positive value of *r*, a network must have some specific additional structure that favors assortative mixing.”
- Special structure that explains networks with $r > 0$:
  - Large clustering coeff compared to random network with same degree sequence
  - Community structure
- (No special structure needed to explain $r < 0$)

Physics Coauthors Network

41 separate clusters

coauthconn:
r = 0.0159



coauthwhole:
r = 0.1538

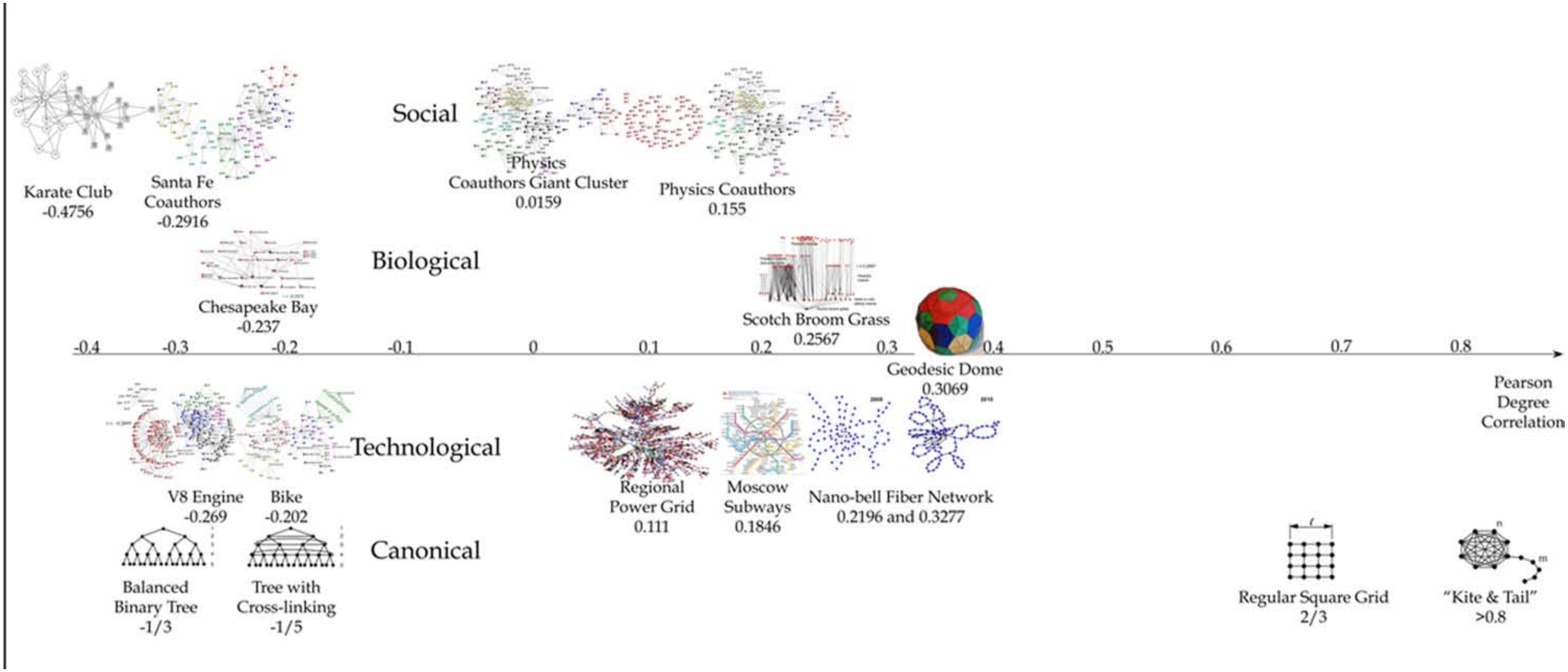
Work with David Alderson

- "Are Social Networks Really Different?"
- ICCS 2006 paper, published in NECSI journal

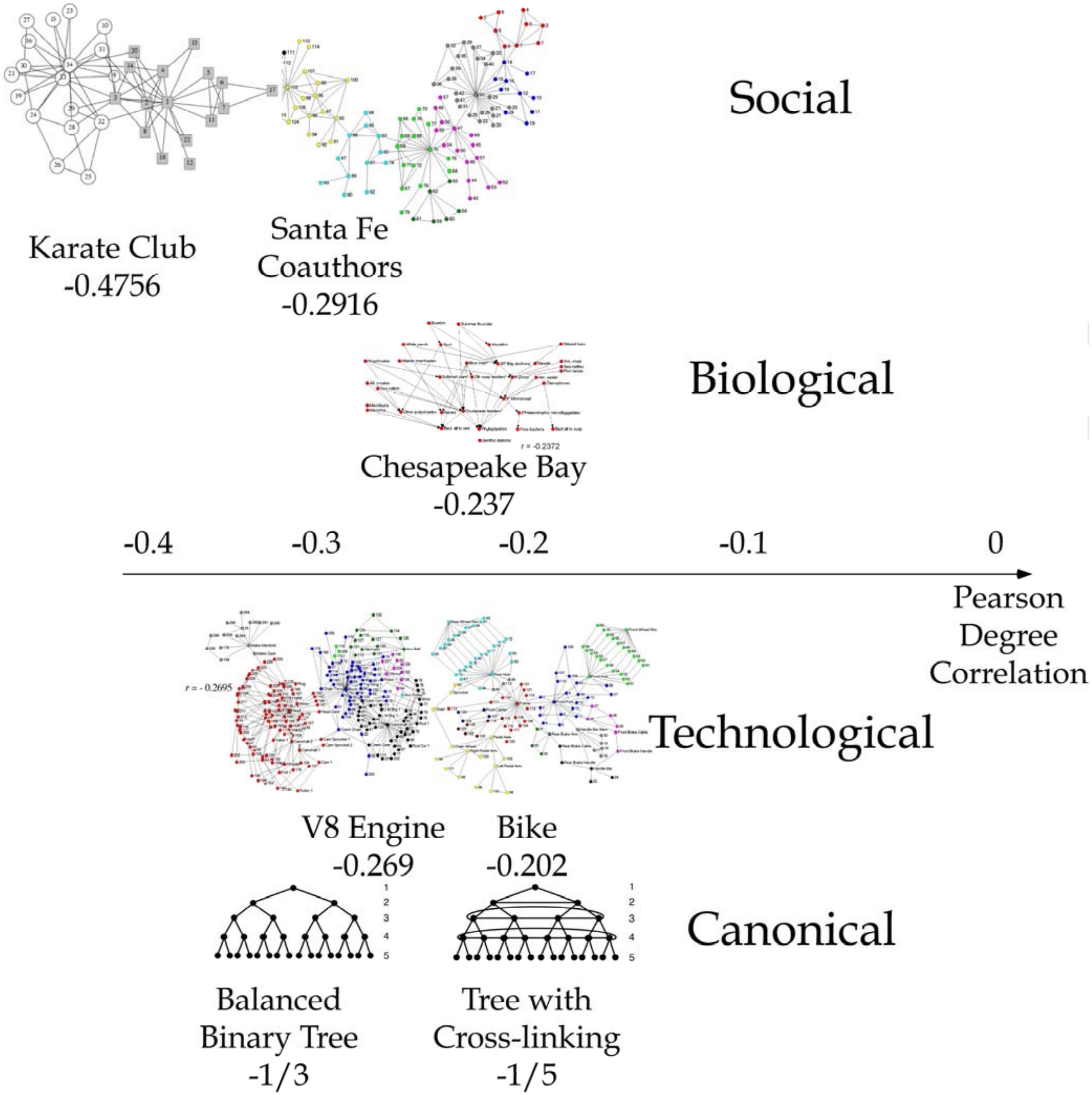
Our Observations

- Data show a mixture of $r > 0$ and $r < 0$ for all kinds of networks (38 simple connected networks)
- Many with $r < 0$ have community structure
- There is a structural explanation for this, based on a structural property that all networks have: the variability of the degree sequence, but not related to category: social - technological
- Can use it to show that certain networks cannot possibly have $r > 0$
- Also, some canonical structures have $r < 0$ or $r > 0$ and real networks share properties with these canonical structures: trees have $r < 0$ and grids have $r > 0$

