Supplementary Information for the paper:

Deciphering the global organization of clustering in real complex networks

Pol Colomer-de-Simón[1], M. Ángeles Serrano[1], Mariano
G. Beiró[2], J. Ignacio Alvarez-Hamelin[2], & Marián Boguñá[1]

[1]*Departament de Física Fonamental, Universitat de Barcelona,*
Martí i Franquès 1, 08028 Barcelona, Spain and

[2]*INTECIN (CONICET–U.B.A.), Facultad de Ingeniería,*
Universidad de Buenos Aires, Paseo Colón 850, C1063ACV Buenos Aires, Argentina

(Dated: July 19, 2013)

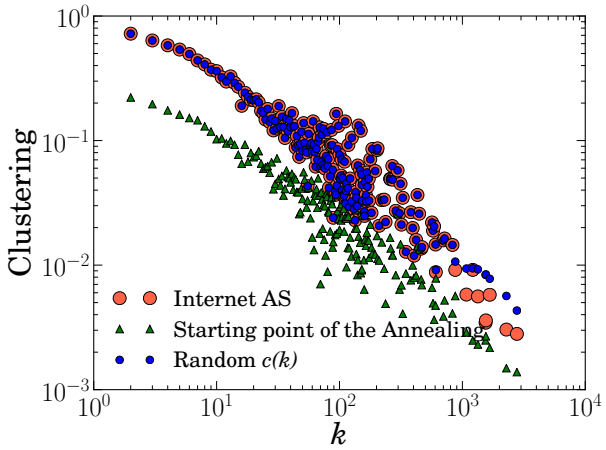

FIG. SI-1: Comparison of the clustering spectrum between the real Internet AS network, the randomized version from where we started the annealing process, and the maximally random model.

I. DATA SETS

A. Internet AS

We use the autonomous system Internet topology of June 2009 extracted from data collected by the archipelago active measurement infrastructure developed by the Cooperative Association for Internet Data Analysis [1, 2]. The AS topology contains 23752 ASs and 58416 AS links, yielding the average AS degree $\bar{k} = 4.92$, clustering coefficient $\bar{C} = 0.51$ and maximum degree $k_{max} = 2778$. Its clustering spectrum is represented in figure SI-1.

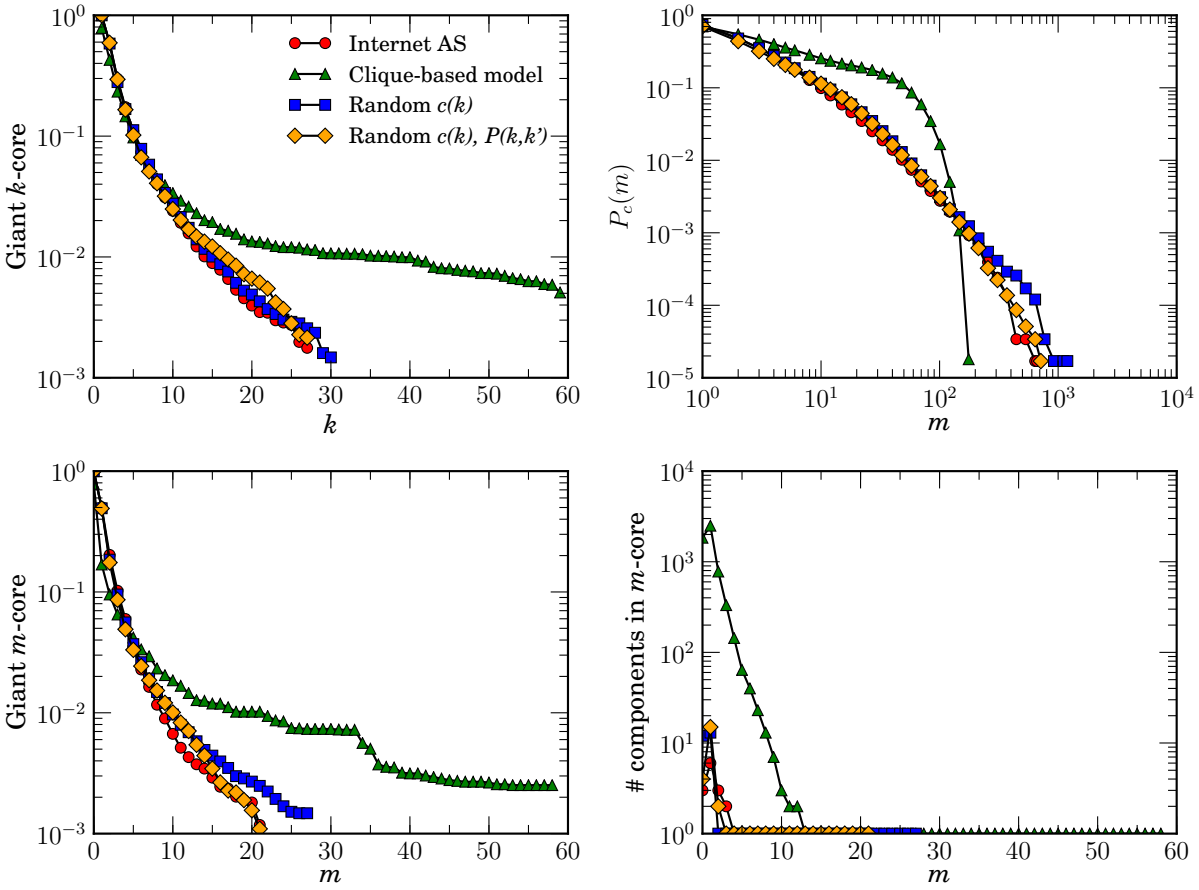

FIG. SI-2: Comparison of the k-core and m-core decompositions between the real Internet AS network, the clique based model, and maximally random models.

B. Pretty-Good-Privacy (PGP)

Pretty-Good-Privacy (PGP) is the most popular encryptor algorithm aimed to maintain privacy in communication between peers in internet. This algorithm makes use of a pair of keys, one of them to encrypt the message, and its counterpart to decrypt the message. Both keys are generated in such a way that it is computationally infeasible to deduce one key from the other. Provided that everyone can generate a PGP key by himself, if anybody wants to know if a given key belongs really to the person stated in the key, he has to verify that. Hence exists a ”signing procedure” where a person signs the public key of another, meaning that she trusts that the other person is who she claims to be. This procedure generates a web of peers that have signed public keys of another based on trust, and this is the so- called web of trust of PGP [3].

Here, we analyze the web of trust as it was on July 2001, when it comprised 191.548 keys and

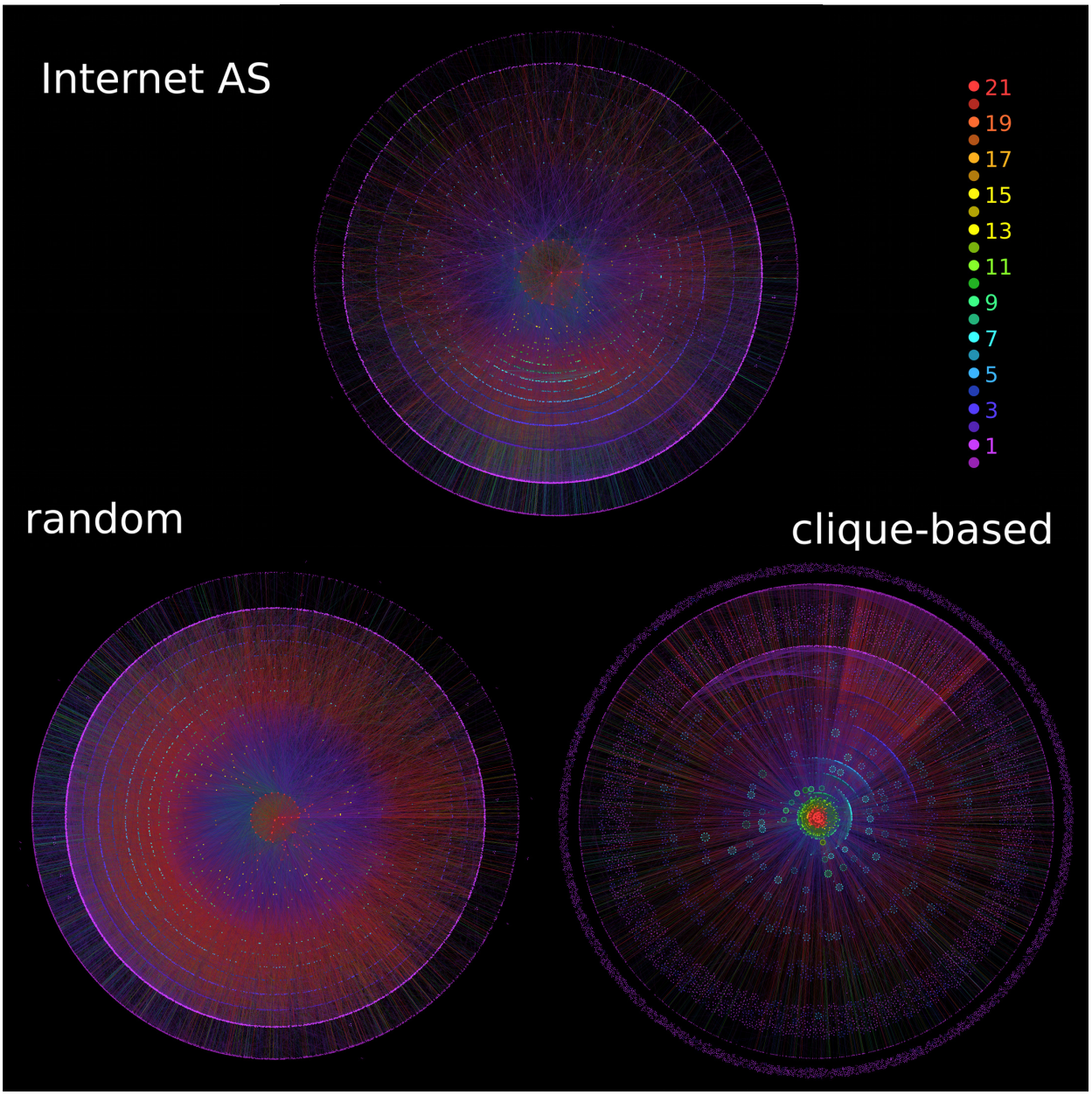

FIG. SI-3: m-core decomposition of the Internet network and its random versions.

