Supplementary materials

Supplementary Methods

Short description of the Bron-Kerbosch algorithm

The Bron-Kerbosch algorithm for counting all maximal cliques (Bron & Kerbosch, 1973) in a graph is a backtracking algorithm which uses the branch and bound technique to cut off branches that cannot lead to a maximal clique. The algorithm uses three sets: *compsub*, *candidates* and *not*. The set *compsub* contains the nodes which are to be examined as maximal cliques, when traversing along a branch of the backtracking tree. The set *candidates* contain the nodes which will enlarge the set of nodes in *compsub*. The set *not* contains nodes which already served as an extension to the current *compsub* nodes, and are now explicitly excluded. The core of the algorithm is generating all possible extensions to *compsub* using the possible nodes in *candidates*, without using the nodes in *not*. The nodes in *not* were already generated in previous stages, and therefore should not be generated again. Whenever *candidates* is empty – this means that the current *compsub* could not be extended. If *not* is empty as well – this means that *compsub* is a maximal clique. In case that *not* is not empty – this means that there was already a larger clique identified, which contains both *compsub* and some extensions from the *not* set. The algorithm has different versions of how to generate the possible extensions, in order to minimize the traverse on the tree (Cazals & Karande, 2008; Koch, 2001; Tomita, Tanaka, & Takahashi, 2006).

Maximal-cliques counting algorithm

We have used the algorithm for counting maximal cliques implemented in the Boost Graph library ("Boost C++ Source Libraries,"). This algorithm is based on the Bron-Kerbosch original algorithm (Bron & Kerbosch, 1973) with a backtracking method to reduce run time. The algorithm counts the total number of maximal cliques and we added a counter for every clique size.

All-cliques counting algorithm

We have changed the Boost algorithm for counting the maximal cliques, to count all cliques in the graph. This was done by deleting the back-tracking part of the algorithm. This way the recursion gets to all the "leaves" of the tree, and therefore counts all the cliques combinations. The original maximal cliques counting algorithm, backtrack back up when it recognizes that there are no maximal cliques down the branch, but now we are interested in all the cliques, not only the maximal ones.

Simulations based on the Bipartite Model

For each real-world network, we simulated another network with the same number of nodes and the same number of edges as in the original one, based on the Bipartite model (M. E. J. Newman, Watts, & Strogatz, 2002).

We first created a bipartite network with V nodes in one partition and N nodes in the other partition. We generated edges between the nodes in the different partitions, such that the probability of an edge between a pair of nodes in the two partitions is: $p(edge) = \sqrt{\frac{2e}{v^2 N_{hidden}}}$ where e is the number of edges in the original network, v is the number of nodes in the original network and N_{hidden} is the number of nodes in the hidden partition.

We then transformed the bipartite network into a unipartite network by connecting all nodes in the first partition that have a common neighboring node in the second partition. We then counted the number of cliques in the final unipartite network. In addition, we generated a shuffled version for that generated network (see "Network shuffling") and counted the number of cliques in the shuffled network. We looked for the best fit to the number of cliques in the original and shuffled networks, where the free parameter was the number of nodes in the second partition.

Simulations based on the Hierarchical Model

For each real-world network, we simulated another network with the same number of nodes and the same number of edges as in the original one, based on the hierarchical model (Kleinberg, 2002; Watts, Dodds, & Newman, 2002). We first created a binary tree whose leaves are the nodes in the network graph. The probability for an edge between two nodes i and j (which are two leaves in the binary tree) is proportional to $e^{-\alpha \cdot LCA(i,j)}$, where LCA is the height of their lowest common ancestor in the tree, and alpha is a free parameter. Once the network has been created, we counted the number of cliques in the resulting network. In addition, we generated a shuffled version of the same network (see "Network shuffling"), and counted the number of cliques in the shuffled network as well. We looked for the best fit to the number of cliques in the original and shuffled network, with alpha as a free parameter.

Simulations based on the Gravitation Model

For each real-world network, we simulated another network, with the same number of nodes and the same number of edges as in the original one, based on the gravitation model (R. Itzhack

& Louzoun, 2010; Kalveram, 1992; Zhang & Jarrett, 1998). In a network with V nodes, each node i is assigned a random location $X_i(\mu)$ from a given distribution (exponential or Gaussian) with a mean distribution variable μ in the exponential case or with a zero mean and standard deviation distribution variable of μ in the Gaussian case. The probability for an edge to exist between node i and node j is proportional to $e^{-\propto|X_i-X_j|}$. In addition, we generated a shuffled version of the same network (see "Network shuffling"), and counted the number of cliques in the shuffled network as well. We looked for the best fit to the number of cliques in the original and shuffled network, with alpha and μ as free parameters. The graphs generated by the gravitation model are actually, when choosing the parameters adequately, unit disk graphs. The problem of clique partitioning in unit disk graphs is already discussed and has some fast approximations (Dumitrescu & Pach, 2011). However, in our case there was no need to use these algorithms, since given the size of the network used, the Bron Kerbosch algorithm is rapid enough.

Comparison between toy models and real networks

In order to compare each model with the observed clique distribution, we defined a cost function to be the sum of squares of the difference between the log of the original network's clique distribution and the log of the simulated network's clique distribution plus the sum of squares between the log shuffled curves of the two networks (original and simulated). We minimized the cost function and found the optimal value(s) of the parameters in the simulated network(s), which gives the minimal cost. Since the number of cliques of different networks even with the same parameter can vary widely, we averaged the number of cliques of 20 runs and calculated the error of the averaged number of cliques.