Full Range of r

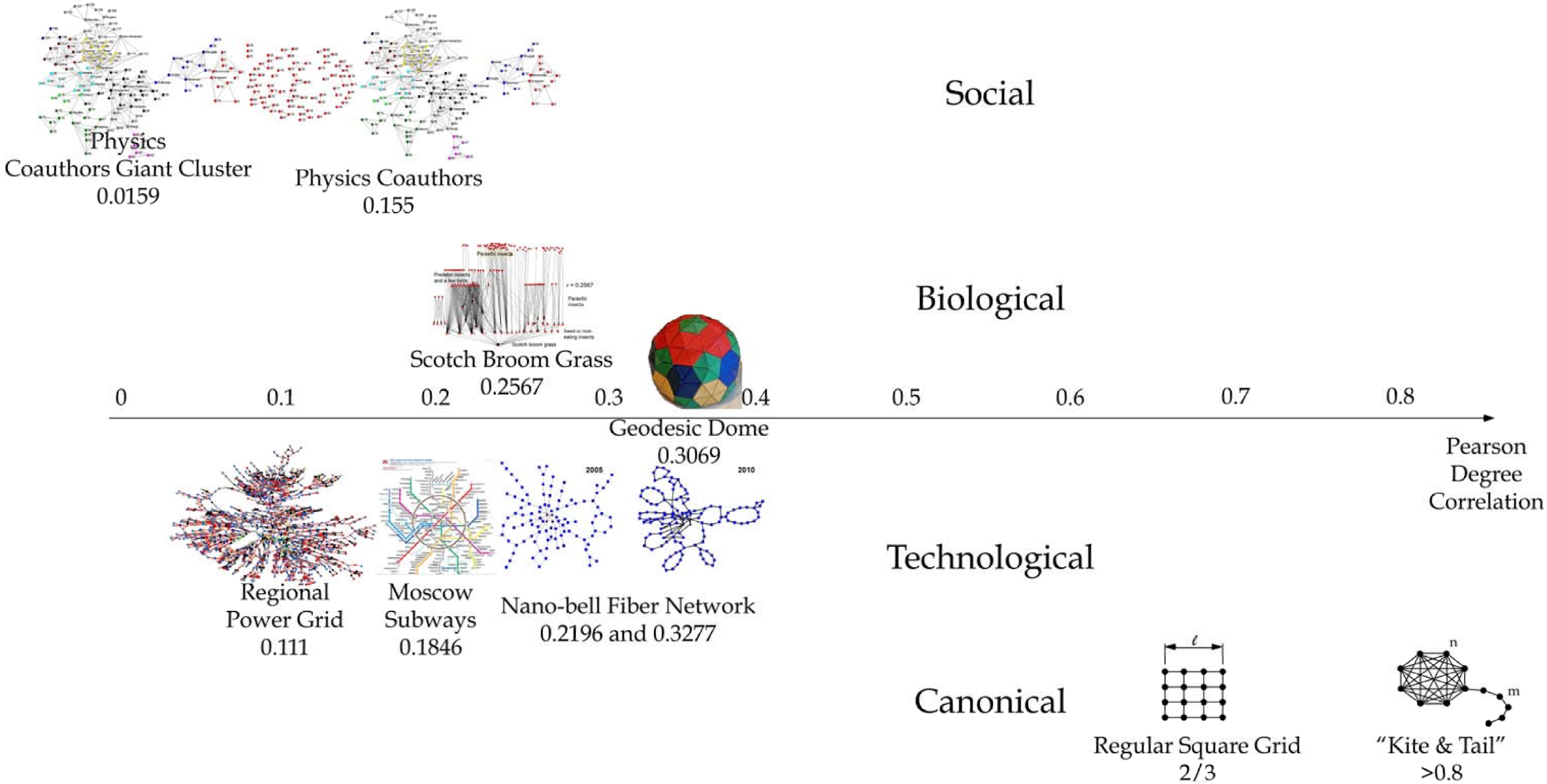


Negative r



Social
Karate Club
-0.4756
Santa Fe
Coauthors
-0.2916
Biological
Chesapeake Bay
-0.237
-0.4
-0.3
-0.2
-0.1
0
Pearson
Degree
Correlation
Technological
V8 Engine
-0.269
Bike
-0.202
Canonical
Balanced
Binary Tree
-1/3
Tree with
Cross-linking
-1/5

Positive r



Social
Physics
Coauthors Giant Cluster
0.0159
Physics Coauthors
0.155
Biological
Scotch Broom Grass
0.2567
Geodesic Dome
0.3069
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Pearson
Degree
Correlation
Technological
Regional
Power Grid
0.111
Moscow
Subways
0.1846
Nano-bell Fiber Network
0.2196 and 0.3277
2005
2010
Canonical
Regular Square Grid
2/3
"Kite & Tail"
>0.8

More Data on *r* vs Network Type

Network	n	m	$<k>$	fraction of nodes: $k > \bar{x}$	r	C	Random clustering coeff $<k>/n$	Generalized random clustering coeff $\frac{<k>}{n}\left[\frac{<k^2>-<k>}{<k>^2}\right]^2$
Karate Club	34	78	4.5882	0.1471	-0.4756	0.5879	0.1349	0.2937
J. Tirole ŅErdos NetworkÓ2	93	149	3.204	0.0645	-0.4412	0.5124	0.0345	0.2097
J. Stiglitz ŅErdos NetworkÓ2	68	85	2.50	0.0882	-0.4366	0.7019	0.0368	0.1768
Scheduled Air Routes, US	249	3389	27.22		-0.39	0.64	0.109	
Littlerock Lake food web*	92	997	10.837	0.337	-0.3264	0.256	0.117	0.1909
Grand Piano Action 1 key	71	92	2.59	0.197	-0.3208	0.1189	0.0365	0.0275
Santa Fe coauthors	118	198	3.3559	0.0593	-0.2916	0.729	0.0284	0.1044
V8 engine	243	367	3.01	0.0122	-0.269	0.2253	0.0124	0.192
Grand Piano Action 3 keys	177	242	2.73	0.2034	-0.227	0.1209	0.0154	0.0182
Exercise walker	82	116	2.8293	0.0854	-0.2560	0.4345	0.0345	0.1288
Abeline	886	896	2.023	0.0158	-0.2239	0.0076	0.0023	0.0543
Bike	131	208	3.1756	0.0458	-0.2018	0.4155	0.024	0.082
Six speed transmission	143	244	3.4126	0.1	-0.1833	0.2739	0.0238	0.0413
ŅHOTÓ	1000	1049	2.098	0.0170	-.1707	0	0.0021	0.0353
Car Door DSM*	649	2128	3.279		-0.1590		0.0051	
Jet Engine DSM*	60	639	10.65		-0.1345		0.1775	
TV Circuit*	329	1050	6.383	0.018	-0.109	0.529	0.0194	0.1157
Tokyo Regional Rail	147	204	2.775	0.3401	-0.0911	0.0783	0.0188	0.0157
FAA Nav Aids, Unscheduled	2669	7635	5.72		-0.0728		0.0021	
Canton food web*	102	697	6.833	0.157	-0.0694		0.0670	0.3979

Color code: Ņsocial,Óassemblies, rail lines, trophic food webs, software call graphs, power grids, internet/phone, Design Structure Matrix, air routes, electric circuits