286.290 signatures. Since we are mainly interested in the social character of the web of trust we only consider bidirectional signatures, i.e., peers who have mutually signed their keys. This filtering process guarantees mutual knowledge between connected peers and makes the PGP network a reliable proxy of the underlying social network. After the filtering process, we are left with an undirected network of 57.243 vertices, 61.837 edges, average degree $\bar{k} = 2.16$, clustering coefficient $\bar{C} = 0.50$ and maximum degree $k_{max} = 205$. Its clustering spectrum is represented in figure SI-4.

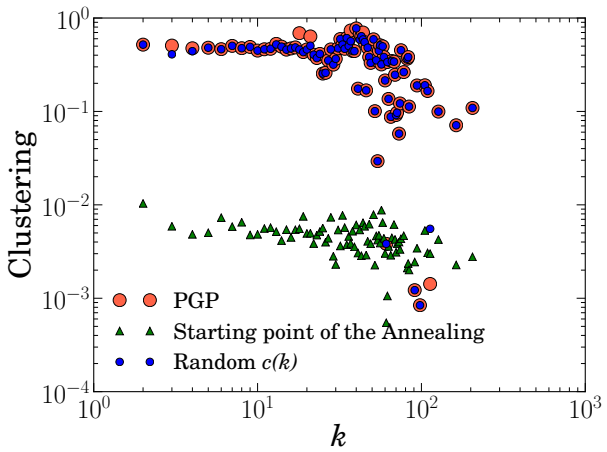

FIG. SI-4: Comparison of the clustering spectrum between the real PGP network, the randomized version from where we started the annealing process, and the maximally random model.

C. E. Coli

A simple abstraction of a given metabolism is given by its bipartite network representation. This amounts to consider metabolites and reactions as belonging to different subsets of nodes, with metabolites (irrespectively considered as reactants and products) linked to all reactions they take part in, thus avoiding connections between nodes of the same kind. Our network data set is the on-mode projection of the metabolism bipartite network of the bacteria Escherichia coli[4]. So nodes accounts for metabolites that are connected whenever they participate in the same reaction. The resulting network has 1.010 nodes, 3.286 edges, average degree $\bar{k} = 6.51$, clustering coefficient $\bar{C} = 0.48$ and maximum degree $k_{max} = 54$. Its clustering spectrum is represented in figure SI-7.

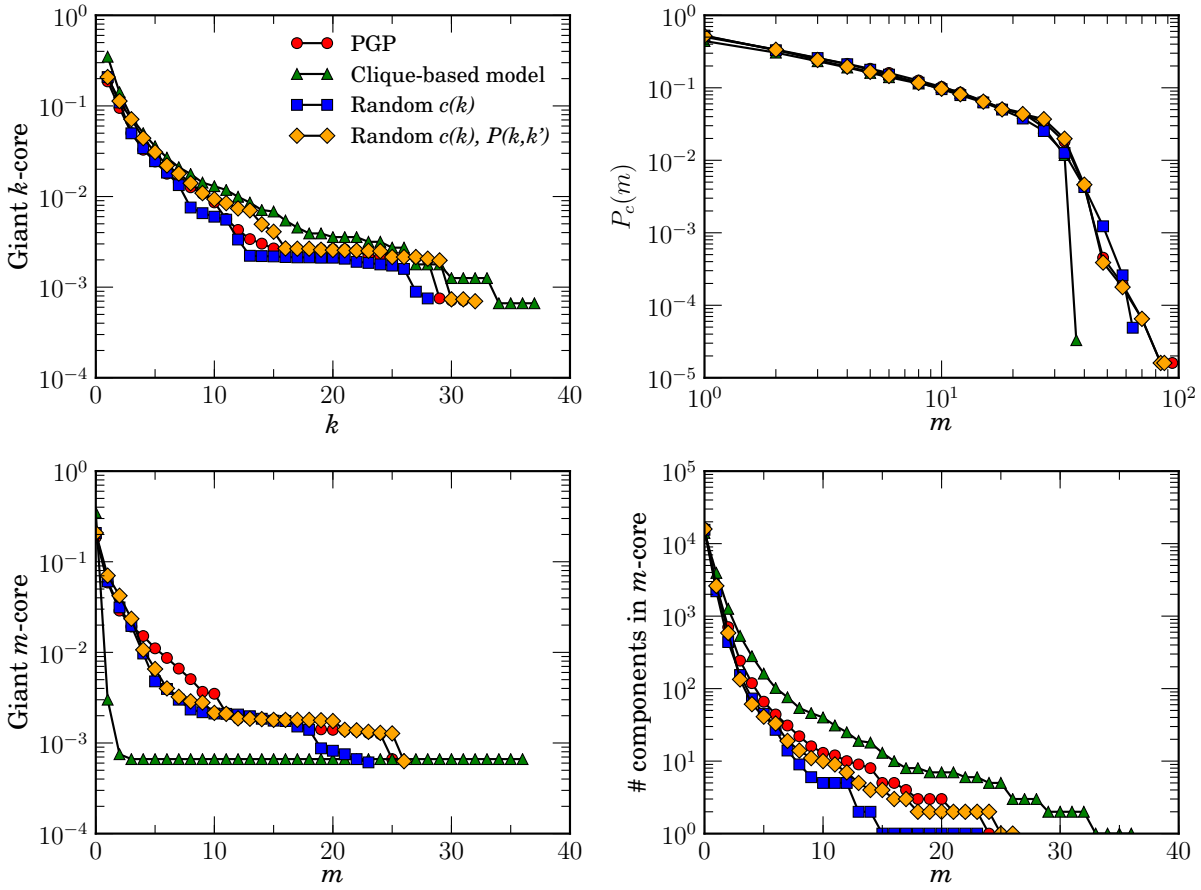

FIG. SI-5: Comparison of the k-core and m-core decompositions between the real PGP network, the clique based model, and maximally random models.

D. CondMat

Condense Matter (CondMat) Physics Arxiv collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers submitted to Condense Matter category. Nodes are authors and there is an edge between them if they co-authored at least one paper. The data covers papers in the period from January 1993 to April 2003 [5].

The Network has 23.133 nodes, 93.439 links, an average degree of $\bar{k} = 8.08$, a clustering coefficient of $\bar{C} = 0.71$ and a maximum degree of $k_{max} = 279$. Its clustering spectrum is represented in figure SI-10.

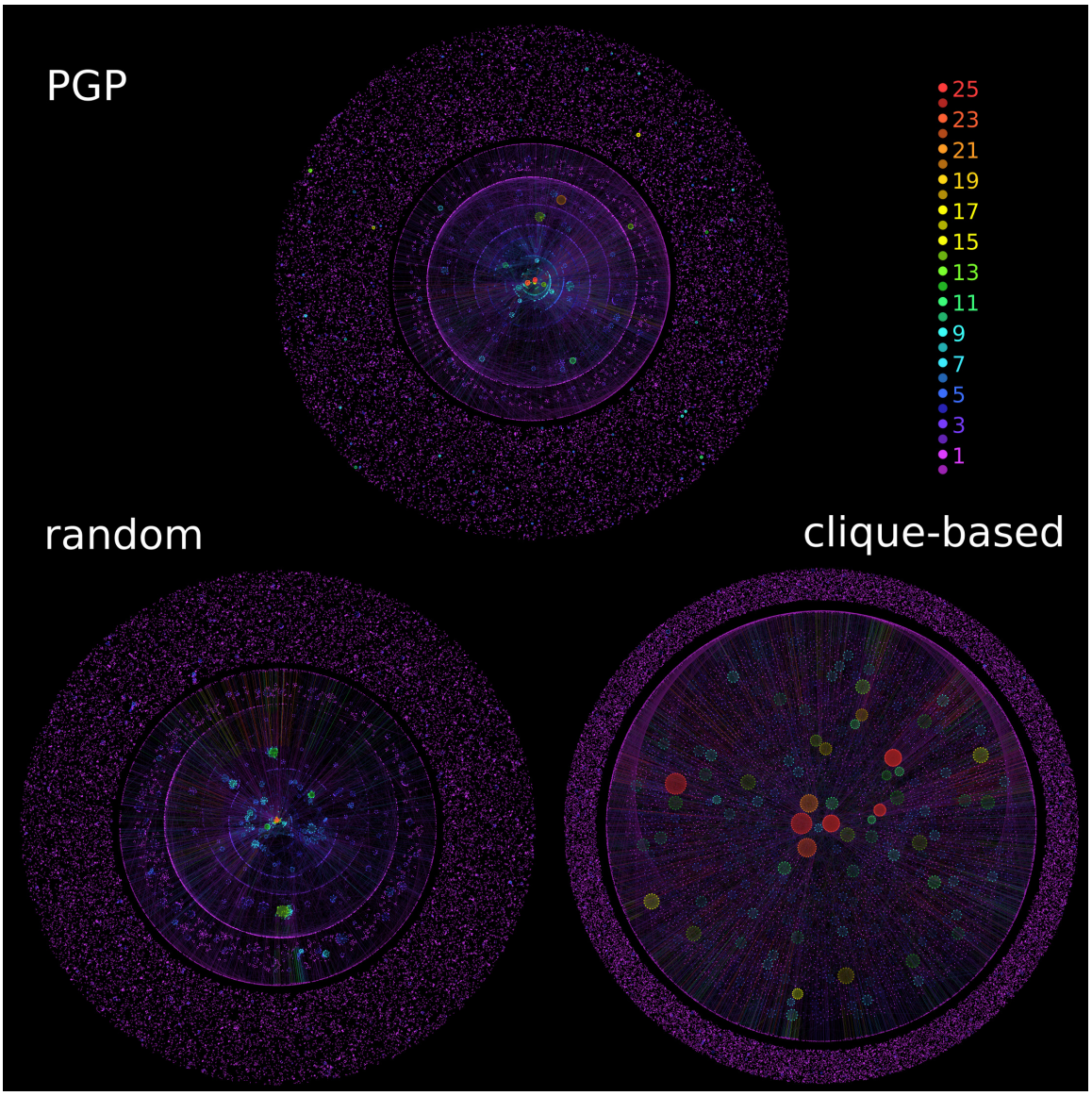

FIG. SI-6: m-core decomposition of the PGP network and its random versions.

