Algorithm for counting large directed loops

2008 J. Phys. A: Math. Theor. 41 224003

(http://iopscience.iop.org/1751-8121/41/22/224003)

J. Phys. A: Math. Theor. **41** (2008) 224003 (8pp) doi:10.1088/1751-8113/41/22/224003

Algorithm for counting large directed loops

Ginestra Bianconi[1] **and Natali Gulbahce**[2]

[1] The Abdus Salam International Center for Theoretical Physics, Strada Costiera 11, 34014 Trieste, Italy
[2] Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, NM 87545, USA

Received 10 September 2007
Published 21 May 2008
Online at stacks.iop.org/JPhysA/41/224003

Abstract
We derive a Belief-Propagation algorithm for counting large loops in a directed network. We evaluate the distribution of the number of small loops in a directed random network with given degree sequence. We apply the algorithm to a few characteristic directed networks of various network sizes and loop structures and compare the algorithm with exhaustive counting results when possible. The algorithm is adequate in estimating loop counts for large directed networks and can be used to compare the loop structure of directed networks and their randomized counterparts.

PACS numbers: 89.75−k, 89.75.Fb, 89.75.Hc

1. Introduction

The structure of complex networks highly affects the critical behavior of different cooperative models [1] and the nonlinear dynamical process that takes place on the network [2].

In particular, both the directionality of the links which suggest a nonsymmetric interaction [3–5] and the local loop structure [6] of the network which correlates neighboring nodes have important dynamical consequences. In fact, directionality of links becomes particularly important when a transport process of mass or information takes place in the network [3] and the loop structure in these directed networks is crucial for assessing the networks' robustness characteristics and determining the load distribution.

Directed networks are ubiquitous in both man-made and natural systems. Some examples of directed networks are the Texas power-grid, the World-Wide-Web, the foodwebs and in biological networks, such as the metabolic network, the transcription network and the neural network. The local structure of the directed network is radically different from the structure of their undirected version [7]. While many undirected networks are characterized but large clustering coefficient [8] and large number of short loops [9, 10] this is not a general trend for directed networks. For example, the *C.elegans* neural network has a over-representation of short loops compared to a randomized network if the direction of the links is not considered

while it has an under-representation of the number of loops when the direction of the links is taken into account [7].

Nevertheless, while counting small loops in a given network is a relatively easy computation, counting large loops in a real world network is a very hard task. In fact, the number of large loops can, and usually does grow exponentially with the number of nodes N in the network. The known efficient exhaustive algorithms [11, 12] for counting loops still have a time bound of $O(N * M * (L + 1))$ where N, M, L are respectively the number of nodes, links and loops in the network. This task becomes computationally inapplicable for counting large loops in many real networks. Two different approaches for the study of long loops have been proposed: devising Monte Carlo algorithms, or using Belief-Propagation (BP) algorithms. The two approaches have both been pursued in the case of undirected networks [13–15]. The BP algorithm [14] is a heuristic algorithm which does not have sampling bias as the Monte Carlo algorithm [13] does and is observed to give good results as the size of the network increases.

In this paper, we generalize the BP algorithm proposed by [14, 15] to directed networks. We analytically derive the outcome of the algorithm in an ensemble of random uncorrelated networks with given degree sequence of in/out degrees in agreement with the prediction for the average number of nodes in this ensemble [7]. We finally study the particular limitations of the algorithm for small network sizes and small number of loops in the graph. The paper is divided into four further sections. In section 2, we derive the BP algorithm for directed networks following the similar steps as described in [15]. In section 3, we derive the distribution of the small loops in uncorrelated random ensembles. In sections 4 and 5, we describe the steps in the algorithm and its application to a few characteristic directed networks.

2. Derivation of the BP algorithm

Given a network G of N nodes and M links, we define a partition function $Z(u)$ as the generating functions of the number $\mathcal{N}_L$ of loops of length L in the network,

$$Z(u) = \sum_L u^L \mathcal{N}_L(G). \tag{1}$$

Starting with this partition function, we can define a free energy $f(u)$ and an entropy $\sigma(\ell)$ of the loops of length $L = N\ell\sigma(\ell)$ as the follows:

$$f(u) = \frac{1}{N} \ln Z(u) \qquad \sigma(\ell) = \frac{1}{N} \ln \mathcal{N}_{L=\ell N}. \tag{2}$$