Data Continued

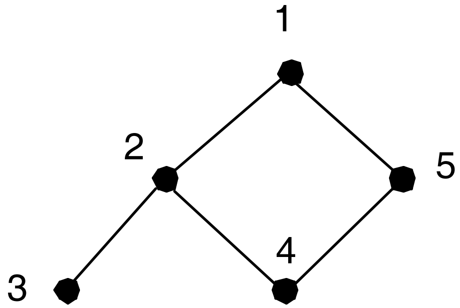
Network	n	m	$<k>$	fraction of nodes: $k > \bar{x}$	r	C	Random clustering coeff $<k>/n$	Generalized random clustering coeff $\frac{<k>}{n}\left[\frac{<k^2>-<k>}{<k>^2}\right]^2$
TV Circuit*	329	1050	**6.383**	0.018	-0.109	**0.529**	**0.0194**	**0.1157**
Tokyo Regional Rail	147	204	**2.775**	0.3401	-0.0911	**0.0783**	**0.0188**	**0.0157**
FAA Nav Aids, Unscheduled	2669	7635	**5.72**		-0.0728		**0.0021**	
Canton food web*	102	697	**6.833**	0.157	-0.0694		**0.0670**	**0.3979**
Mozilla19980331 *	811	4077	**5.0271**	0.0259	-0.0499		**0.0062**	
Mozilla all comp*	1187	4129	**3.4785**		-0.0393		**0.0029**	
Munich Schnellbahn	50	65	2.6	0.34	-0.0317	0.0892	0.052	0.0545
FAA Nav Aids, Scheduled	1787	4444	4.974		-0.0166		0.0028	
St. Marks food web *	48	221	4.602	0.146	-0.0082		0.0959	
Western Power Grid	4941	6594	2.6691	0.2022	0.0035	0.1065	0.00054	0.000625
Unscheduled Air Routes, US	900	5384	11.96		0.0045		0.0133	
Apache call list*	62	365	5.88		0.007		0.095	
Physics coauthors	145	346	4.7724	0.1517	0.0159	0.6905	0.0329	0.0578
Tokyo Regional Rail plus Subways	191	300	3.1414	0.4188	0.0425	0.0897	0.0164	0.0156
Traffic Light controller*	133	255	**1.9173**		0.0614		**0.0144**	
Berlin U- & S-Bahn	75	111	2.96	.48	0.0957	0.1171	.00395	0.032
London Underground	92	139	3.02	0.413	0.0997	0.2223	0.0328	0.0296
Regional Power Grid	1658	2589	3.117	0.1695	0.1108	0.1683	0.002	0.0027

Data Continued

Network	n	m	$<k>$	fraction of nodes: $k > \bar{x}$	r	C	Random clustering coeff $<k>/n$	Generalized random clustering coeff $\frac{<k>}{n}\left[\frac{<k^2>-<k>}{<k>^2}\right]^2$
Moscow Subways	51	82	3.216	0.1765	0.1846	0.1061	0.0631	0.0595
Nano-bell	104	121	2.327		0.2196	0.0262	0.022	
Broom food web*	82	223	2.623		0.2301		0.0309	
Company directors	6731	50775	15.09	0.1703	0.2386	0.8682	0.0022	0.0041
Moscow Subways and Regional Rail	129	204	3	.4191	0.2601	0.0803	0.0232	0.0186

How to Calculate r in Closed Form for Canonical Structures

Review: Calculating *r* from the Edge List

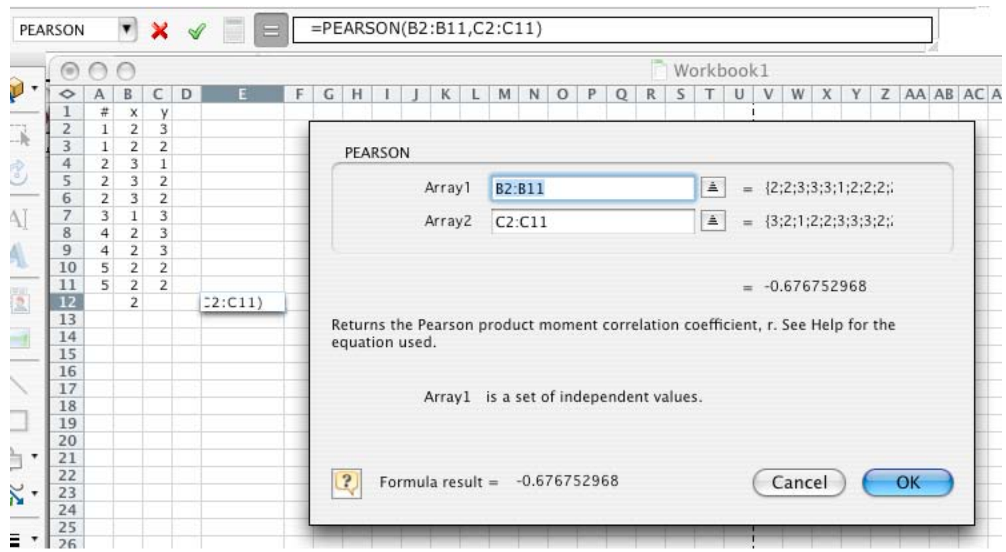


$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

	k_{out}	k_{in}
node	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
average	2.2	pearson -0.67675297

$\bar{x} = 2.2$

$\bar{y} = 2.2$



r = -0.676752968 using Pearson function in Excel

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

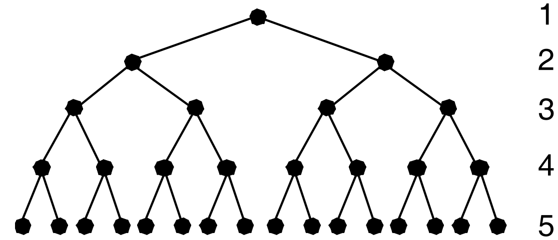
Census of Pairs for Pure Binary Tree

n=1
n=2
n=3
n=4
n=5

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

2 rows like this - ignore

6 rows like this - ignore



All other rows like this:
3-3
(except last two sets)

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
$(3-\bar{x})(1-\bar{x})$

Ksum total rows
2^n total rows of 3-1
~Ksum - 2^n rows of 3-3

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

Result of Census

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

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

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

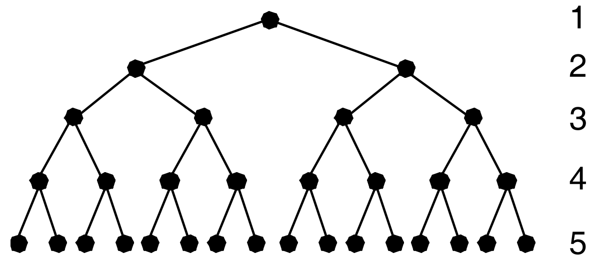
Also <k>=2

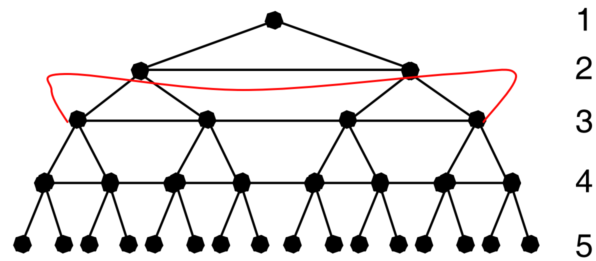
Total 2^n rows of 3-1
Approx (ksum - 2^n) rows of 3-3

Denominator = $\sqrt{(x-\bar{x})^2 (y-\bar{y})^2} = \sqrt{(x-\bar{x})^4} = (x-\bar{x})^2$
This is just one column's entries squared

r = - 0.4122 for this tree with 5 layers

Closed Form Results

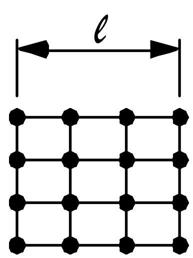




Property	**Pure Binary Tree**	**Binary Tree with Cross-linking**
ksum	$2^{n+1}-4$	$3*2^n-10$
ksqsum	$10*2^{n-1}-14$	$13*2^n-64$
$\bar{x}$	$\rightarrow 2.5$ as n becomes large (>~ 6)	$\rightarrow \frac{13}{3}$ as n becomes large (>~ 6)
Pearson numerator	$\sim 2^n(3-\bar{x})(1-\bar{x})+(ksum-2^n)(3-\bar{x})^2$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