After finding the best network, other network properties were evaluated: the degree distribution, the distance distribution (using the Complex Networks Package for MatLab (Royi Itzhack et al., 2010)) and number of motifs. In order to measure the “similarity” of each of the networks generated by each model to the original network, we compared the differences between the degree/distance distributions as in Eq. S1:

(S1) $$diff(model) = \sum \left(\log_{10}(p_{orig} + 0.00001) - \log_{10}(p_{model} + 0.00001) \right)^2$$

where p_{orig} is the degree/distance distribution for the original network and p_{model} is the degree/distance distribution for the simulated network.

Un-directed motif count

We have checked the number of undirected motifs (Kashtan, Itzkovitz, Milo, & Alon, 2004) of sizes 3 and 4 in the networks. We have counted the number of instances of every motif in the original network (Royi Itzhack, Mogilevski, & Louzoun, 2007), and compared it to the number of motifs found in the network generated by each of the models described (Bipartite, Hierarchical, Gravitation).

Supplementary tables

Table s1. List of the networks used.

Name of network	Number of nodes	Average degree	Description of the network.
Political books	105	8.4	A network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent copurchasing of books by the same buyers. http://www-personal.umich.edu/~mejn/netdata/polbooks.zip
Word adjacencies (M. E. J. Newman, 2006)	112	7.6	adjacency network of common adjectives and nouns in the novel *David Copperfield* by Charles Dickens.
CEneural (Watts & Strogatz, 1998)	297	14.5	A directed, weighted network representing the neural network of C. Elegans.
Les Miserables (Knuth, 1993)	77	6.6	coappearance network of characters in the novel *Les Miserables*.
Florida (Ulanowicz, Bondavalli, & Egnotovich, 1998)	128	32.4	Food Web data collection (http://vlado.fmf.uni-lj.si/pub/networks/data/bio/foodweb/foodweb.htm).
Foldoc (Batagelj,	13,356	13.7	Foldoc is a searchable dictionary. In the network,

Mrvar, & Zaveršnik, 2002a, 2002b)			an arc (X,Y) from term X to term Y exists in the network iff in the FOLDOC dictionary the term Y is used to describe the meaning of term X (http://vlado.fmf.uni-lj.si/pub/networks/data/dic/foldoc/foldoc.htm).
PairsP (Nelson, McEvoy, & Schreiber, 1998)	10,617	12	Free Associations norms (cue X is associated with target Y).
eatSR (Kiss, Armstrong, Milroy, & Piper, 1973)	23,218	26.3	The Edinburgh Associative Thesaurus (EAT) is a set of word association norms showing the counts of word association as collected from subjects. http://monkey.cis.rl.ac.uk/Eat/htdocs/eat.zip
American College Football (Girvan & Newman, 2002)	117	10.7	Network of American football games between Division IA colleges during regular season Fall 2000.
CEmeta (Duch & Arenas, 2005)	453	9	List of edges of the metabolic network of C.elegans.
political blogs(Adamic & Glance, 2005)	1224	27.3	A directed network of hyperlinks between weblogs on US politics, recorded in 2005 by Adamic and Glance. Please cite L. A. Adamic and N. Glance, "The political blogosphere and the 2004 US Election", in Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem (2005).

Autonomous systems (M. Newman)	22963	4.2	A symmetrized snapshot of the structure of the Internet at the level of autonomous systems, reconstructed from BGP tables posted by the University of Oregon Route Views Project. This snapshot was created by Mark Newman from data for July 22, 2006 and is not previously published.
High energy theory collaborations (M. E. J. Newman, 2001)	7610	4.1	Weighted network of coauthorships between scientists posting preprints on the High-Energy Theory E-Print Archive between January 1, 1995 and December 31, 1999.

Table s2. Quantification of similarity of the different attributes checked (number of all cliques, number of maximal cliques, distance distribution, degree distribution and connectivity distribution) for the different models and for an Erdős–Rényi network with the same number of nodes and the same number of edges. The similarity for number of cliques/maximal cliques is $\sum\left(\log_{10}(Ncliques_{currModel}+1)-\log_{10}(Ncliques_{original}+1)\right)^2$. The similarity for distance, degree and connectivity distributions is $\sum\left(\log_{10}(Value_{currModel}+0.00001)-\log_{10}(Value_{original}+0.00001)\right)^2$. The sum was performed on 20 logarithmic bins in the case of the distance and in the case of the degree and on 21 linear bins (0 to 1 in jumps of 0.05) in the case of the connectivity.