For each directed link in the network, $l = \langle ij \rangle$ from node i to node j, if we define a variable $S_l = 0, 1$ which indicates if a given loop passes through the link l, the partition function $Z(u)$ can then be written as

$$Z(u) = \sum_{\{S_l\}} w(\{S_l\}) u^{\sum_{l=1}^{M} S_l}, \tag{3}$$

where $w(\{S_l\})$ is an indicator function of the loops, i.e. it is 1 if the variables $S_l = 1$ have a support which forms a closed loop, and it is zero otherwise. As in [14, 15] we take for simplicity a relaxed local form of the indicator function $w(\{S_l\})$ which is 1 also if the assignment of the link variables S_l is compatible with a few disconnected loops. In particular, we take $w(\{S_l\})$ as

$$w(\{S_l\}) = \prod_{i=1}^{N} w_i(\{S\}_i), \tag{4}$$

where $\{S\}_i = \{S_{\langle ij\rangle}\}_{j\in\partial i}$, and ∂i indicates the set of nodes either pointing to i or pointed by i and where $w_i(\{S\}_i)$ is defined as

$$w_i(\{S\}_i) = \begin{cases} 1 & \text{if} \quad \sum_{j\in\partial_+ i} S_{\langle ij\rangle} = 1 \quad \text{and} \quad \sum_{j\in\partial_- i} S_{\langle ij\rangle} = 1 \\ 1 & \text{if} \quad \sum_{j\in\partial_+ i} S_{\langle ij\rangle} = 0 \quad \text{and} \quad \sum_{j\in\partial_- i} S_{\langle ij\rangle} = 0 \\ 0 & \text{otherwise} \end{cases}$$

with $\partial_+ i$ and $\partial_- i$ indicating the sets of nodes j which point to i or which are pointed by i, respectively. Finding the free energy $f(u)$ associated with the partition function (3) can be cast into finding normalized distributions $p_v(\{S_l\})$ which minimize the Kullback distance

$$F_{\text{Gibbs}}[p_v] = \sum_{\{S_l\}} p_v(\{S_l\}) \ln\left(\frac{p_v(\{S_l\})}{w(\{S_l\})u^{\sum_l S_l}}\right). \tag{5}$$

In fact, it is straightforward to show that F_{Gibbs} assumes its minimal value when $p_v(\{S_l\}) = w(\{S_l\})u^{\sum_l S_l}/Z$. If the given network is a tree, the trial distribution $p_v(\underline{S})$ takes the form

$$p(\{S_l\}) = \left(\prod_l p_l(S_l)\right)^{-1} \left(\prod_i p_i(\underline{S}_i)\right) \tag{6}$$

with $p_l(S_l)$ and $p_i(\{S\}_i)$ being the marginal distributions

$$p_\ell(S_\ell) = \sum_{\{S_l\}\setminus S_l} p(\{S_l\}) \qquad p_i(\{S\}_i) = \sum_{\{S\}\setminus\{S\}_i} p(\{S\}). \tag{7}$$

In a real case, when the network is not a tree, we can always take a variational approach and try a given trial distribution of form (6). After taking this variational approach, we then have to minimize the Bethe free energy F_{Bethe} as

$$F_{\text{Bethe}}[\{p_i\},\{p_l\}] = \sum_i \sum_{\{S\}_i\setminus S_l} p_i(\{S\}_i) \ln\left(\frac{p_i(\{S\}_i)}{w_i(\{S\}_i)}\right) - \sum_l \sum_{S_l} p_l(S_l)\ln(p_l(S_l)u^{S_l}). \tag{8}$$

For each link $l\langle ij\rangle$ starting from i and ending in j, there are the constraints

$$p_l(S_l) = \sum_{\{S\}_i} p_i(\{S\}_i) \qquad p_l(S_l) = \sum_{\{S\}_j} p_j(\{S\}_j). \tag{9}$$

Introducing the Lagrangian multipliers enforcing the conditions (9) and the normalization of the probabilities it is easy to show that a possible parametrization of the marginals is as follows:

$$\begin{aligned} p_l(S_l) &= \frac{1}{C_l}(u y_{i\to j}\hat{y}_{j\to i})^{S_l} \\ p_i(\{S\}_i) &= \frac{1}{C_i} w_i(\{S\}_i) \prod_{j\in\partial_+ i} (u y_{i\to j})^{S_{\langle ij\rangle}} \prod_{j\in\partial_- i} (u\hat{y}_{i\to j})^{S_{\langle ij\rangle}}. \end{aligned} \tag{10}$$