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

$\bar{x}=4$ so all terms in $(4-\bar{x})$ disappear

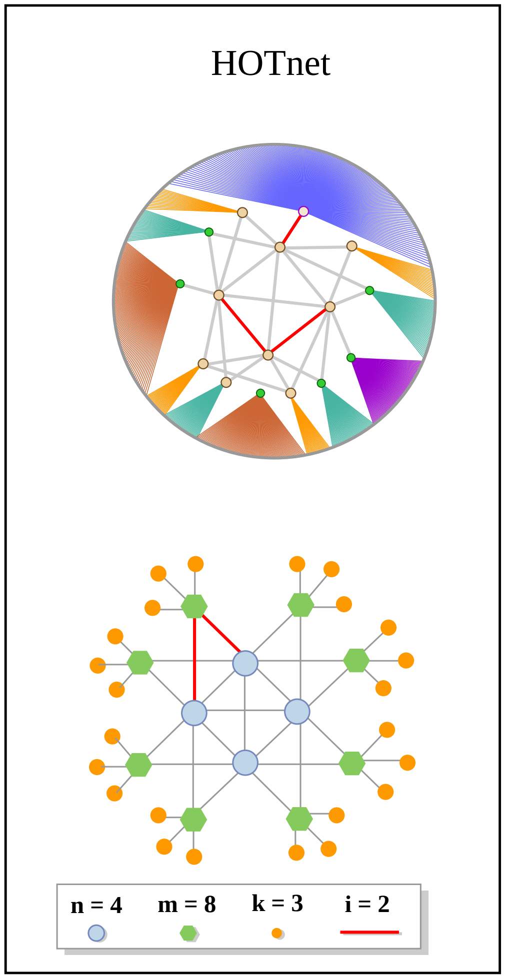
"HOT" Network: A WAN

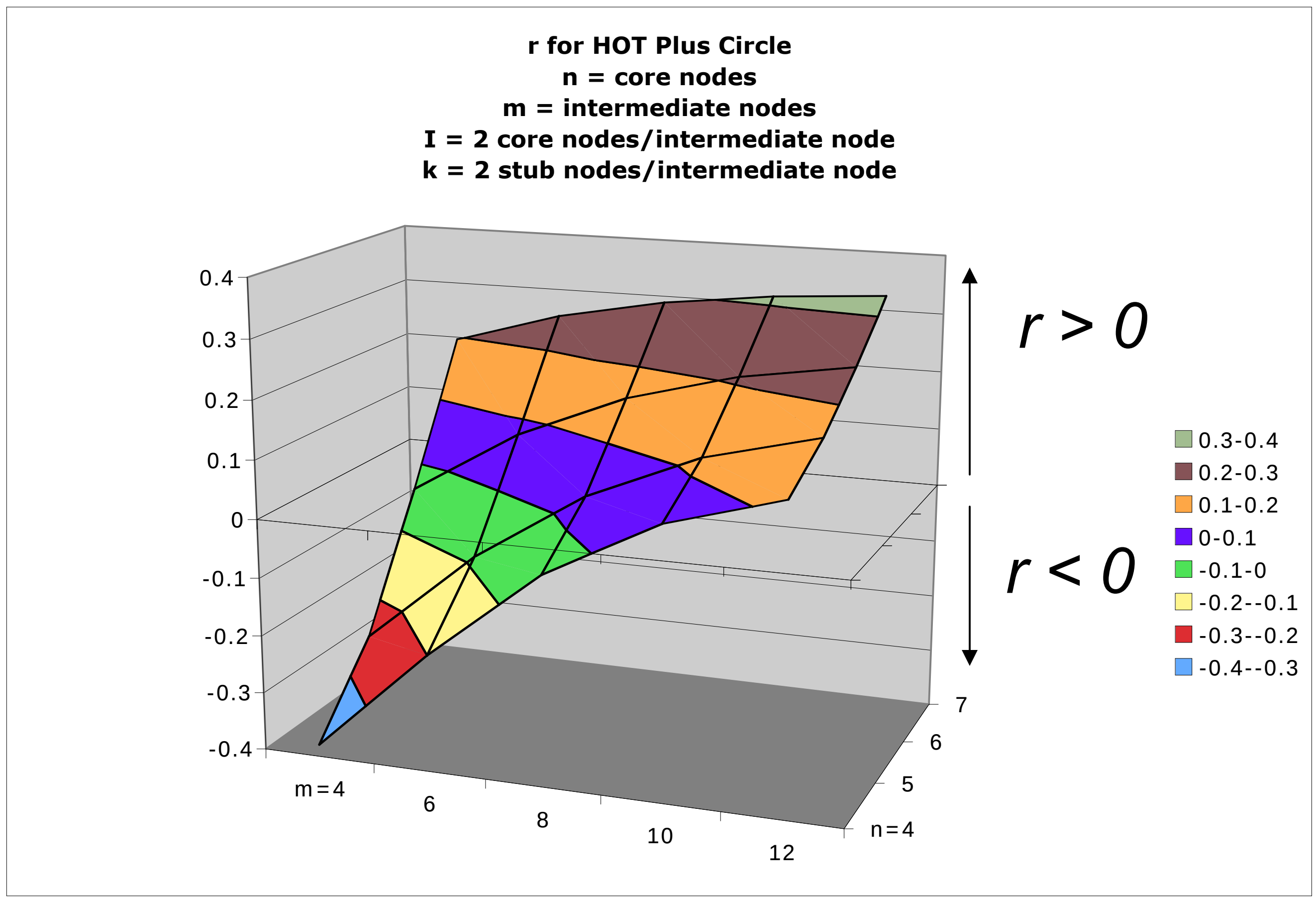
Num_rows=n*(n-1+m*I/n)+m*(I+k+2)+m*k

Row_sum=n*(n-1+m*I/n)^2+m*(I+k+2)^2+m*k

numerator=(n-1+m*I/n-xbar)^2*n*(n-1)+2*(n-1+m*I/n-xbar)*(I+k+2-xbar)*m*I
+(I+k+2-xbar)^2*2*m+(I+k+2-xbar)*(1-xbar)*2*k*m