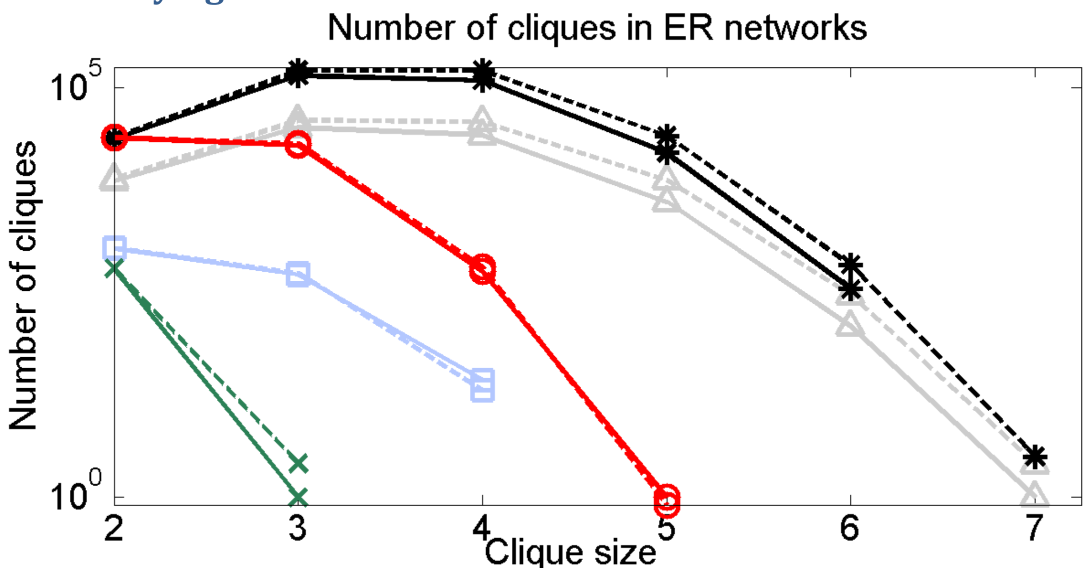
		exp	gauss	bipartite	hierarchical	ER
CEmeta	cliques	38.12445	0.526036	3.903792	31.80707635	41.8841
	maximal cliques	4.204133	0.965977	2.25388	16.83305905	17.00752
	distance	54.3227	279.0178	0.417263	136.2590108	3.492258
	degree	125.3849	154.8057	120.294	356.5183431	156.2681
	connectivity	67.63719	100.5453	15.82171	0	0
CEneural		exp	gauss	bipartite	hierarchical	ER
	cliques	0.222053	0.878272	3.441292	7.693195753	23.39759
	maximal cliques	0.639551	1.239018	1.467425	8.385082828	12.994
	distance	201.0002	227.678	0.05767	89.41065109	2.244398
	degree	89.30931	89.10089	120.4451	327.5917488	229.7808
	connectivity	76.65822	76.5002	8.011063	0	0
lesmis		exp	gauss	bipartite	hierarchical	ER
	cliques	6.03727	4.363946	12.17891	15.85587523	33.88578

	maximal cliques	0.547642	0.624482	0.523501	13.05641655	2.782671
	distance	144.4159	145.7518	1.411752	125.2473688	0.246215
	degree	58.64144	58.81799	64.81995	258.697885	134.6167
	connectivity	92.39437	97.88812	20.85202	0	0
polbooks		exp	gauss	bipartite	hierarchical	ER
	cliques	3.20421	0.24494	4.725936	4.794936831	9.113107
	maximal cliques	2.264349	0.240985	0.764138	4.671503487	5.343049
	distance	0.160112	79.26078	51.61488	124.542932	53.0961
	degree	60.82552	63.94694	60.81471	199.1013798	55.18073
	connectivity	78.49295	79.08468	0	0	0

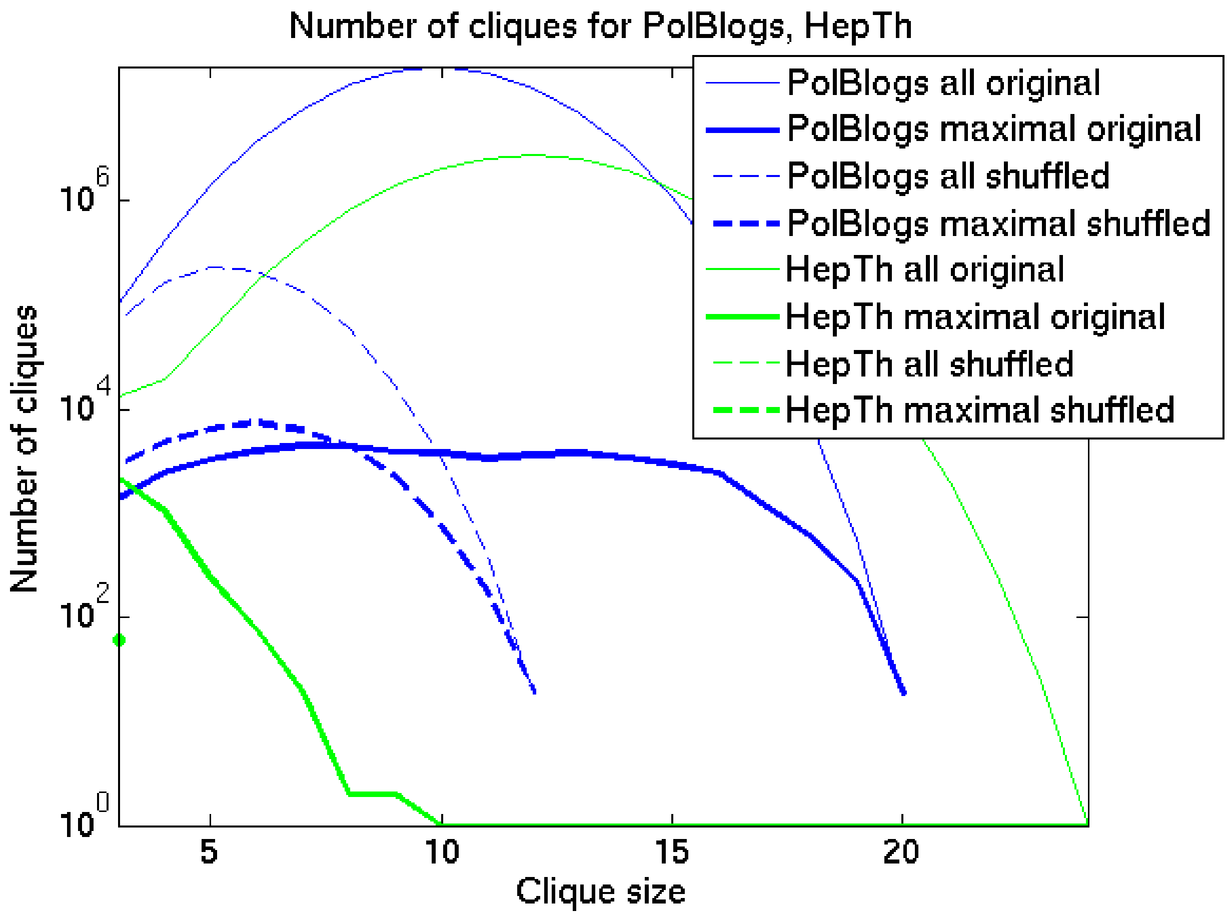
Table s3. The assortativity (the correlation between the degrees of neighboring nodes) for the original networks, for the networks generated by each of the models and for an Erdős–Rényi network with the same number of nodes and the same number of edges.