For every directed link $\langle ij\rangle$ from node i to node j the values of the messages $y_{i\to j}$ and $\hat{y}_{j\to i}$ are fixed by the constraints in equation (9) to satisfy the following BP equations:

$$\begin{aligned} y_{i\to j} &= \frac{u\sum_{k\in\partial_- i} y_{k\to i}}{1+u^2\sum_{k'\in\partial_+(i)\setminus j}\hat{y}_{k'\to i}\sum_{k\in\partial_- i} y_{k\to i}} \\ \hat{y}_{j\to i} &= \frac{u\sum_{k\in\partial_+ j}\hat{y}_{k\to j}}{1+u^2\sum_{k'\in\partial_- j\setminus i} y_{k'\to j}\sum_{k\in\partial_+ i}\hat{y}_{k\to j}}. \end{aligned} \tag{11}$$

The normalization constants for the marginals are consequently given by

$$C_l = 1 + u y_{i\to j}\hat{y}_{j\to i}. \qquad C_i = 1 + u^2 \sum_{k'\in\partial_- i} y_{k'\to i} \sum_{k\in\partial_+ i} \hat{y}_{k\to i}. \tag{12}$$

The Bethe free energy density $f_{\text{Bethe}} = \frac{1}{N} F_{\text{Bethe}}$ becomes

$$N f_{\text{Bethe}}(u) = -\sum_{l=1}^{M} \ln C_l + \sum_{i=1}^{N} \ln C_i . \tag{13}$$

For any given value of u the loops length is given by

$$\ell(u) = \frac{1}{N} \sum_{l=1}^{M} p_l(1) = \frac{1}{N} \sum_{l} \frac{u y_{i\to j} \hat{y}_{j\to i}}{1 + u y_{i\to j} \hat{y}_{j\to i}}. \tag{14}$$

The function $\ell(u)$ can be inverted giving the function $u(\ell)$ and finally proving an expression for the entropy of the loops in the graph under a Bethe variational approach,

$$\sigma_{\text{Bethe}}(\ell) = f(u(\ell)) - \ell \ln u(\ell). \tag{15}$$

3. Derivation of the typical number of short loops in a random directed network with a given degree sequence

We consider an ensemble of random directed networks with a given degree sequence $\{k_i^{\text{in}}, k_i^{\text{out}}\} \forall i = 1, \ldots, N$. If the maximal in/out connectivities $K^{\text{in}}/K^{\text{out}}$ of the network satisfy the inequality $K^{\text{in}} K^{\text{out}} < (\langle k_{\text{in}} \rangle) N$, the network is uncorrelated. By $q_{k_{\text{in}},k_{\text{out}}}$ we indicated the degree distribution of the ensemble. In [7], an expression for the average number $\mathcal{N}_L$ of small loops was given,

$$\langle \mathcal{N}_L \rangle \simeq \frac{1}{L} \left(\frac{\langle k_{\text{in}} k_{\text{out}} \rangle}{\langle k_{\text{in}} \rangle} \right)^L \tag{16}$$

valid as long as

$$L \ll N \frac{\langle k_{\text{in}} k_{\text{out}} \rangle^2}{\langle (k_{\text{in}} k_{\text{out}})^2 \rangle}. \tag{17}$$

It an interesting exercise to see what is the distribution of the number of small loops in the ensemble of directed networks by solving the BP equation for a random directed ensemble in parallel with the distribution found in the undirected case [15]. In a directed network ensemble the BP messages y and $\hat{y}$ along each link are equally distributed depending only on the value of u. Given the BP equations (11), the distribution $P(y; u)$ of the field y has to satisfy the self-consistent equation

$$\begin{aligned} P(y; u) = & \sum_{k_{\text{out}}=1}^{\infty} \frac{k_{\text{out}}}{\langle k_{\text{out}} \rangle} q_{0,k_{\text{out}}} \delta(y) + \sum_{k_{\text{in}}=1}^{\infty} \sum_{k_{\text{out}}=1}^{\infty} \frac{k_{\text{out}}}{\langle k_{\text{out}} \rangle} q_{k_{\text{in}},k_{\text{out}}} \\ & \times \int_0^{\infty} \mathrm{d}y_1 P(y_1; u) \ldots, \mathrm{d}y_{k_{\text{in}}} P(y_{k_{\text{in}}}; u) \\ & \times \int_0^{\infty} \mathrm{d}\hat{y}_1 P(\hat{y}_1; u) \ldots \mathrm{d}\hat{y}_{k_{\text{out}}} P(\hat{y}_{k_{\text{out}}}; u) \delta(y - g_k(\{y\}, \{\hat{y}\})) \end{aligned} \tag{18}$$

with

$$\begin{aligned} g_1 &= u y_1 \\ g_k &= \frac{u \sum_{k \in \partial_- i} y_{k\to i}}{1 + u^2 \sum_{k' \in \partial_+(i) \setminus j} \hat{y}_{k'\to i} \sum_{k \in \partial_- i} y_{k\to i}} \qquad \text{for} \qquad k \geqslant 2. \end{aligned} \tag{19}$$