denom=(n-1+m*I/n-xbar)^2*n*(n-1+m*I/n)
+(I+k+2-xbar)^2*m*(I+k+2)+(1-xbar)^2*m*k





r = - 0.1707

HOT is a Tree with the Core at the Top

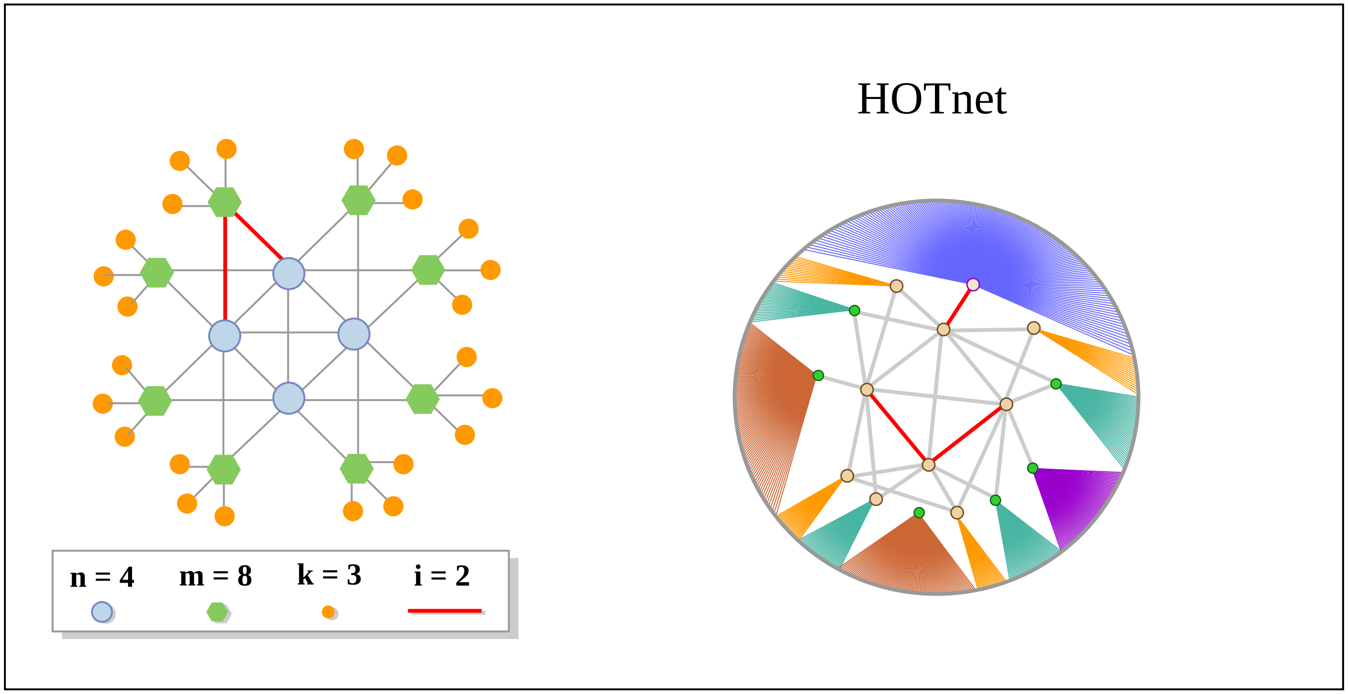
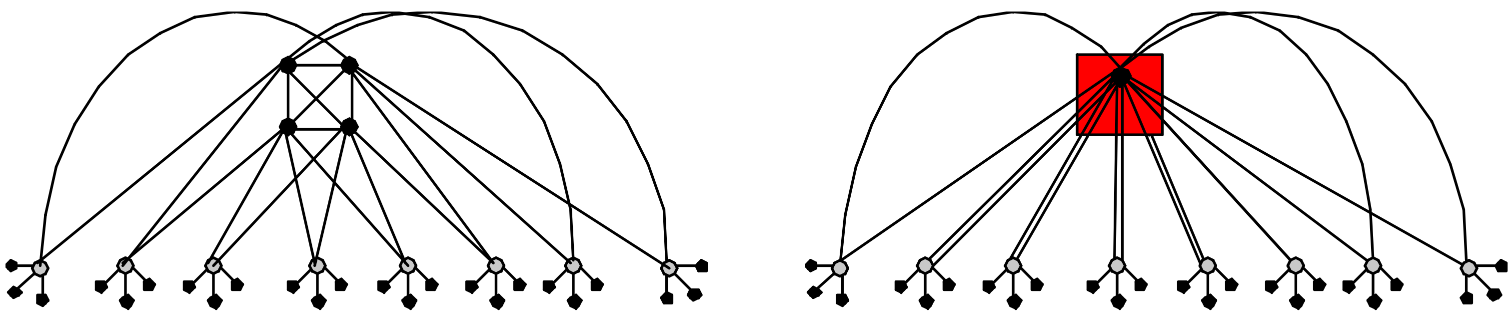
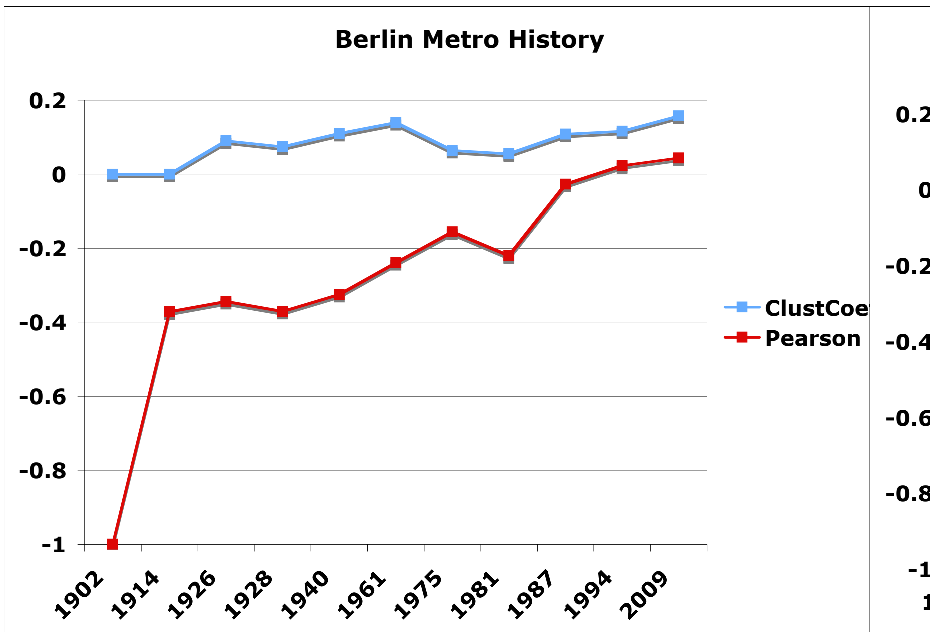
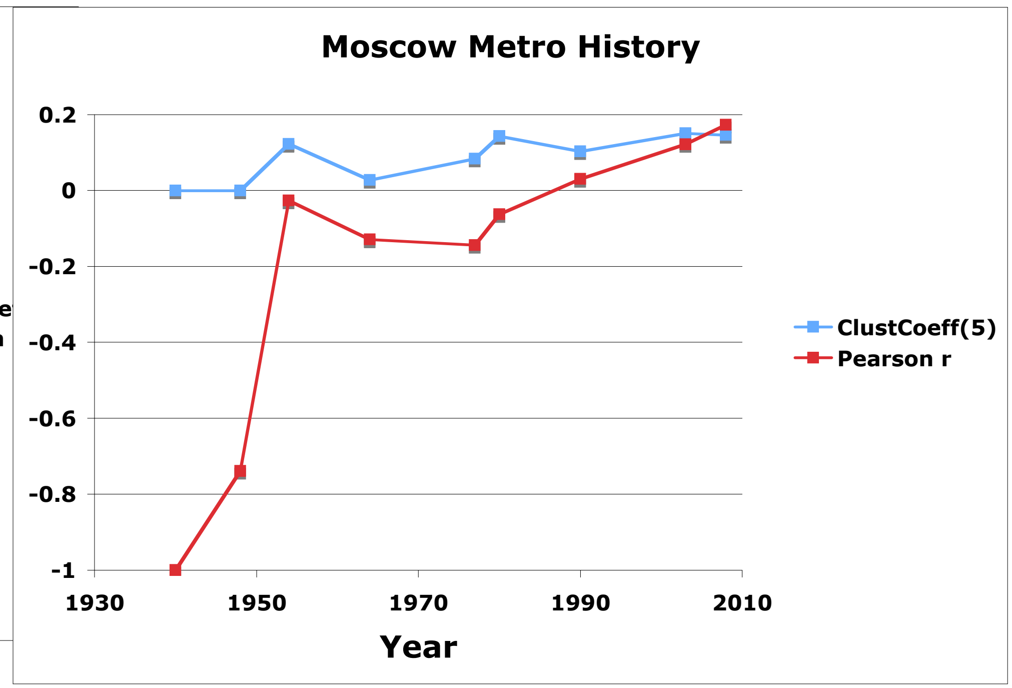


Image by MIT OpenCourseWare.