E. Musical discourse

The dataset contains the beat-based music descriptions of the audio rendition of a musical piece or score. For pitch, these descriptions reflect the harmonic content of the piece, and encapsulate all sounding notes of a given time interval into a compact representation, independently of their articulation (they consist of the 12 pitch class relative energies, where a pitch class is the set of all pitches that are a whole number of octaves apart, e.g. notes C1, C2, and C3 all collapse to pitch class C). All descriptions are encoded into music codewords, using a binary discretization in the case of pitch. Codewords are then used to perform frequency counts, and as nodes of a complex

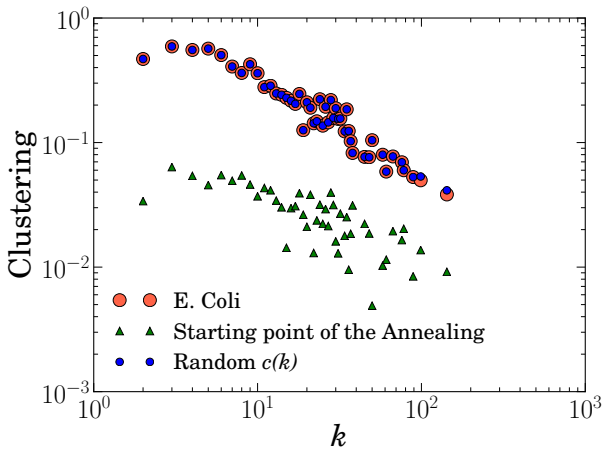

FIG. SI-7: Comparison of the clustering spectrum between the real E. Coli network, the randomized version from where we started the annealing process, and the maximally random model.

network whose links reflect transitions between subsequent codewords [6].

The resulting network have 2476 nodes, 20624 edges, an average degree of $\bar{k} = 16.66$, a clustering coefficient of $\bar{C} = 0.82$ and a maximum degree of $k_{max} = 1566$.Its clustering spectrum is represented in figure SI-13.

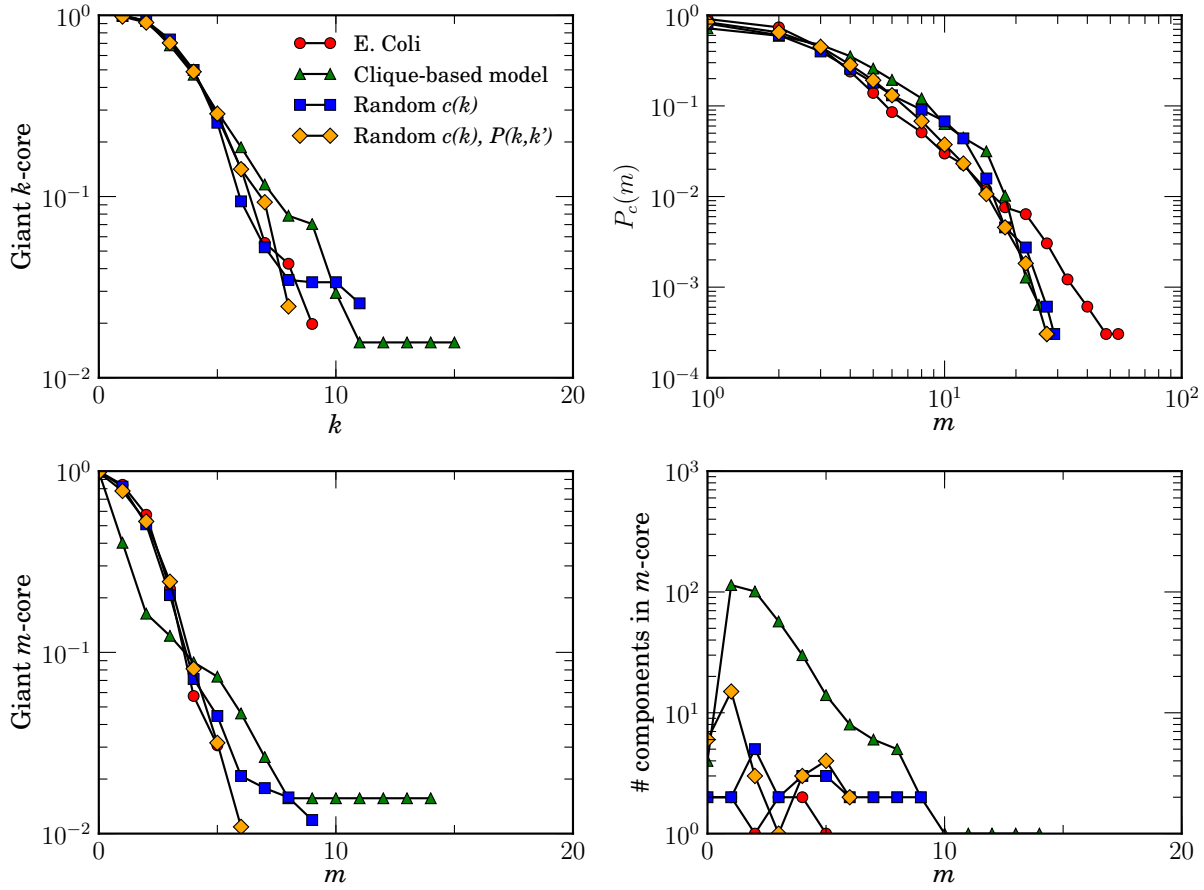

FIG. SI-8: Comparison of the k-core and m-core decompositions between the real E. Coli metabolism network, the clique based model, and maximally random models.

F. Email

An e-mail network can be built regarding each e-mail address as a node and linking two nodes if there is an e-mail communication between them. We take as a case study the e-mail network of University Rovira i Virgili (URV) in Tarragona, Catalonia, containing 1669 users [7]. After deleting the isolated nodes we have a network with 1144 nodes, 6004 edges, an average degree of $\bar{k} = 10.50$, a clustering coefficient of $\bar{C} = 0.27$ and a maximum degree of $k_{max} = 71$. Its clustering spectrum is represented in figure SI-16.

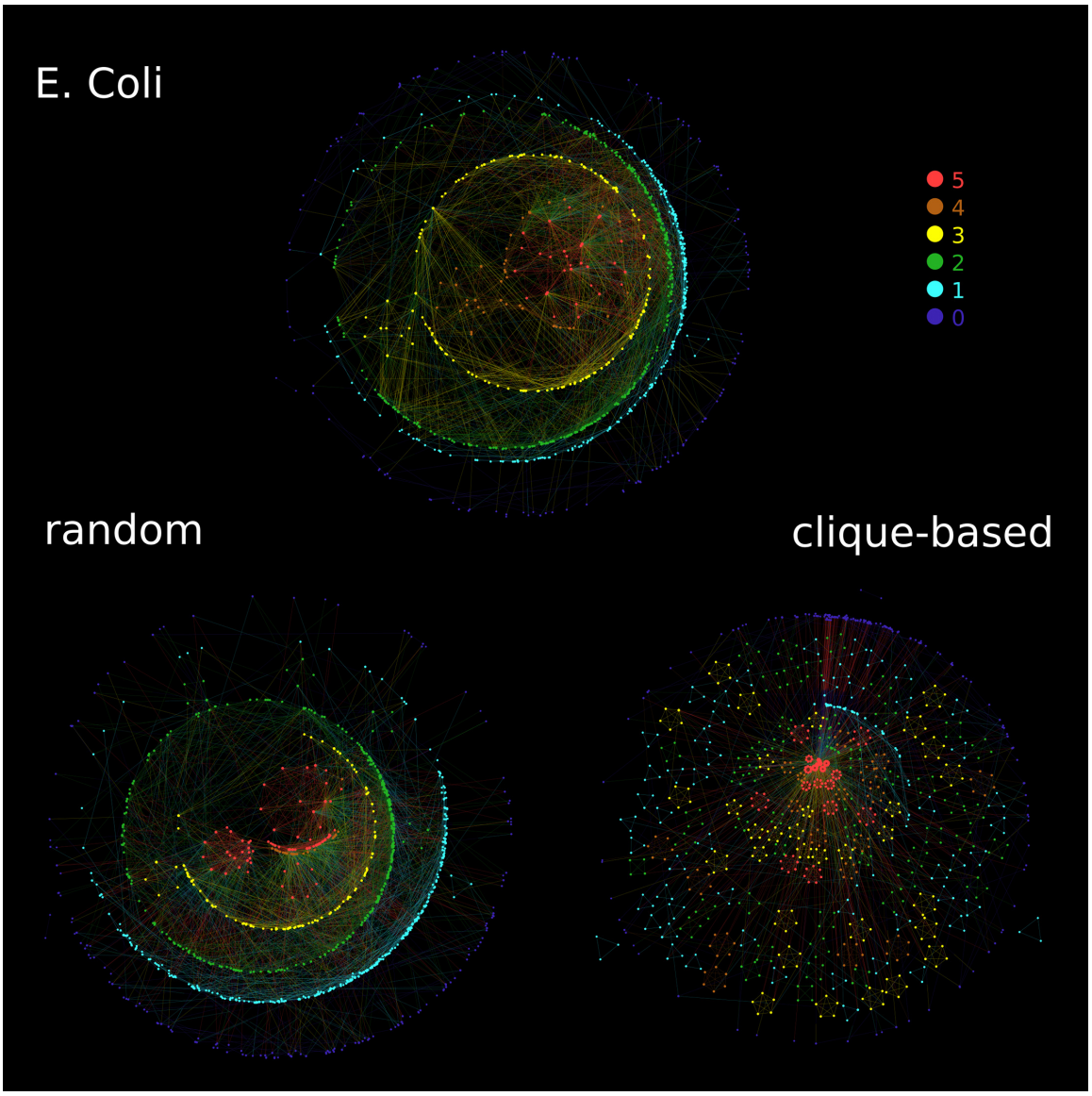

FIG. SI-9: m-core decomposition of the E. Coli metabolism network and its random versions.