In fact, given a random edge the probability that its starting node i has connectivity $(k_{\rm out}, k_{\rm in})$ is given by $\frac{k_{\rm out}}{\langle k_{\rm out}\rangle} q_{k_{\rm in},k_{\rm out}}$. The fields $\hat{y}$ have to satisfy a similar recursive equation, i.e.

$$P(\hat{y};u) = \sum_{k_{\rm in}=1}^{\infty} \frac{k_{\rm in}}{\langle k_{\rm in}\rangle} q_{k_{\rm in},0}\delta(y) + \sum_{k_{\rm in}=1}^{\infty}\sum_{k_{\rm out}=1}^{\infty} \frac{k_{\rm in}}{\langle k_{\rm in}\rangle} q_{k_{\rm in},k_{\rm out}} \times \int_0^{\infty} {\rm d}y_1 P(y_1;u)\ldots {\rm d}u_{k_{\rm in}} y_{k_{\rm in}} P(y_{k_{\rm in}};u) \times \int_0^{\infty} {\rm d}\hat{y}_1 P(\hat{y}_1;u)\ldots {\rm d}u_{k_{\rm out}} \hat{y}_{k_{\rm out}} P(\hat{y}_{k_{\rm out}+};u)\,\delta(y-\hat{g}_k(\{y\},\{\hat{y}\})) \tag{20}$$

with

$$\hat{g}_1 = u\hat{y}_1$$
$$\hat{g}_k = \frac{u\sum_{k\in\partial_+ i}\hat{y}_{k\to i}}{1+u^2\sum_{k'\in\partial_+(i)\setminus j}\hat{y}_{k'\to i}\sum_{k\in\partial_- i} y_{k\to i}} \qquad \text{for} \qquad k\geqslant 2. \tag{21}$$

For a given small value of $u = u_m + \epsilon$, the two coupled equations in equations (18) and (20) become independent. By proceeding as in [15], we find that the number of small loops in the ensemble is given by

$$\langle N_L\rangle \simeq \frac{1}{L}\left(\frac{\langle k_{\rm in}k_{\rm out}\rangle}{\langle k_{\rm in}\rangle}\right)^L \tag{22}$$

with Poisson fluctuations for loops of size $L \ll \log(N)$. For larger loop sizes up to the boundary limit given by (17), the average number of loops in the ensemble is still given by (22) but with significant fluctuations in the number of loops.

4. The BP algorithm

The study of the partition function equation (3) carried on in section 2 is such that a new algorithm for counting large loops in a directed network can be formulated. In particular, given a network with N nodes and M links, the algorithm is:

- Initialize the messages $y_{i\to j}$, $\hat{y}_{j\to i}$ for every directed link between i and j to random values.
- For every value of u, iterate the BP equations in equation (11)

$$y_{i\to j} = \frac{u\sum_{k\in\partial_- i} y_{k\to i}}{1+u^2\sum_{k'\in\partial_+(i)\setminus j}\hat{y}_{k'\to i}\sum_{k\in\partial_- i} y_{k\to i}}$$
$$\hat{y}_{j\to i} = \frac{u\sum_{k\in\partial_+ j}\hat{y}_{k\to j}}{1+u^2\sum_{k'\in\partial_- j\setminus i} y_{k'\to j}\sum_{k\in\partial_+ i}\hat{y}_{k\to j}}. \tag{23}$$

  until convergence.
- Calculate $\ell(u)$ and $f(u)$ from equations (14) and (13) which we recall here for convenience

$$\ell(u) = \frac{1}{N}\sum_{l=1}^{M} p_l(1) = \frac{1}{N}\sum_l \frac{u y_{i\to j}\hat{y}_{j\to i}}{1+u y_{i\to j}\hat{y}_{j\to i}}. \tag{24}$$

$$Nf_{\rm Bethe}(u) = -\sum_{l=1}^{M}\ln(1+u y_{i\to j}\hat{y}_{j\to i}) + \sum_{i=1}^{N}\ln\left(1+u^2\sum_{k'\in\partial_- i} y_{k'\to i}\sum_{k\in\partial_+ i}\hat{y}_{k\to i}\right). \tag{25}$$