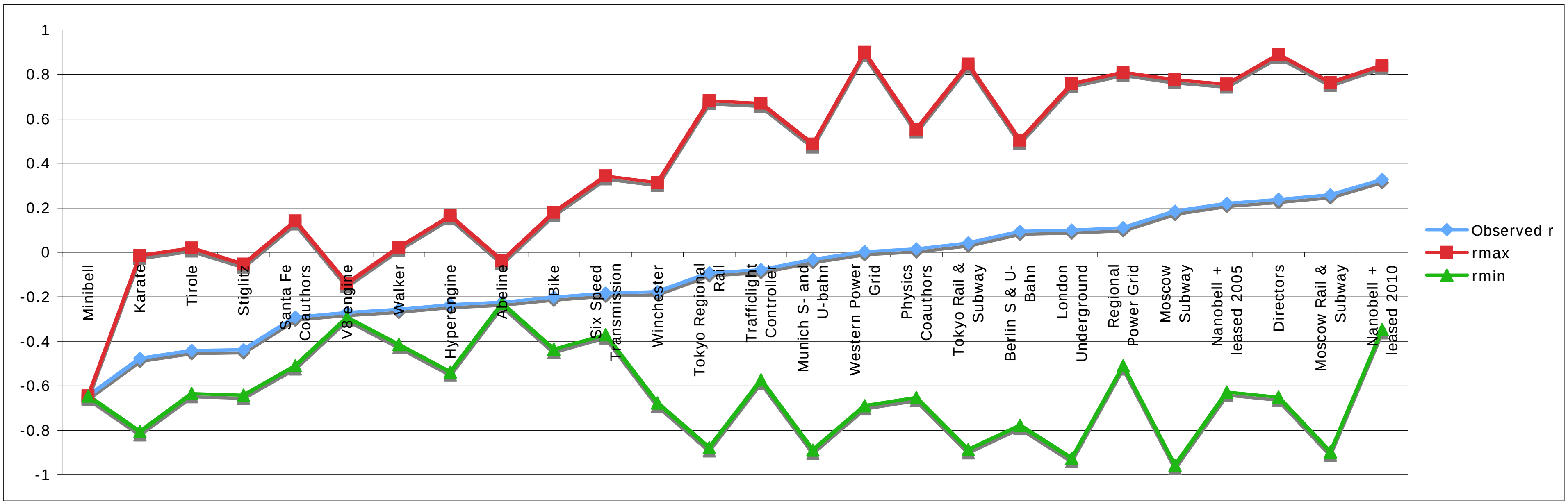


Actual Metro System Histories





Networks' Observed *r* Against Background Found by Rewiring While Preserving the Degree Sequence

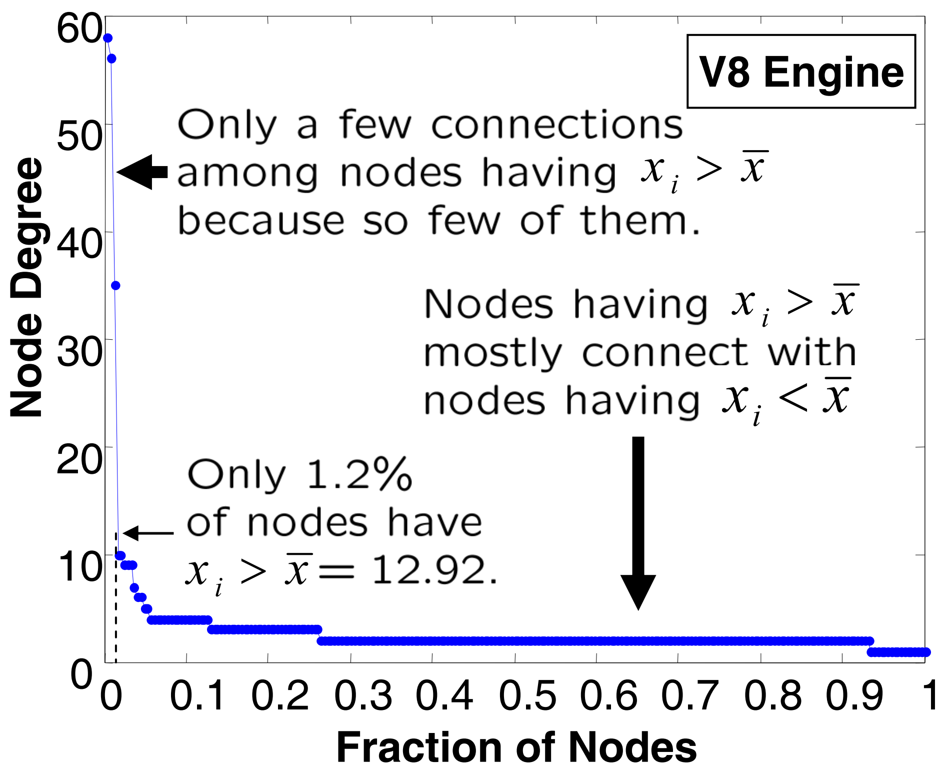


For r < -0.1, the background range is restricted to mostly negative values

For r > -0.1, the background covers most of [-1, 1]

How Degree Sequence Constrains r for One Network But Not for Another

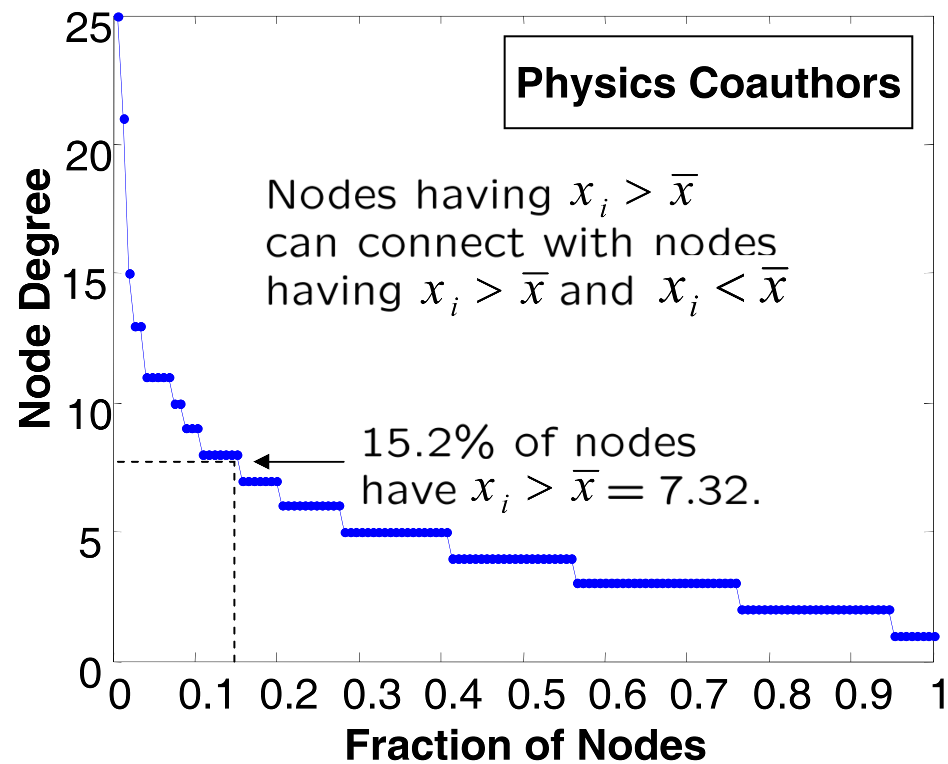
Almost no $x_i > \bar{x}$



$r = -0.269$

Range = [- 0.2932, - 0.1385]