G. Epinions

Epinions network is a who-trust-whom online social network of a consumer review site Epinions.com. Nodes are memebers of this site and an edge connects them if they "trust" each other [8]. The resulting network has 31100 nodes, 103097 edges, an average degree of $\bar{k} = 6.63$, a clustering coefficient of $\bar{C} = 0.24$ and a maximum degree of $k_{max} = 443$. Its clustering spectrum is represented in figure SI-19.

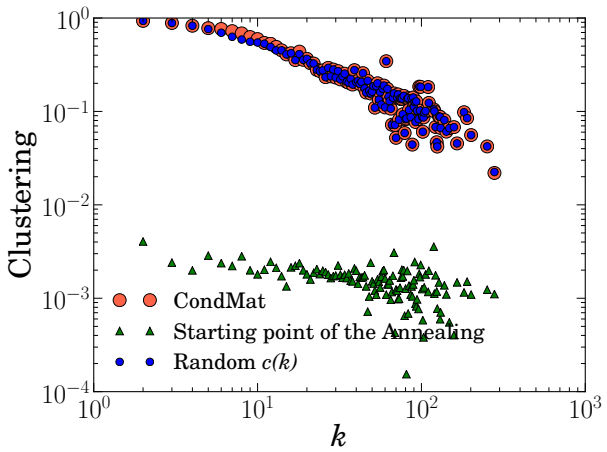

FIG. SI-10: Comparison of the clustering spectrum between the real CondMat network, the randomized version from where we started the annealing process, and the maximally random model.

H. Power grid

This power grid dataset corresponds to an undirected, unweighted network representing the topology of the Western States Power Grid of the United States [9]. The resulting network has 4941 nodes, 6594 edges, an average degree of $\bar{k} = 2.67$, a clustering coefficient of $\bar{C} = 0.11$ and a maximum degree of $k_{max} = 19$. Its clustering spectrum is represented in figure SI-22.

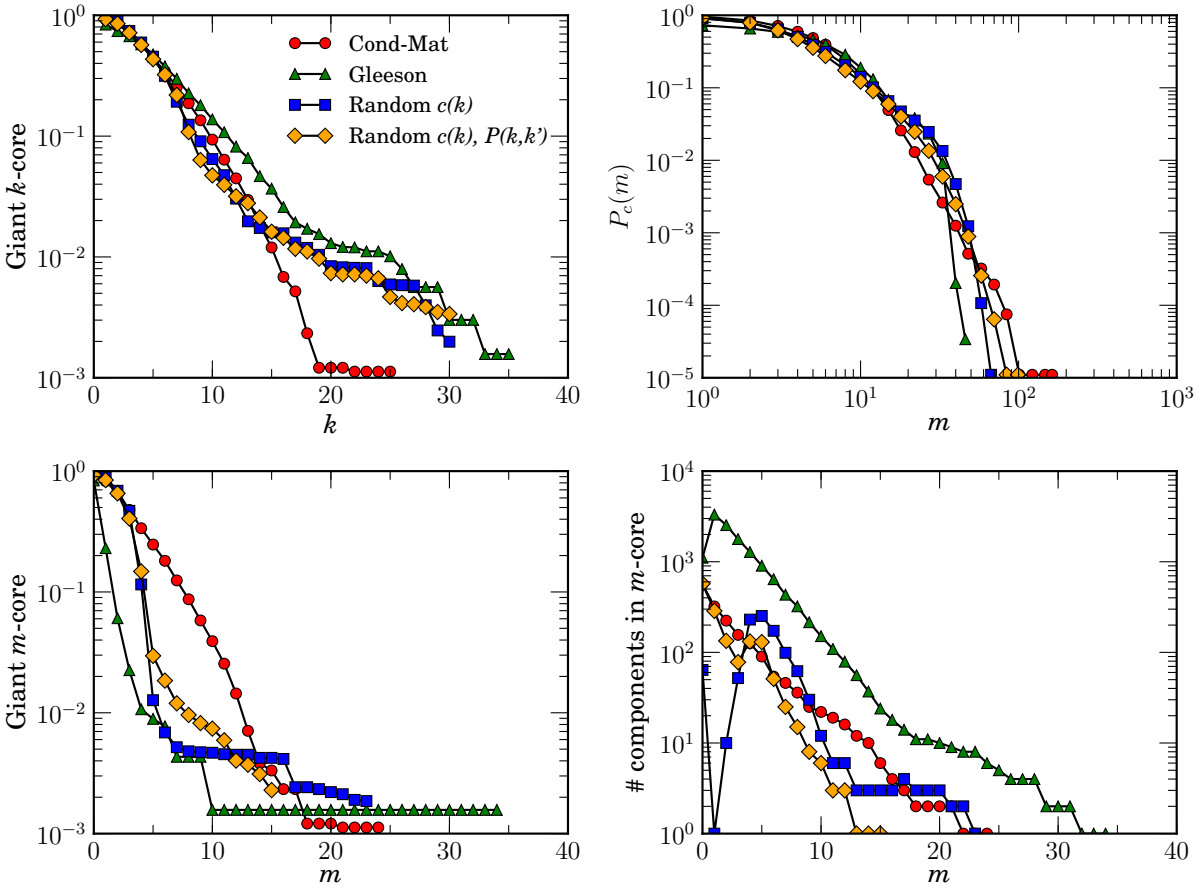

FIG. SI-11: Comparison of the k-core and m-core decompositions between the real CondMat network, the clique based model, and maximally random models.

II. m-CORES AND THEIR RELATION WITH THE k-CORE DECOMPOSITION

The m-core decomposition was introduced in[10] under the name of k-dense decomposition. In that paper, a procedure for obtaining the m-cores is described, which implies computing the k-cores of the graph several times. Here we introduce a more efficient procedure which allows us to obtain the m-cores at a cost similar to that of computing the clustering coefficients, about $\mathcal{O}(n \cdot m_{\max}^2)$, where $m_{\max}$ is the level of the maximal m-core.

Given a simple undirected network graph $G = (V, E)$ we construct an *hypergraph* $G^* = (V^*, E^*)$ whose nodes are the edges e_{ij} of the original graph, i.e., $V^* = E$. Now, for each triangle of connected vertices (v_i, v_j, v_k) in G, we add a 3-tuple (e_{ij}, e_{jk}, e_{ik}) in G^* (this construction has a reminiscence of the *triangular line graph* [11] in ordinary graphs). Given $v = e_{ij} \in E$, the multiplicity of this edge, $M_G(e_{ij})$ is the number of triangles in G in which it takes part. Then we

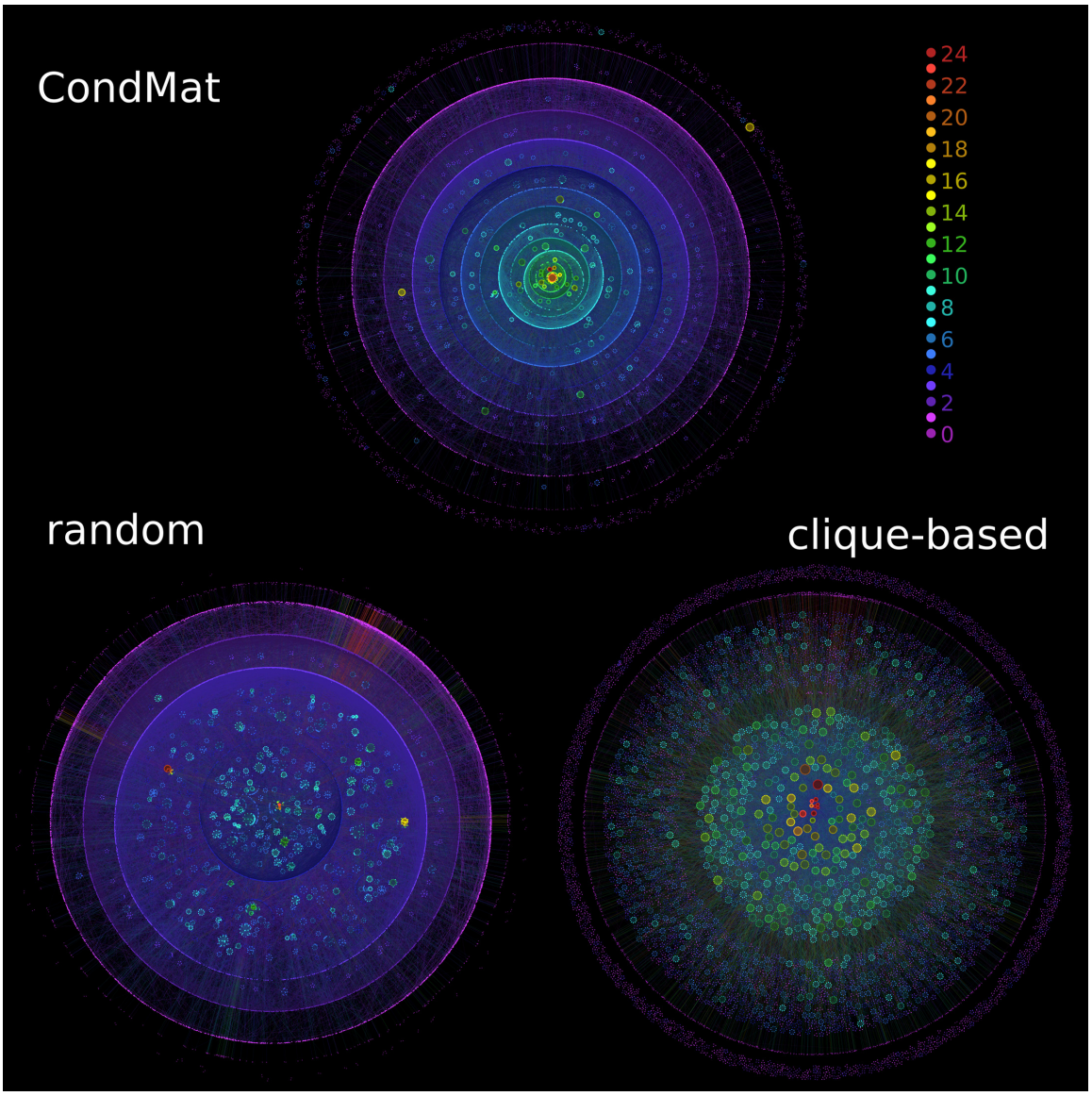

FIG. SI-12: m-core decomposition of theCondMat metabolism network and its random versions.

have that $deg_{G^*}(v) = 2 \cdot |M_G(e_{ij})|$.