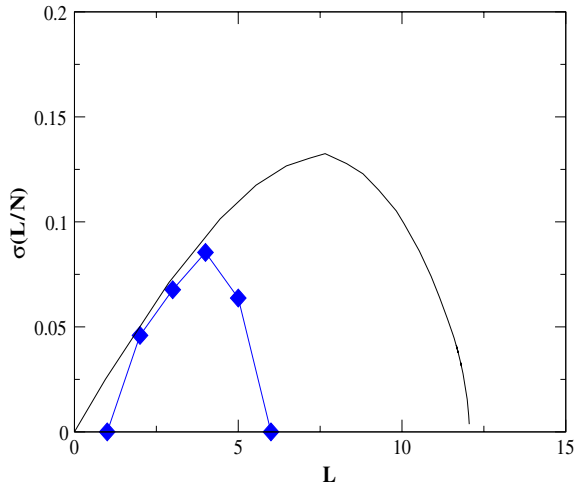

Figure 1. Entropy $\sigma(L/N)$ of the loops of length L for the real *Chesapeake* food-web (solid line) and the entropy of the loops counted by exact enumeration (diamonds).

- Evaluate $\sigma(\ell)$ by equation (15) which again we repeat here for convenience

$$\sigma_{\text{Bethe}}(\ell(u)) = f(u) - \ell(u)\ln u. \tag{26}$$

5. Application of the algorithm to real directed networks

We applied the formulated algorithm to a large set of directed networks [17]. For some of these networks we calculated the number of loops $\mathcal{N}_L$ of length L directly by exact enumeration [12]. We then compare the entropy of the loops $\sigma(\ell)$ found by the BP algorithm with the entropy of the loops $\sigma_0(\ell)$ found by the directed enumeration of the number of loops

$$\sigma_0(\ell) = \frac{1}{N}\ln\left(\mathcal{N}^{\text{exact}}_{L=\ell N}\right). \tag{27}$$

We note that for the foodweb with small number of nodes the algorithm does not provide a good approximation for the number of loops present in the graph. A dramatic example is the Chesapeake foodweb. In this case, we were able to count all the loops in the network exhaustively since the network contains very few loops. In this case, the BP algorithm, since the loops are few the BP algorithm highly overestimates the largest loop in the network (see figure 1). In fact, it predicts a largest loop of lend $L_{\max} = 12$ where the largest loop is of length $L_{\max} = 7$. This effect is observed to be present also in the undirected BP algorithm [14].

The discrepancy is predicted to be strong only in cases where the size of the network is small and the number of loops in the network is small just as in the Chesapeake case. When the network has a larger number of loops and the entropy of the loops is larger, much better results are expected. In the case of the *C. elegans* neural network ($N = 306$) the entropy for small number of loops is overlapping with the results of exact enumeration as it can clearly be seen in figure 2. We further compare the results of the algorithm on a given network and on a randomized network ensemble. A typical example is the metabolic network of *E. coli* [17] (see figure 3) in which we could compare the entropy provided by the BP algorithm with the entropy of a series of 100 random networks with the same degree distribution.

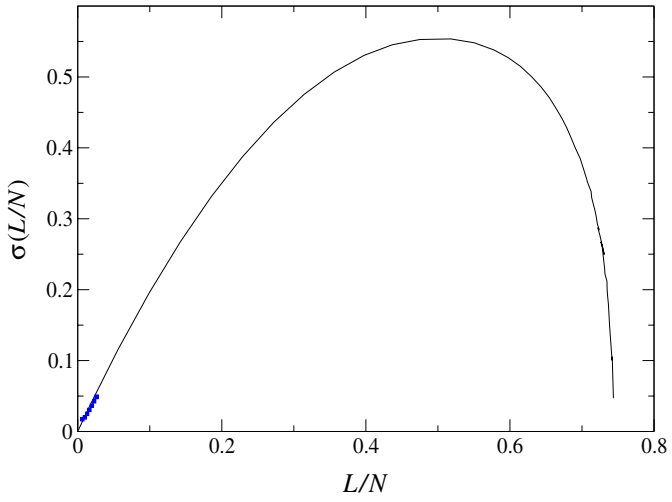

Figure 2. Entropy $\sigma(L/N)$ of the loops of length L for the real *C.elegans* neural network (solid line) and the entropy of the loops counted by exact enumeration for small loops (small diamonds).