	original	exp	gauss	bipartite	hierarchical	ER
CEmeta	-0.19395	-0.05669	0.659752	0.054955	-0.00698	-0.02991
CEneural	-0.10679	0.639954	0.642821	0.056169	-0.00507	0.001562
lesmis	-0.01147	0.474783	0.540804	-0.07357	-0.03007	0.108019
polbooks	-0.10411	0.306476	0.492816	0.034314	-0.50206	-0.02696

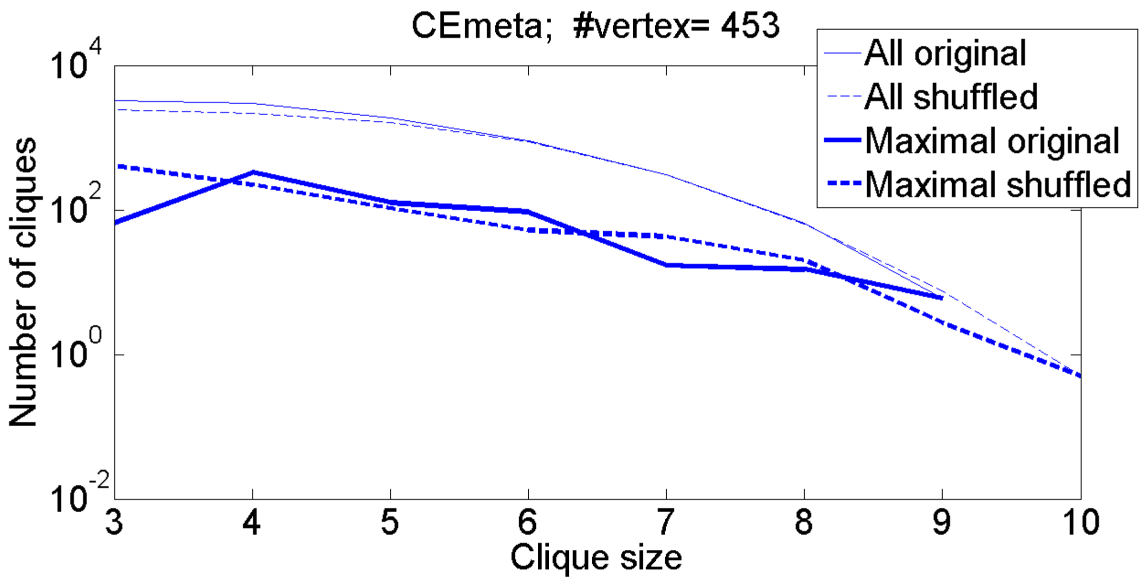
Supplementary figures



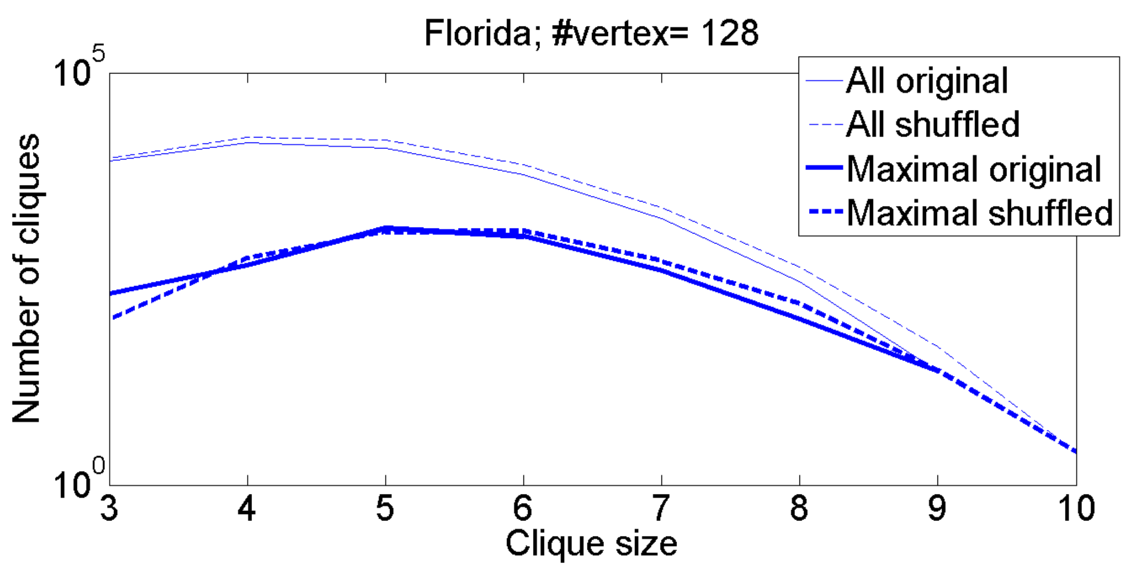
S1. Number of cliques counted by the algorithm we used for various ER networks (solid lines) compared to the expected number of cliques (dashed lines). Blue (squares): v=150, p=0.1; gray(triangles): v=250, p=0.25; black(stars): v=500, p = 0.2; green('x' signs): v=500, p=0.005; red(circles): v=1000,p=0.05.