Now we compute the k-core decomposition[12] of G^*. Observe that the concept of k-core extends naturally to hypergraphs, where the degree of a node is the number of edges in which it takes part.

We observe the following fact: The l-(k-core) $(V_l^*, E_l^*) = \mathcal{K}^l(G^*)$ maps directly to the same l-(m-core) $(V_l, E_l) = \mathcal{M}^l(G)$, by the following relations:

$$E_l = V_l^*$$

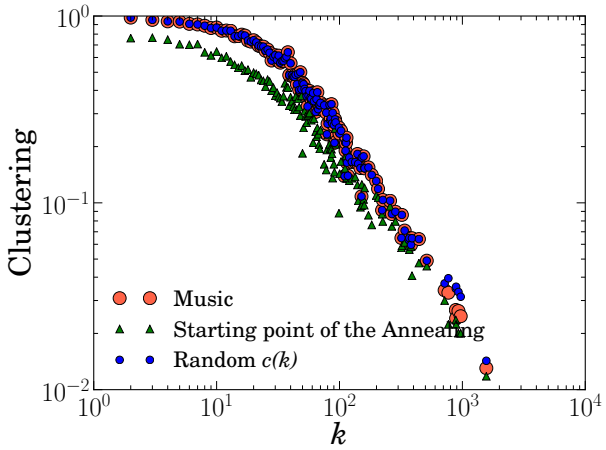

FIG. SI-13: Comparison of the clustering spectrum between the real Music network, the randomized version from where we started the annealing process, and the maximally random model.

$$V_l = \{v_i \in V : \exists j : e_{ij} \in E_l\}$$

The proof of this fact is as follows: Given the l-(k-core) of G^*, and $v = e_{ij} \in V_l^*$, e_{ij} has degree at least l in (V_l^*, E_l^*), thus it is part of at least l triangles in (V_l, E_l). Thus, the set of edges V_l^* is part of the l-(m-core) of G.

Similarly, given the l-(m-core) (V_l, E_l) of G, its edges have multiplicity at least l in it. Then, if we construct a hypergraph with $V^* = E_l$ and we add an edge in G^* for each triangle in the l-(m-core) of G, it follows that each node in the hypergraph has degree at least l. Thus, this hypergraph is part of the l-(k-core) of G^*.

Observation: *(Maximality property)* As the l-(k-core) is the maximal subset of V^* such that its vertices have degree at least l in the induced subgraph, the l-(m-core) turns out to be maximal too.

The procedure is illustrated for a sample graph in Figure SI-25.

---

[1] M. Boguñá, F. Papadopoulos, and D. Krioukov, Nature communications **1**, 62 (2010), ISSN 2041-1723, URL `http://www.ncbi.nlm.nih.gov/pubmed/20842196`.

[2] K. Claffy, Y. Hyun, K. Keys, M. Fomenkov, and D. Krioukov, 2009 Cybersecurity Applications & Technology Conference for Homeland Security pp. 205–211 (2009), URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4804445`.

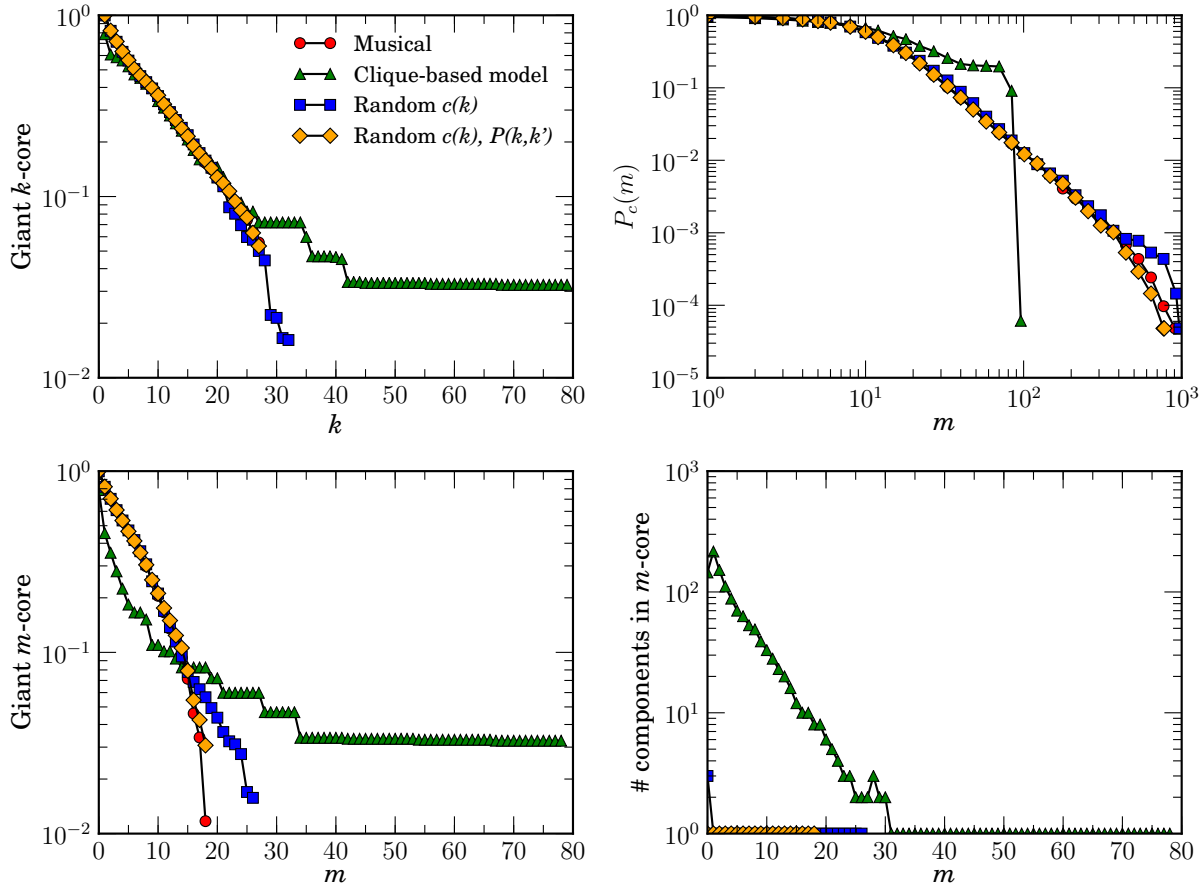

FIG. SI-14: Comparison of the k-core and m-core decompositions between the real musical network, the clique based model, and maximally random models.

[3] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, Physical Review E **70**, 056122 (2004), ISSN 1539-3755, URL http://link.aps.org/doi/10.1103/PhysRevE.70.056122.

[4] M. A. Serrano, M. Boguñá, and F. Sagués, Molecular bioSystems **8**, 843 (2012), ISSN 1742-2051, URL http://www.ncbi.nlm.nih.gov/pubmed/22228307.

[5] J. Leskovec, J. Kleinberg, and C. Faloutsos, ACM Transactions on Knowledge Discovery from Data **1**, 2 (2007), ISSN 15564681, URL http://portal.acm.org/citation.cfm?doid=1217299.1217301.

[6] J. Serrà, A. Corral, M. Boguñá, M. Haro, and J. L. Arcos, Scientific reports **2**, 521 (2012), ISSN 2045-2322, URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3405292&tool=pmcentrez&rendertype=abstract.

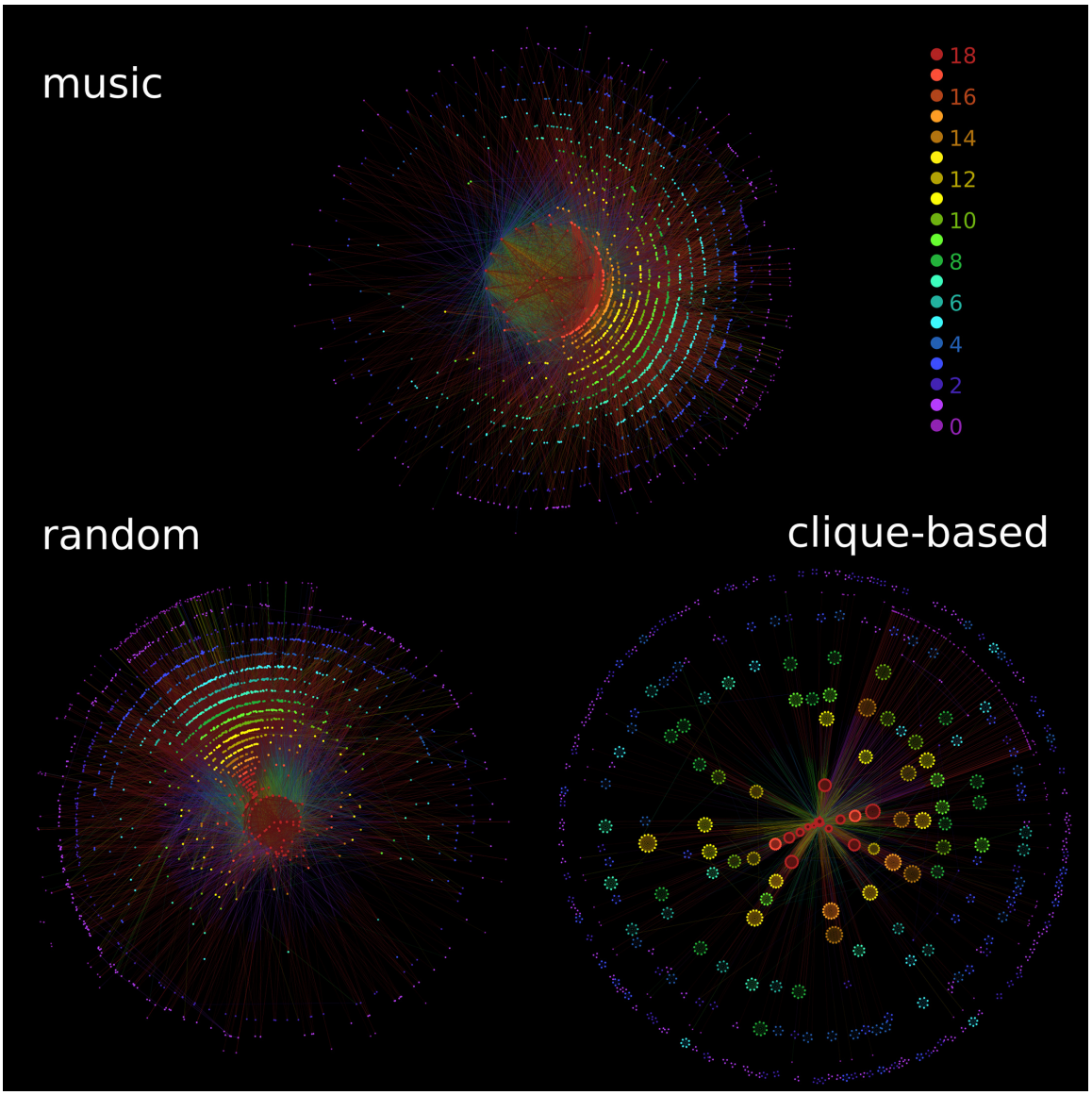

FIG. SI-15: m-core decomposition of the musical network and its random versions.