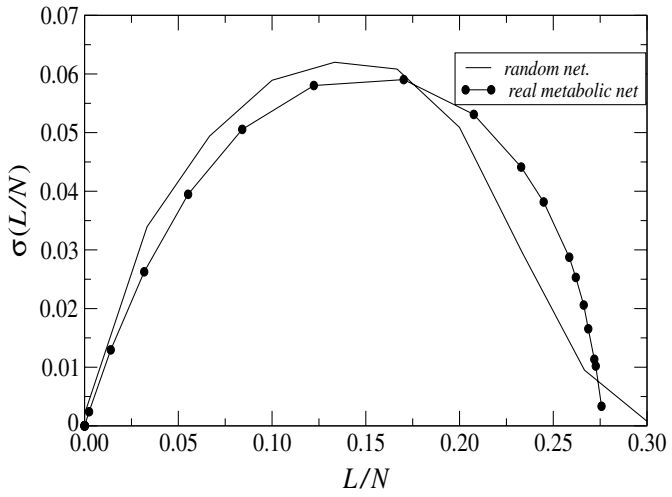

Figure 3. Entropy $\sigma(L/N)$ of the loops of length L for the real metabolic network and average entropy of the loops in the randomized network ensemble with the same degree sequence.

6. Conclusions

In conclusion, we provide a new algorithm for counting large loops in the directed network. The algorithm is predicted to give good results only for large network size N. In this paper we demonstrate cases in which it fails to predict the right entropy and loop structure due to the small size of the network. We propose to study the significance of loop structure in large networks by comparing the results of the algorithm on real networks and randomized networks when networks are large and the number of loops in the network are also large.

Acknowledgments

The work was partially supported by IST STREP GENNETEC contract no. 034952. We acknowledge G Semerjian and A E Motter for interesting discussions.

References

[1] Dorogovtsev S N, Goltsev A V and Mendes J F F 2007 *Preprint* 0705.0010
[2] Motter A E, Matias M A, Kurths J and Ott E 2006 *Physica* D **224** vii
[3] Toroczkai Z and Bassler K E 2004 *Nature (London)* **428** 716
[4] Nishikawa T and Motter A E 2006 *Physica* D **224** 77
[5] Galla T 2006 *J. Phys. A: Math. Gen.* **39** 3853
[6] Klemm K and Bornholdt S 2005 *Proc. Natl Acad. Sci. USA* **102** 18414
[7] Bianconi G, Gulbahce N and Motter A E 2007 *Preprint* 0707.4084
[8] Watts D J and Strogatz S H 1998 *Nature (London)* **393** 440
[9] Bianconi G and Capocci A 2003 *Phys. Rev. Lett.* **90** 078701
[10] Bianconi G and Marsili M 2005 *J. Stat. Mech.* P06005
[11] Johnson D B 1975 *SIAM J. Comput.* **4** 77
[12] Tarjan R 1973 *SIAM J. Comput.* **2** 211
[13] Rozenfeld H D, Kirk J E, Bollt E M and ben-Avraham D 2005 *J. Phys. A: Math. Gen.* **38** 4589
[14] Marinari E, Monasson R and Semerjian G 2006 *Europhys. Lett.* **73** 8
[15] Marinari E and Semerjian G 2006 *J. Stat. Mech.* P06019
[16] Burda Z and Krzywicki A 2003 *Phys. Rev.* E **67** 046118
Boguñá M, Pastor-Satorras R and Vespignani A 2004 *Eur. Phys. J.* B **38** 205
[17] The network data are available at http://vlado.fmf.unilj.si/pub/networks/data/ (Chesapeake and Mondego foodwebs), www.cosinproject.org/ (Littlerock and Seagrass foodwebs), http://cdg.columbia.edu/cdg/ (*C. elegans* neural network), and www.weizmann.ac.il/mcb/UriAlon/ (*S. cerevisiae* transcription network). The Texas power-grid dataset was provided by Ken Werley (LANL), and the *E. coli* metabolic network was generated using flux balance analysis [18] on the reconstructed metabolic model iJE660 available at http://gcrg.ucsd.edu/organisms/. We consider the metabolic network without the metabolites CO_2, NH_3, PP_i, P_i, ATP, ADP, NAD, NADP, and NADH, as in Fell D A and Wagner A 2000 *Nat. Biotech.* **18** 1121
[18] Edwards J S and Palsson B O 2000 *Proc. Natl Acad. Sci. USA* **97** 5528