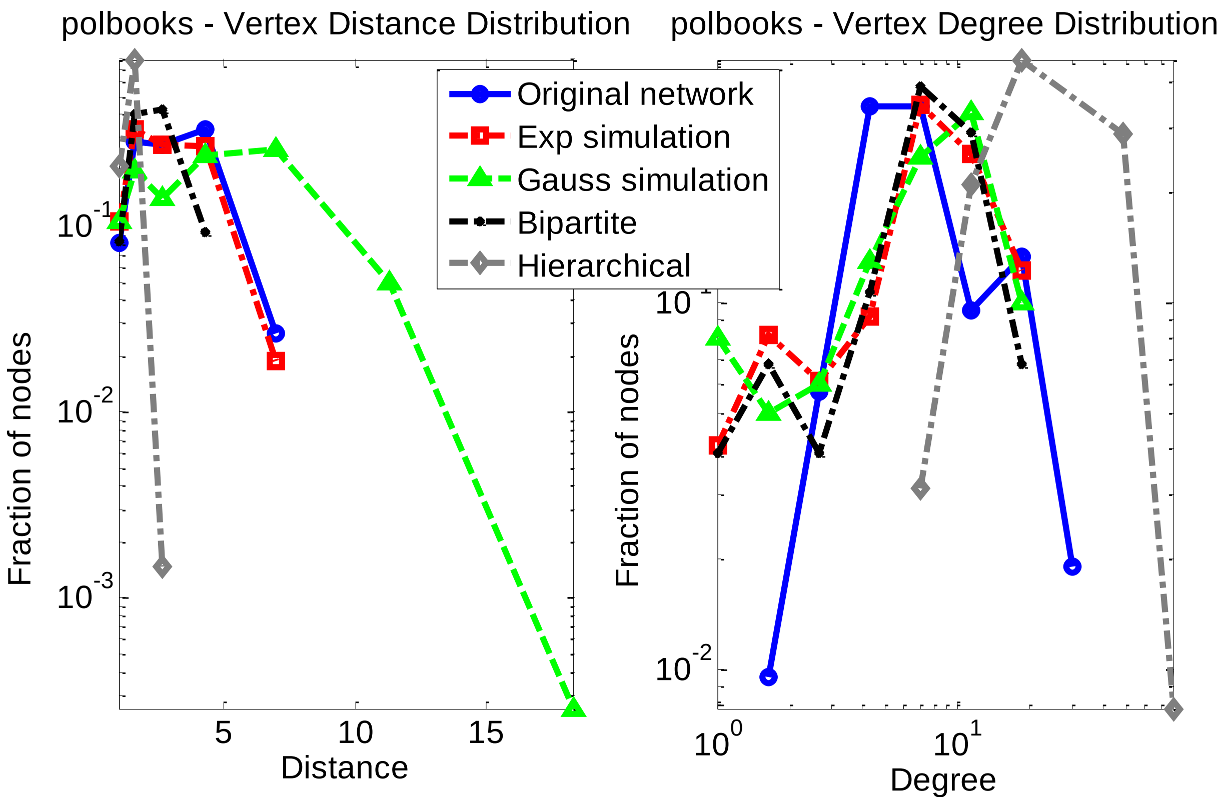
S2. Number of all cliques in the original PolBlogs network (1224 nodes, blue solid lines) and in its counterpart shuffled network (blue dashed line); number of all cliques in the original HepTh network (7610 nodes, green solid lines) and in its counterpart shuffled network (green dashed line). All cliques in thin lines; maximal cliques in thick lines.



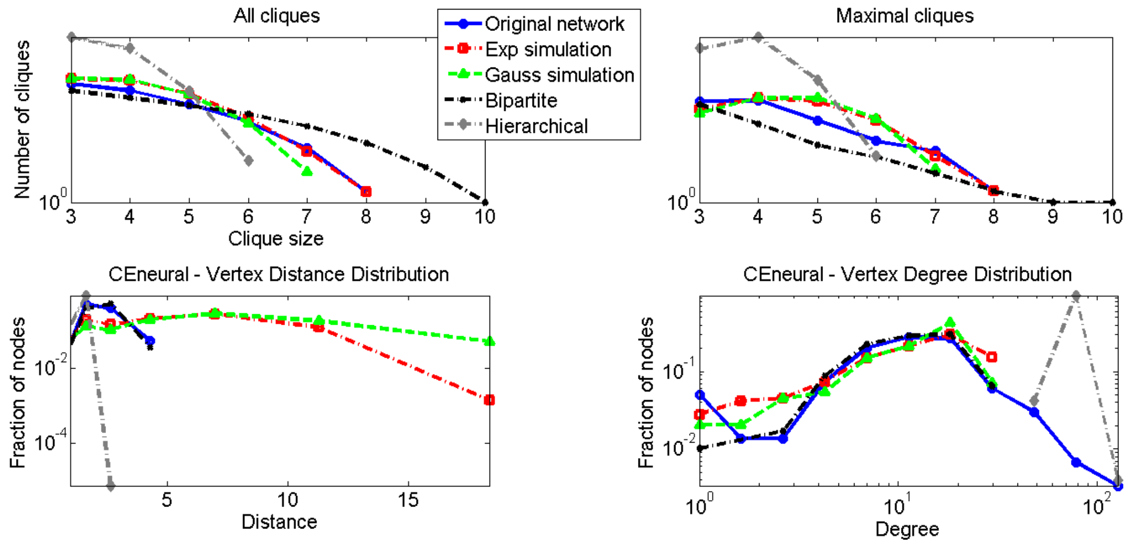
S3. Number of all cliques in the original CEmeta network (453 nodes, solid lines) and in its counterpart shuffled networks (dashed line). All cliques in thin lines; maximal cliques in thick lines.