[7] R. Guimerà, L. Danon, a. Díaz-Guilera, F. Giralt, and a. Arenas, Physical review. E, Statistical, nonlinear, and soft matter physics **68**, 065103 (2003), ISSN 1539-3755, URL `http://www.ncbi.nlm.nih.gov/pubmed/14754250`.

[8] M. Richardson, R. Agrawal, and P. Domingos, pp. 351–368 (2003).

[9] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998), ISSN 0028-0836, URL `http://www.ncbi.nlm.nih.gov/pubmed/9623998`.

[10] K. Saito, T. Yamada, and K. Kazama, pp. 2–6 (2006).

[11] E. B. Jarrett, in *7 th International Kalamazoo Conference in Graph Theory, Combinatorics, Algo-*

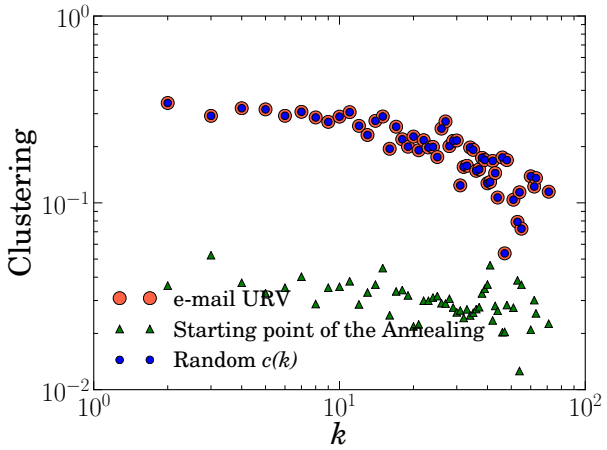

FIG. SI-16: Comparison of the clustering spectrum between the real e-mail URV network, the randomized version from where we started the annealing process, and the maximally random model.

rithms, and Applications (Wiley, 1995), pp. 589–599.

[12] S. Dorogovtsev, a. Goltsev, and J. Mendes, Physical Review Letters **96**, 040601 (2006), ISSN 0031-9007, URL `http://link.aps.org/doi/10.1103/PhysRevLett.96.040601`.

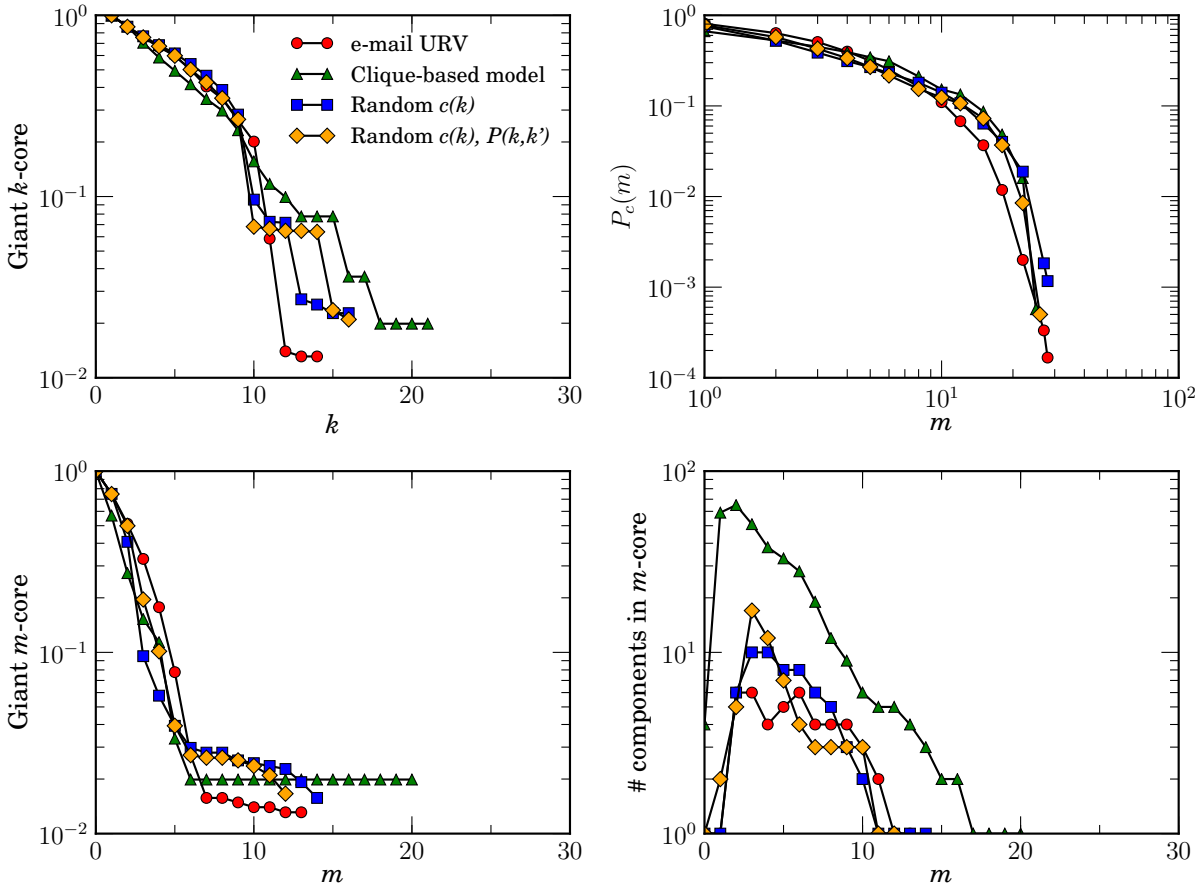

FIG. SI-17: Comparison of the k-core and m-core decompositions between the real musical network, the clique based model, and maximally random models.

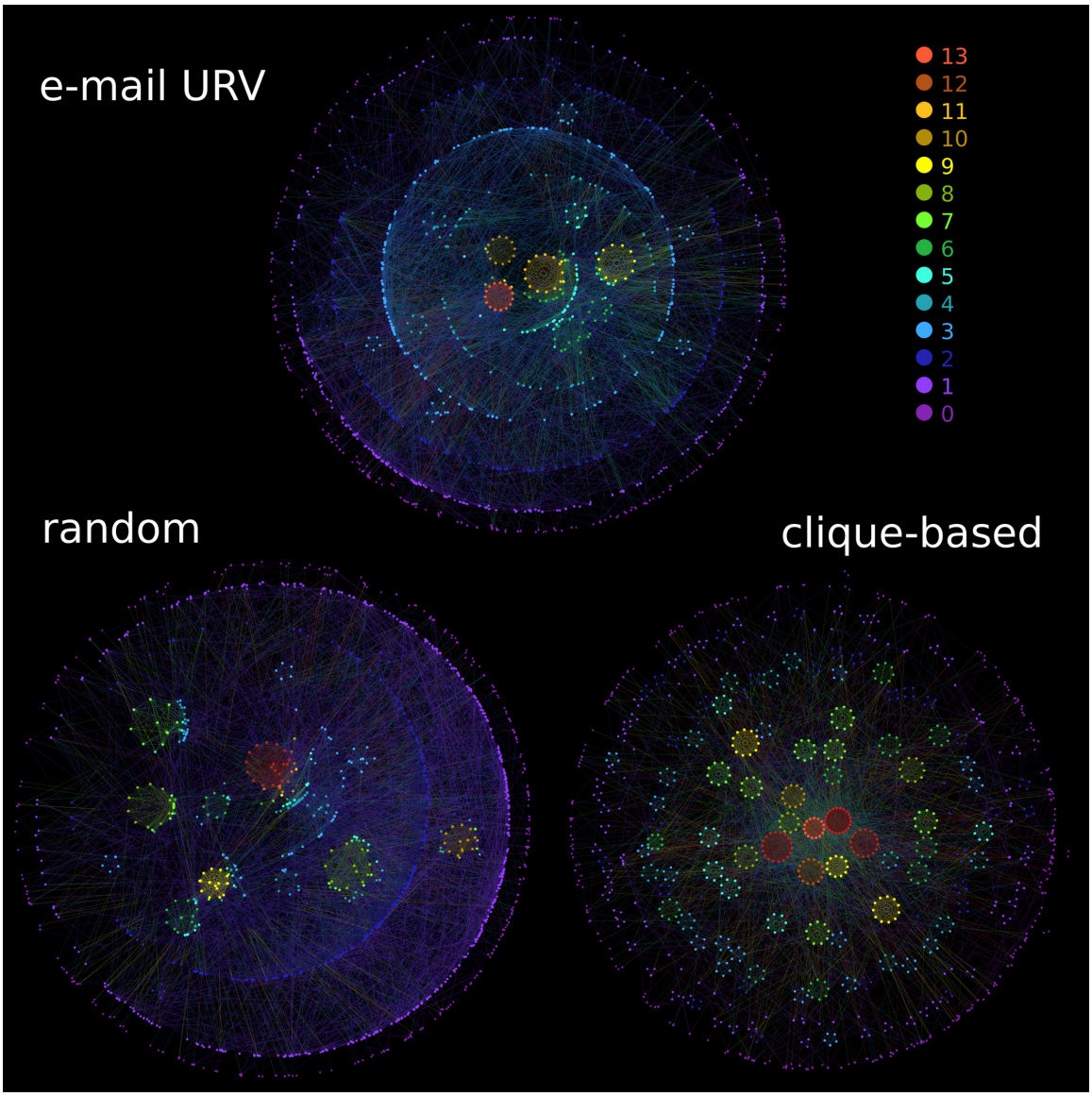

FIG. SI-18: m-core decomposition of the email of the URV network and its random versions.