Many $x_i > \bar{x}$



$r = 0.016$

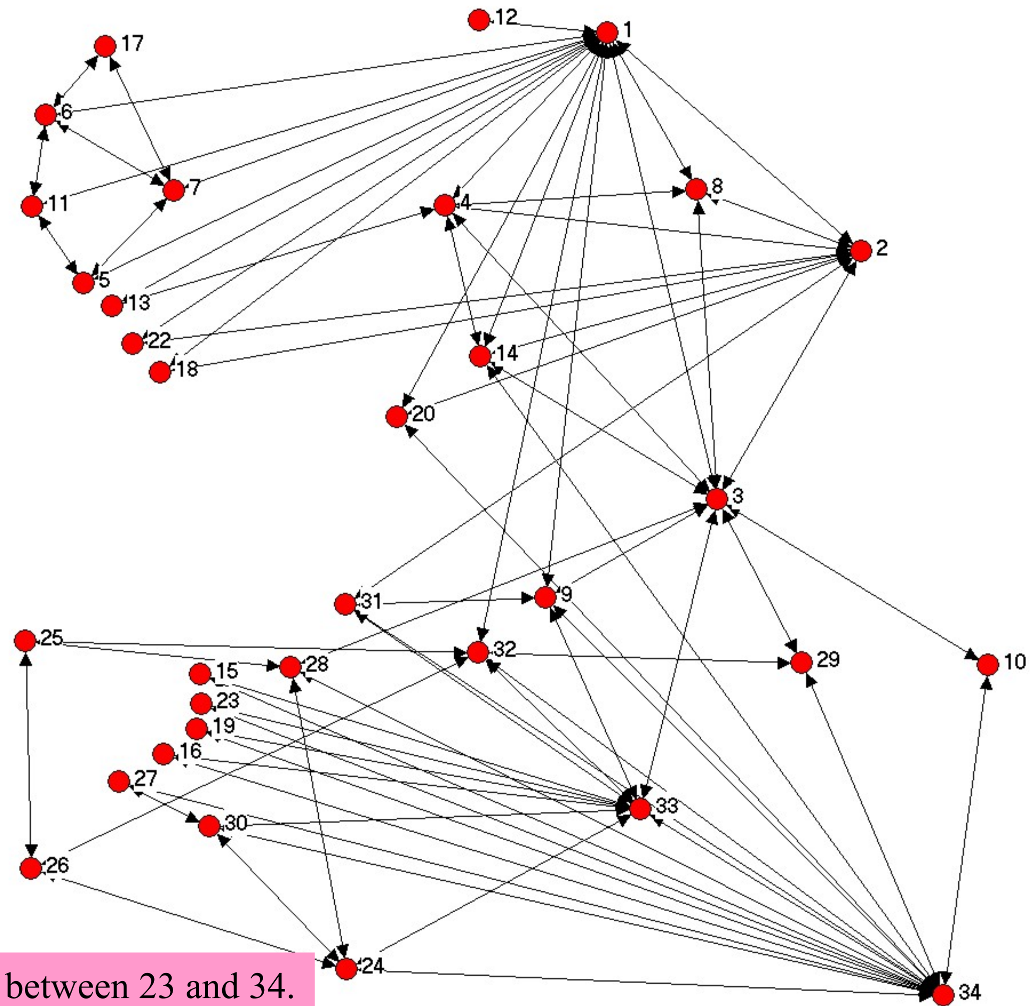
Range = [- 0.652, 0.553]

Conclusions

- A network's domain ("social," "technical," etc.) is not a reliable predictor of the sign of r
- The degree sequence imposes considerable structural constraint on networks whose observed $r < 0$
- But it does not impose much constraint on networks whose observed $r > 0$
- Each network's actual circumstances impose constraint, but circumstances are stronger than the degree sequence when $r > 0$ and vice-versa when $r < 0$
- Example: cost of connection may be high for technological systems but not for social systems like coauthor or movie actor networks
- Similarly, the exact connections matter for the bike but not for the coauthors, who could in principle collaborate with anyone

Backups

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



$r = -0.475$
$C = 0.588$

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

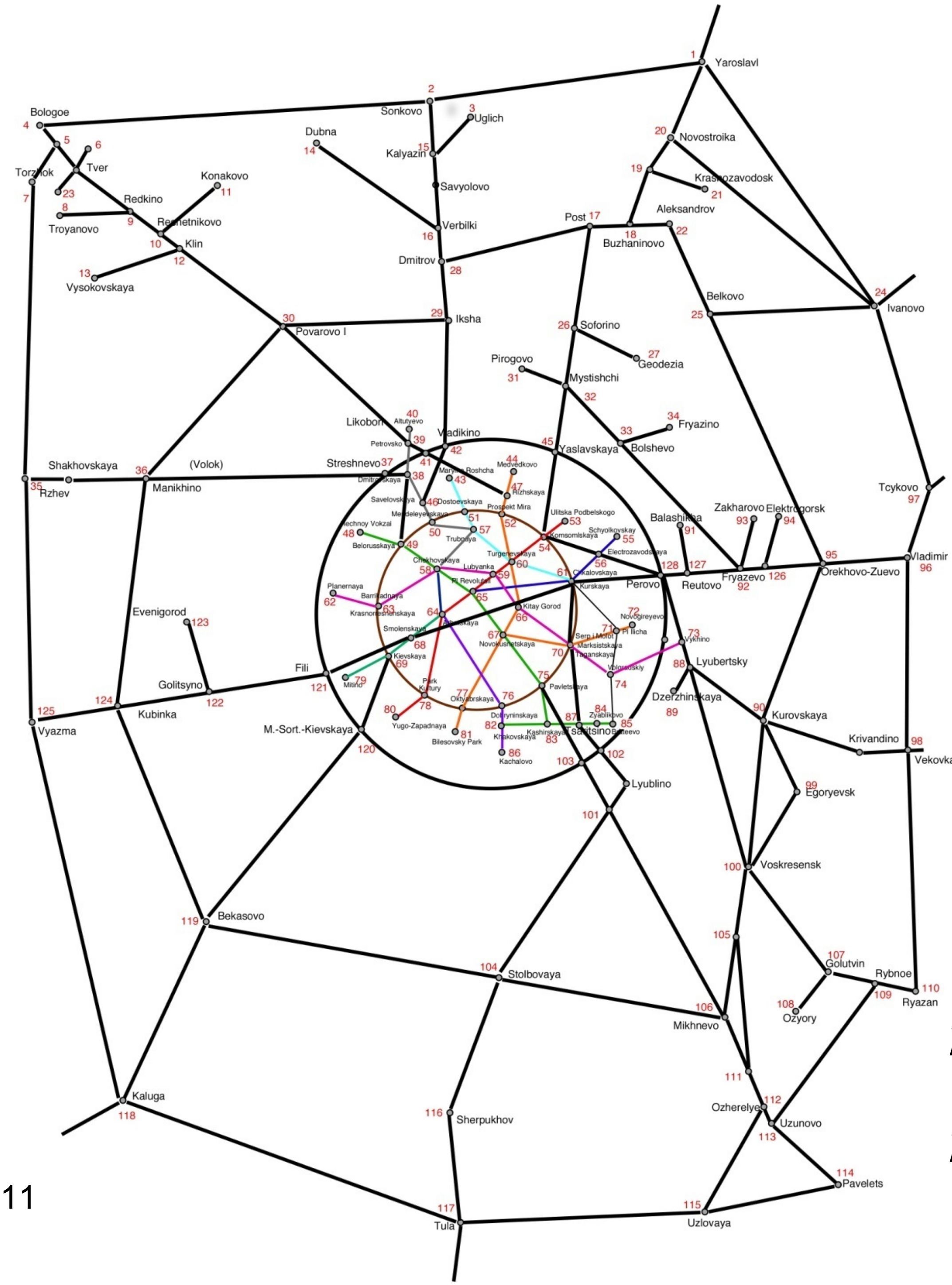
Moscow Metro

Image of Moscow Metro map removed due to copyright restrictions. See Moscow Metro.