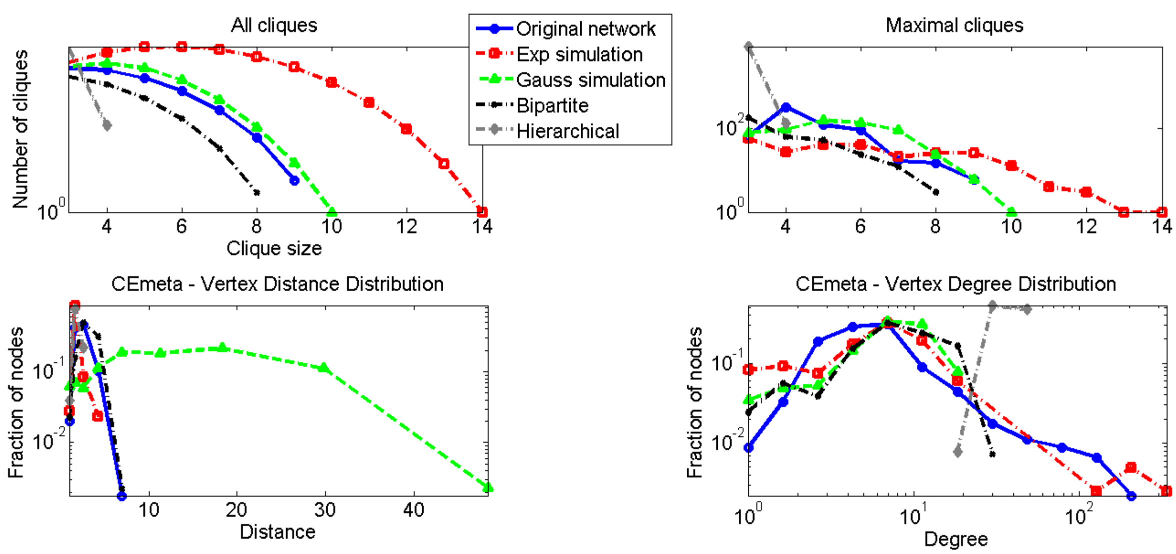
S4. Number of all cliques in the original Florida network (128 nodes, solid lines) and in its counterpart shuffled networks (dashed line). All cliques in thin lines; maximal cliques in thick lines.



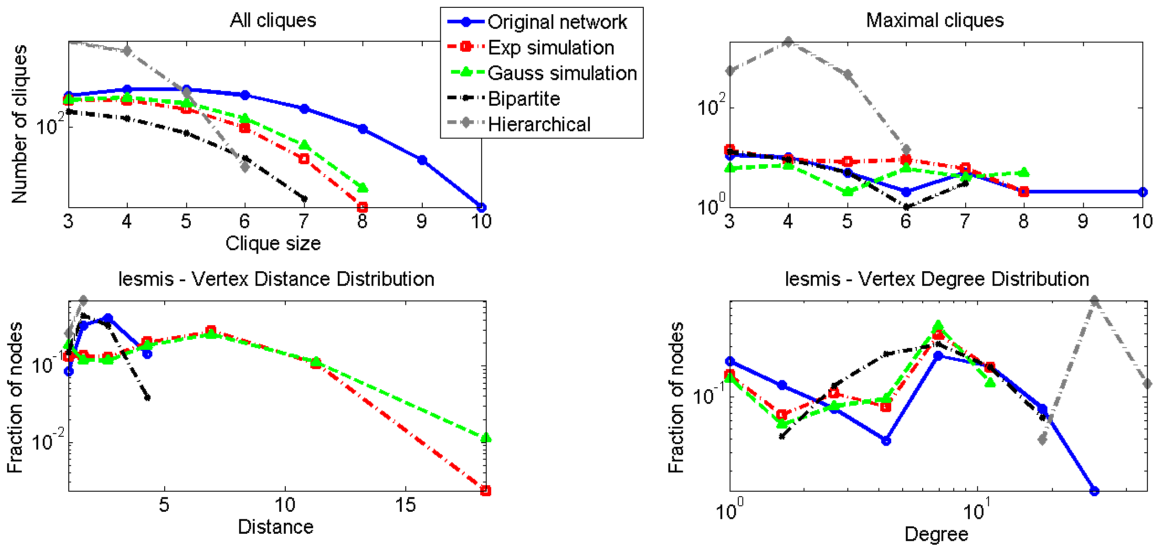
S5. Results of Gravitation model (both Exponential and Gaussian simulations) as well as for bipartite and hierarchical models for the PolBooks network (104 nodes, 416 edges): Left drawing: Shortest distance distributions in the different networks. Right drawing: Degree distribution of the nodes in the different networks.