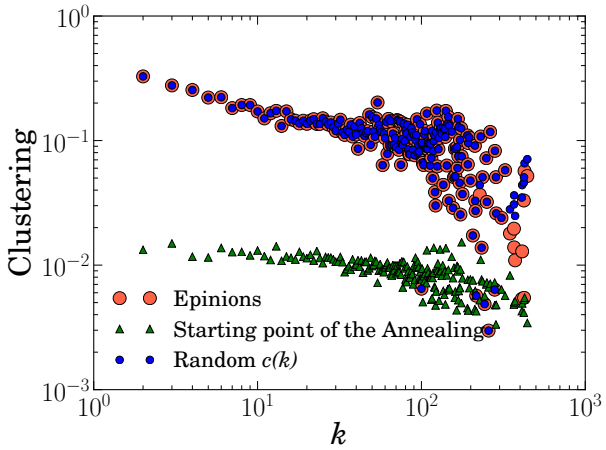

FIG. SI-19: Comparison of the clustering spectrum between the real Epinions network, the randomized version from where we started the annealing process, and the maximally random model.

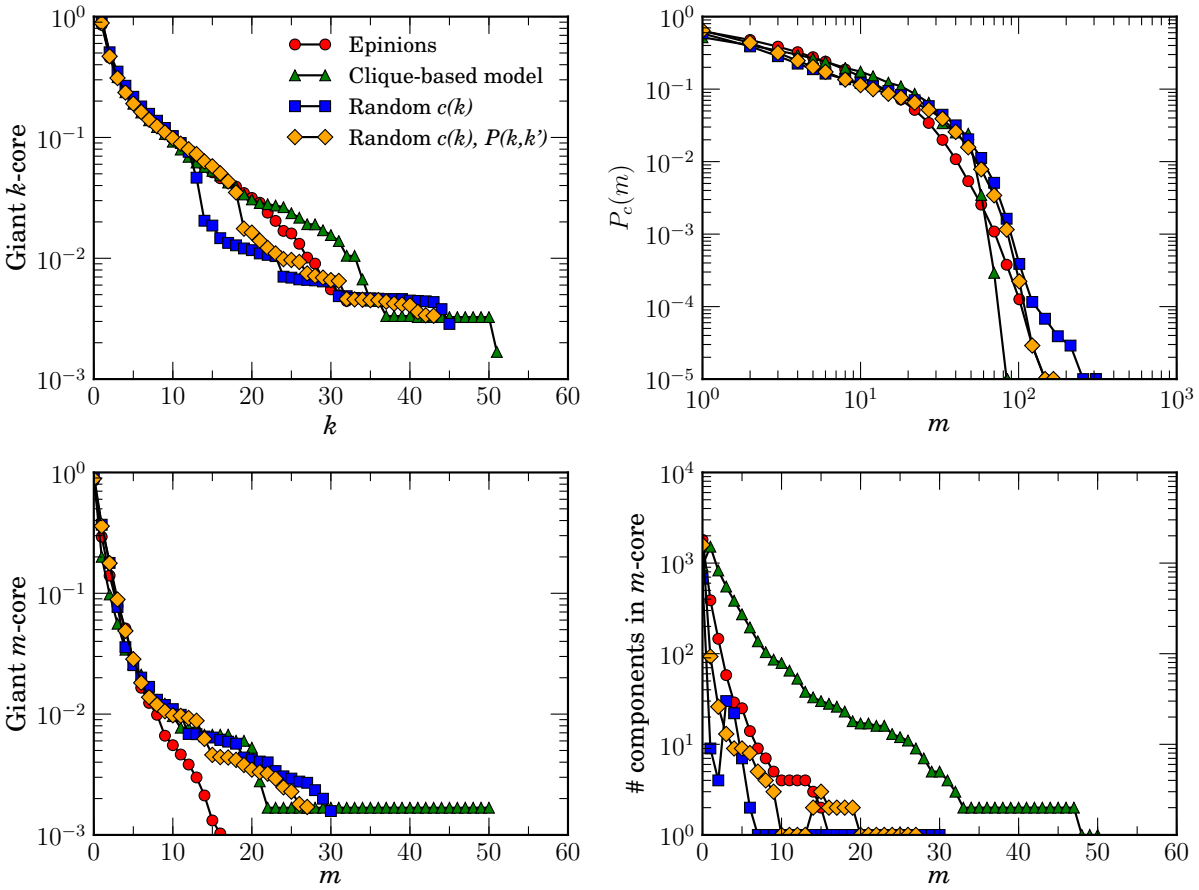

FIG. SI-20: Comparison of the k-core and m-core decompositions between the real Epinions trust network, the clique based model, and maximally random models.

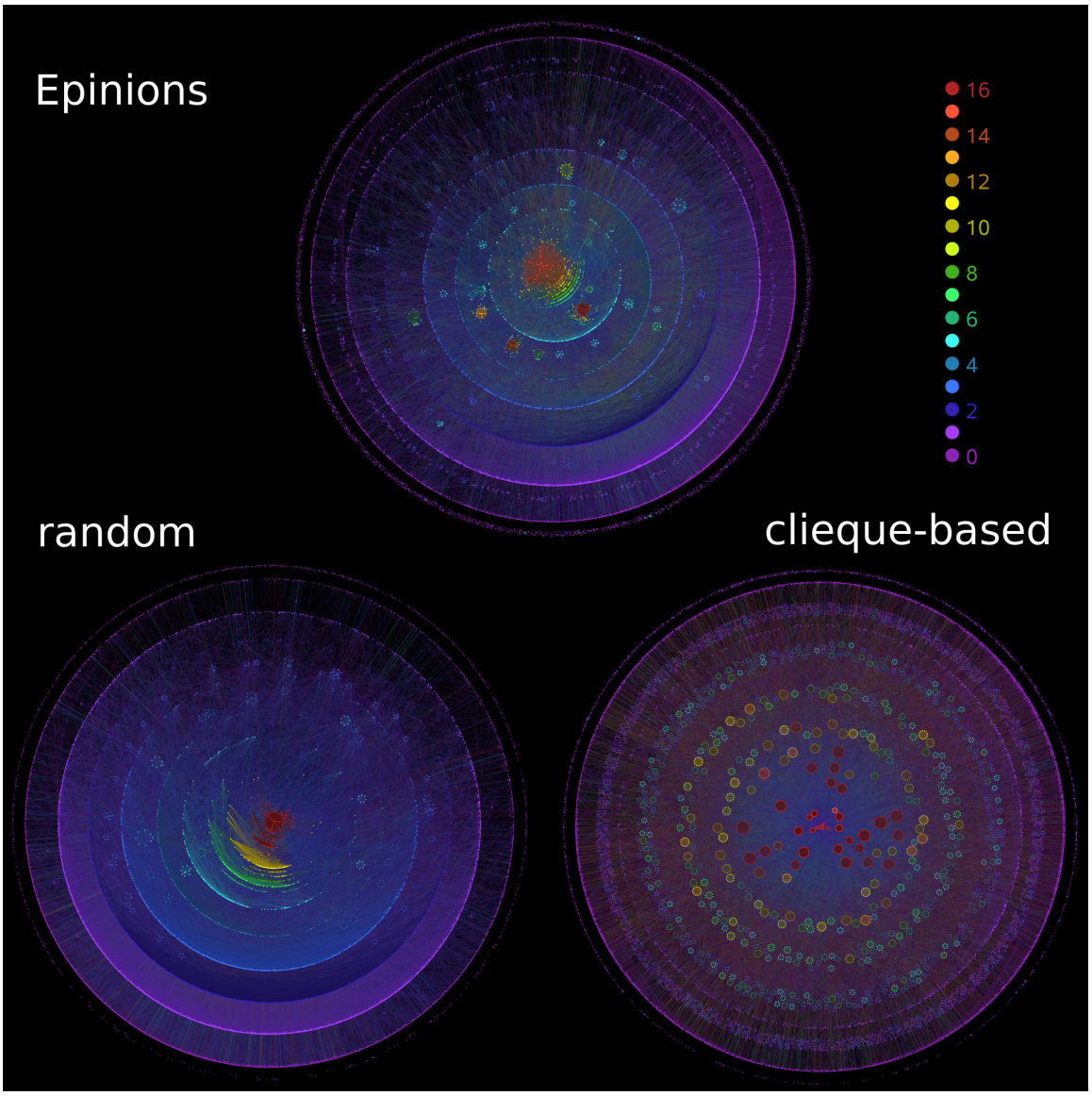

FIG. SI-21: m-core decomposition of the Epinions trust network and its random versions.

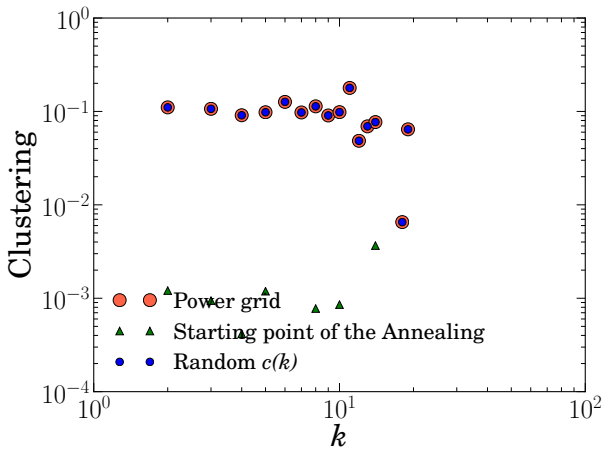

FIG. SI-22: Comparison of the clustering spectrum between the real power grid network, the randomized version from where we started the annealing process, and the maximally random model.