Moscow Regional Rail

Map of Moscow Regional Rail removed due to copyright restrictions. Please refer to: The Mappery

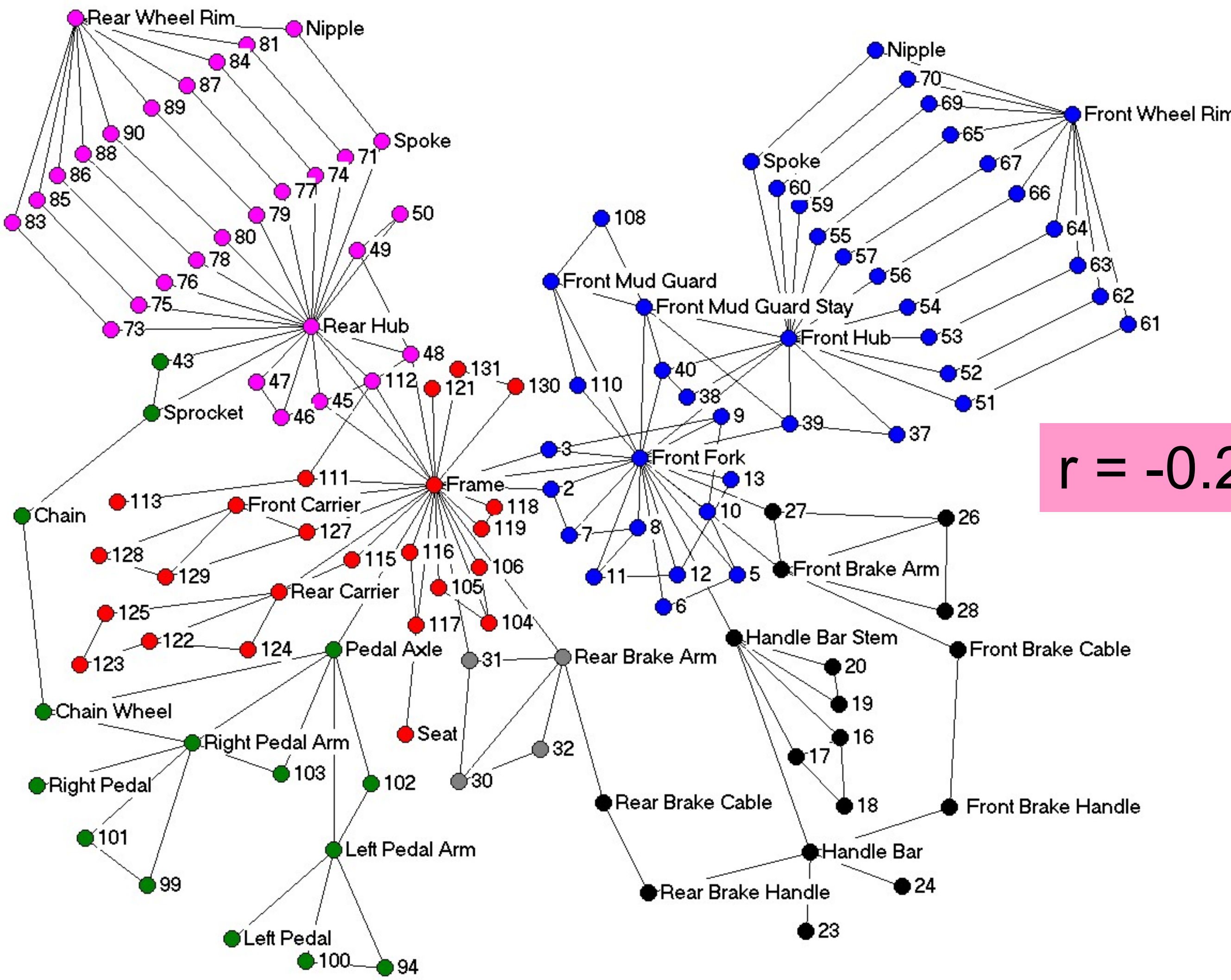
Moscow Metro and Regional Rail



$$r_{\text{subway + rail}} = 0.2601$$

$$r_{\text{subway}} = 0.1846$$

Bike

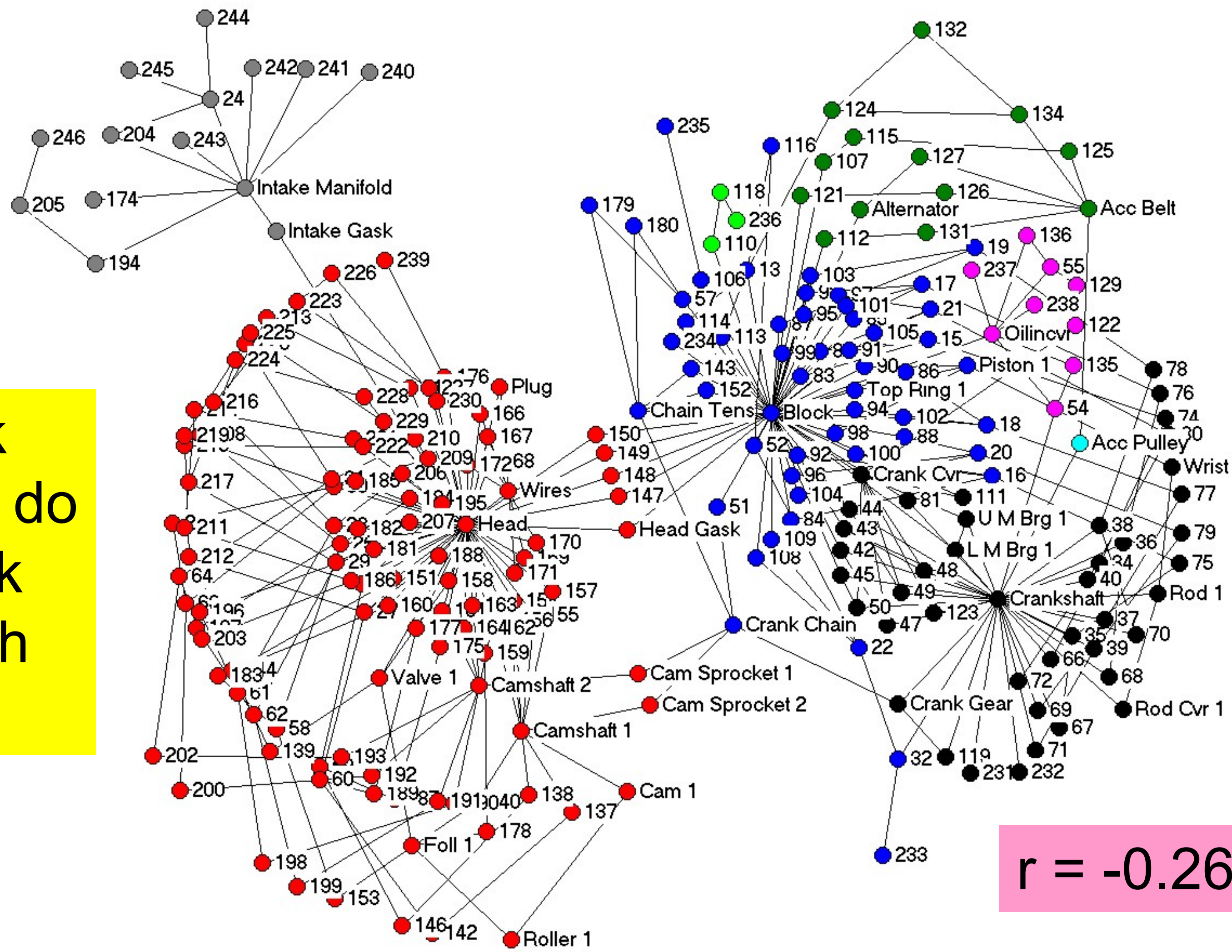


High-k nodes do not link to each other!

r = -0.2

bike newman girvan 2.jpg

V8 Engine



High-k nodes do not link to each other!

r = -0.269

Does a Network Have to Have a Particular Value of r Given its D?

- Technological networks have to perform a function, use scarce resources efficiently, or satisfy some other structural or functional constraint, so their observed wiring and *r* are probably necessary
- Social networks do not have to do any of these things so their structure is more subject to circumstances, such as communication or collaboration habits; thus their observed wiring and *r* are probably circumstantial
- We can test by seeing if rewired versions are plausible

Example Domain Sources for Constraint in D Leading to $r < 0$

- High x with respect to $\bar{x}$ in mechanical assemblies comes from need to provide a foundation part to absorb loads and locate other parts to each other
  - Engine block
  - Bike and walker frame
- High x with respect to $\bar{x}$ in some social networks reflects hierarchy or dominance
  - Karate instructor and club president in Zachary's club
- Tree-like structure of wireline phone networks causes them to have $r < 0$ because trees have $r < 0$
- These networks can't be rewired plausibly

Example Domain Sources for Constraint in D Leading to $r > 0$

- Planar transport networks are grid-like, and grids have $r > 0$
- Modern fiber-optic phone networks are built on loops or chains (trunks) that link clusters (central offices) leading to $r > 0$

ESD.342 Network Representations of Complex Engineering Systems
Spring 2010