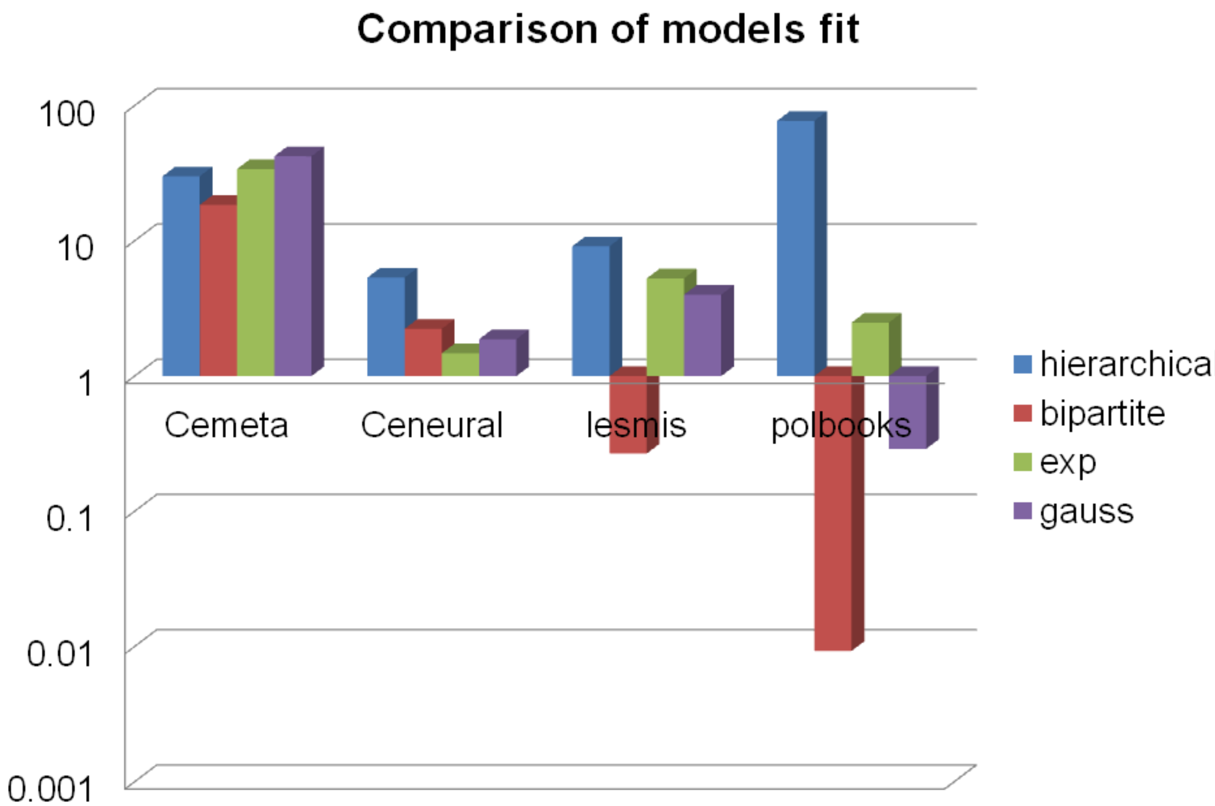
S6. Results of Gravitation model (both Exponential and Gaussian simulations) as well as for bipartite and hierarchical models for the CEneural network (296 nodes, 2072 edges): Number of cliques, number of maximal cliques, shortest distance distribution and degree distribution.



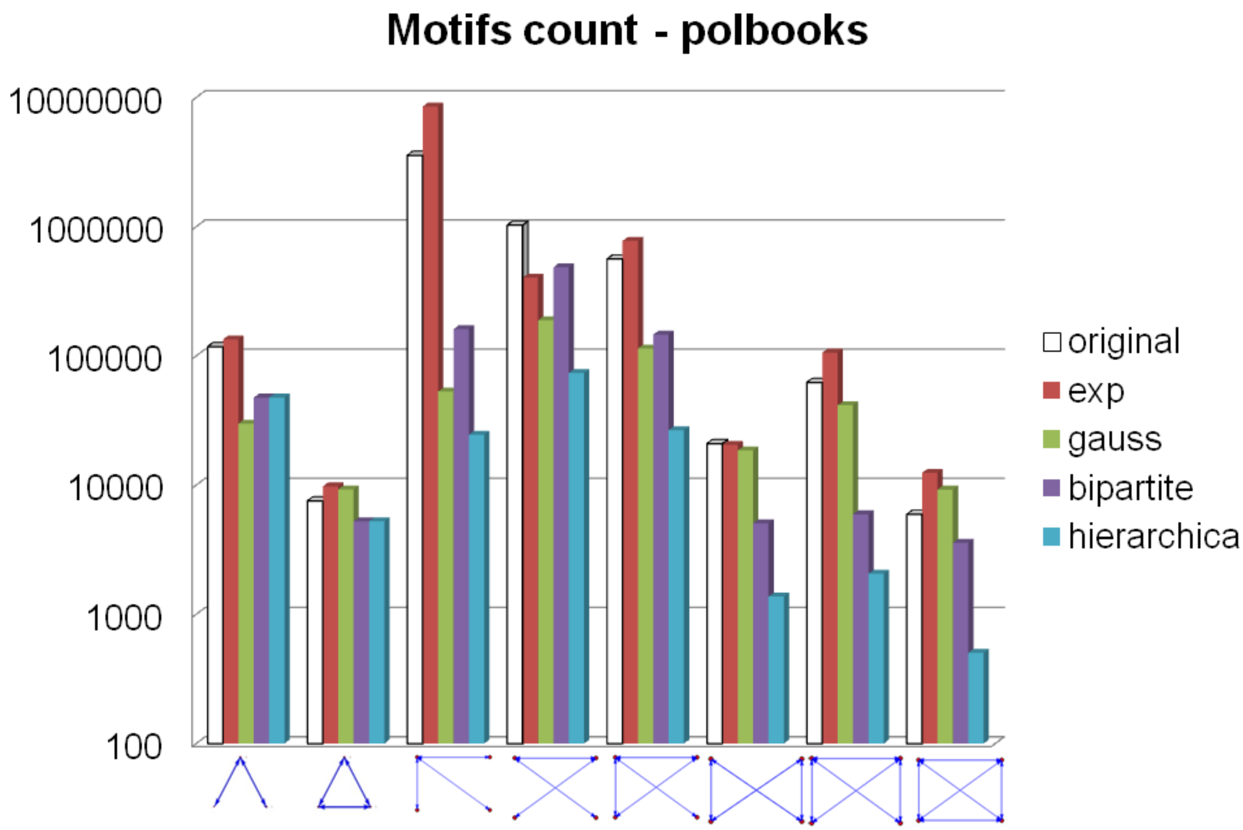
S7. Results of Gravitation model (both Exponential and Gaussian simulations) as well as for bipartite and hierarchical models for the CEmeta network (453 nodes, 1812 edges): Number of cliques, number of maximal cliques, shortest distance distribution and degree distribution.