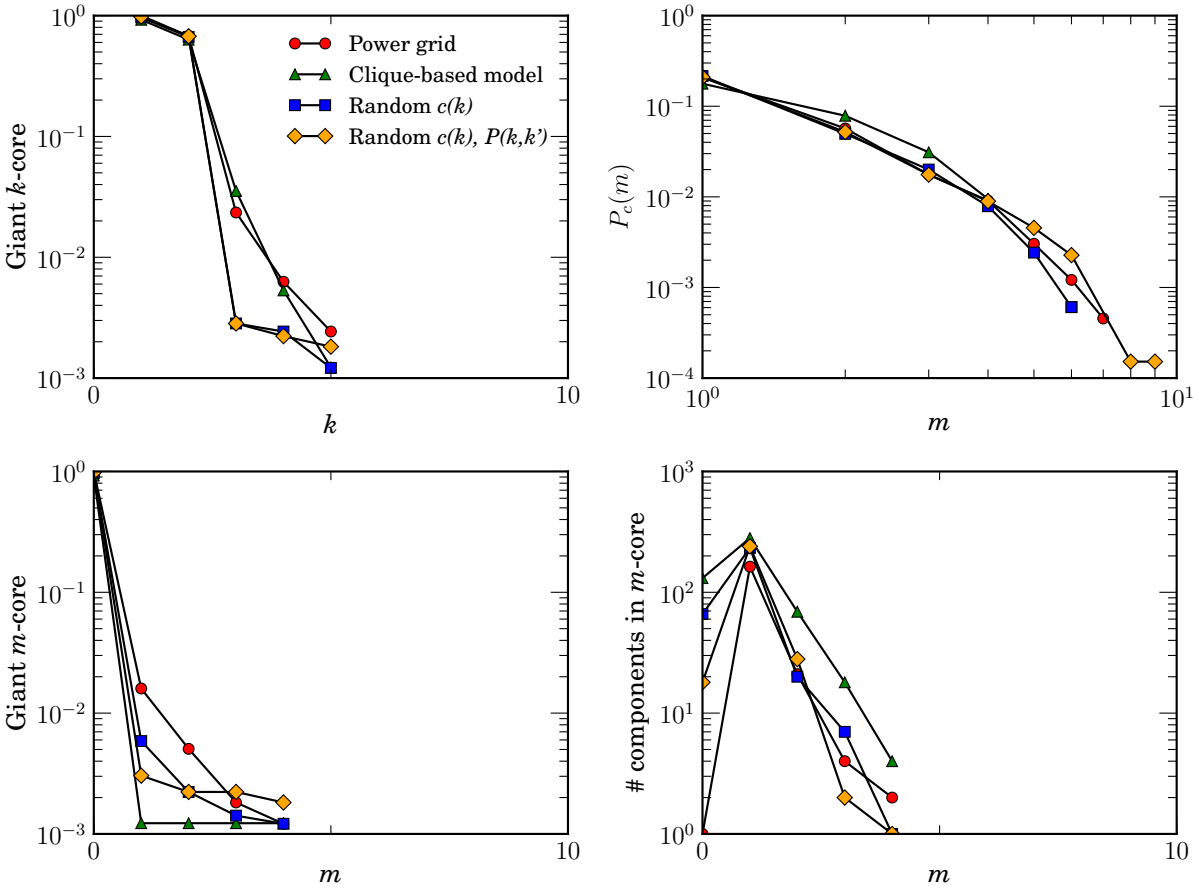

FIG. SI-23: Comparison of the k-core and m-core decompositions between the real power grid network, the clique based model, and maximally random models.

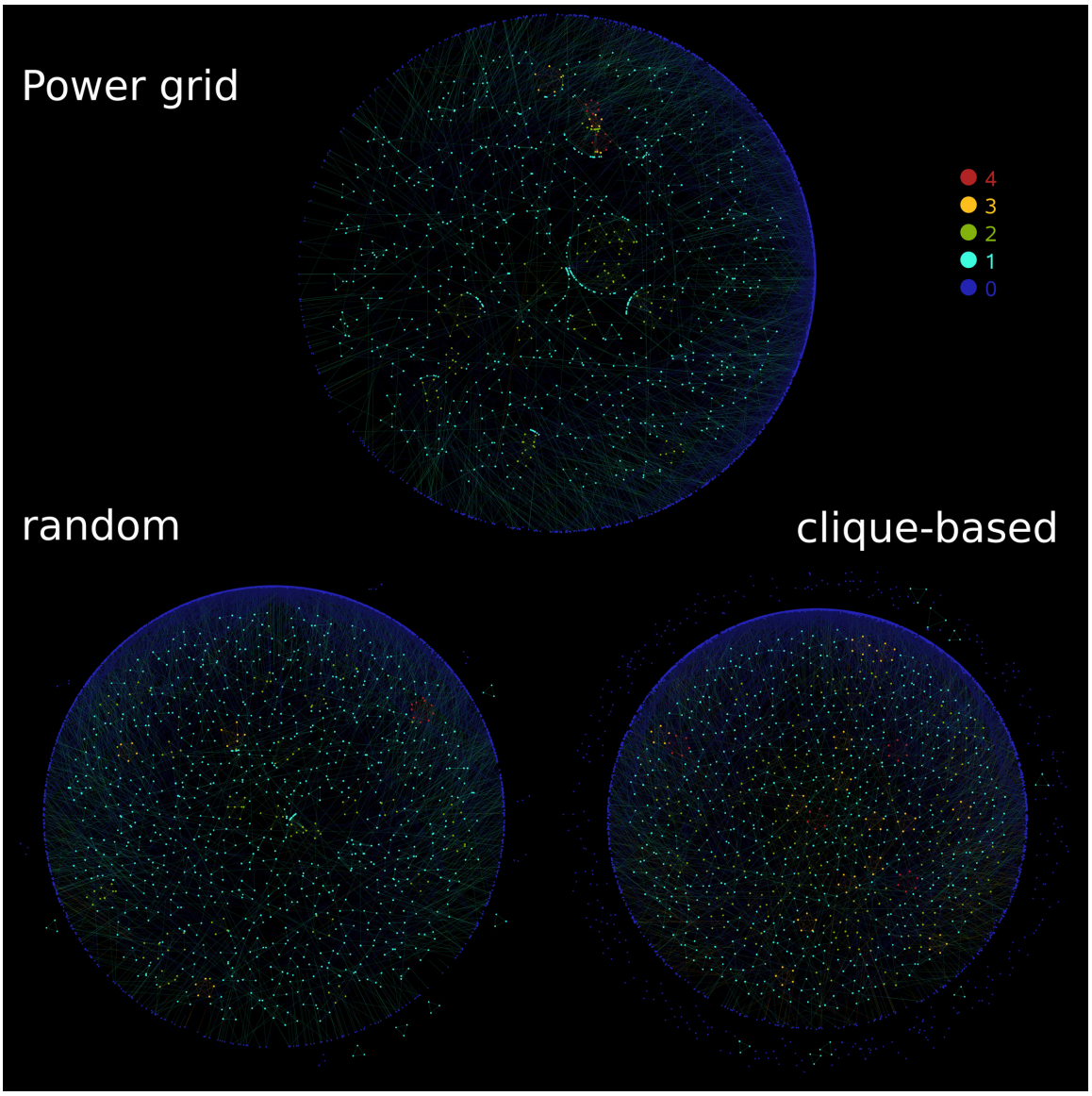

FIG. SI-24: m-core decomposition of the power grid and its random versions.

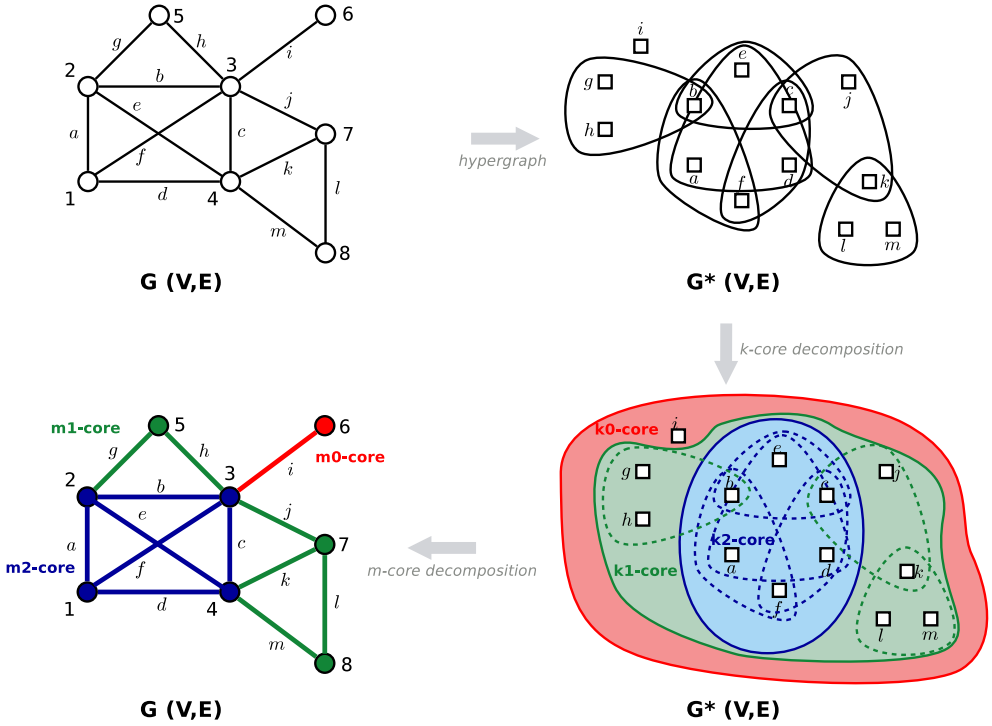

FIG. SI-25: Illustration of the m-cores extraction procedure using a sample graph $G(V, E)$ *(upper left)*. The hypergraph G^* is built *(upper right)*. The nodes in this graph represent edges in the original graph, and an edge (3-tuple) is added for each triangle found in the original graph. Then, the k-core decomposition of G^* is computed *(lower right)*. Finally, the *k-coreness* of each node in G^* maps to the *m-coreness* of an edge in G, and thus the m-core decomposition of G is straightforwardly obtained *(lower left)*.