S8. Results of Gravitation model (both Exponential and Gaussian simulations) as well as for bipartite and hierarchical models for the lesmis network (76 nodes, 228 edges): Number of cliques, number of maximal cliques, shortest distance distribution and degree distribution.



S9. Error values (according to Eq. 5) for the differences in number of cliques between original networks and the networks generated by each one of the models. The smallest error is either for the gravitation model or for the bipartite model.



S10. Comparison of number of motifs of sizes 3 and 4, for the original polbooks network as well as for networks generated by the gravitation model (using either an Exponential or a Gaussian distribution), bipartite model and hierarchical model. The number of motifs for the exponential gravitation model is the closest to the number of motifs in the original real world network.

Bibliography

Adamic, L. A., & Glance, N. (2005). *The political blogosphere and the 2004 US election: divided they blog*.

Batagelj, V., Mrvar, A., & Zaveršnik, M. (2002a). *Network analysis of dictionaries*: Univ. of Ljubljana, Inst. of Mathematics, Physics and Mechanics, Dep. of Theoretical Computer Science.

Batagelj, V., Mrvar, A., & Zaveršnik, M. (2002b). *Network analysis of texts*: Univ. of Ljubljana, Inst. of Mathematics, Physics and Mechanics, Dep. of Theoretical Computer Science.

Boost C++ Source Libraries.

Bron, C., & Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM, 16*(9), 575-577.

Cazals, F., & Karande, C. (2008). A note on the problem of reporting maximal cliques. *Theoretical Computer Science, 407*(1), 564-568.

Duch, J., & Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Physical Review E, 72*(2), 027104.

Dumitrescu, A., & Pach, J. (2011). Minimum clique partition in unit disk graphs. *Graphs and Combinatorics, 27*(3), 399-411.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821.

Itzhack, R., & Louzoun, Y. (2010). Random distance dependent attachment as a model for neural network generation in the Caenorhabditis elegans. *Bioinformatics, 26*(5), 647.

Itzhack, R., Mogilevski, Y., & Louzoun, Y. (2007). An optimal algorithm for counting network motifs. *Physica A: Statistical Mechanics and its Applications, 381*, 482-490.

Itzhack, R., Muchnik, L., Erez, T., Tsaban, L., Goldenberg, J., Solomon, S., & Louzoun, Y. (2010). Empirical extraction of mechanisms underlying real world network generation. *Physica A: Statistical Mechanics and its Applications, 389*(22), 5308-5318.

Kalveram, K. T. (1992). A neural network model rapidly learning gains and gating of reflexes necessary to adapt to an arm's dynamics. *Biol Cybern, 68*(2), 183-191.

Kashtan, N., Itzkovitz, S., Milo, R., & Alon, U. (2004). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics, 20*(11), 1746-1758.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). *An associative thesaurus of English and its computer analysis*: The computer and literary studies. Edinburgh: University Press.

Kleinberg, J. (2002). Small-world phenomena and the dynamics of information. *Advances in neural information processing systems, 1*, 431-438.

Knuth, D. E. (1993). *The Stanford GraphBase: a platform for combinatorial computing*: AcM Press.

Koch, I. (2001). Enumerating all connected maximal common subgraphs in two graphs. *Theoretical Computer Science, 250*(1-2), 1-30.

Nelson, D., McEvoy, C., & Schreiber, T. (1998). The University of South Florida word association, rhyme, and word fragment norms. http. *w3. usf. edu/FreeAssociation*.

Newman, M. Network data, from http://www-personal.umich.edu/~mejn/netdata/

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, 98*(2), 404.

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E, 74*(3), 036104.

Newman, M. E. J., Watts, D. J., & Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America, 99*(Suppl 1), 2566-2572.

Tomita, E., Tanaka, A., & Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science, 363*(1), 28-42.

Ulanowicz, R., Bondavalli, C., & Egnotovich, M. (1998). Network analysis of trophic dynamics in south florida ecosystem, FY 97: The florida bay ecosystem. *Ref. CBL98-123. Chesapeake Biological Laboratory, Solomons, MD, USA*.

Watts, D. J., Dodds, P. S., & Newman, M. E. J. (2002). Identity and search in social networks. *Science, 296*(5571), 1302-1305.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’networks. *Nature, 393*(6684), 440-442.

Zhang, X., & Jarrett, D. F. (1998). Chaos in a dynamic model of traffic flows in an origin-destination network. *Chaos, 8*(2), 503-513.