Weighted networks as randomly reinforced urn processes

Guido Caldarelli,[1,2,3,4] Alessandro Chessa,[2,4] Irene Crimaldi,[1] and Fabio Pammolli[1,5]

[1]*IMT Institute for Advanced Studies, Piazza San Ponziano 6, 55100 Lucca, Italy*
[2]*Istituto dei Sistemi Complessi, Consiglio Nazionale delle Ricerche, Dip. Fisica Università "Sapienza," P.le A. Moro 2, 00185 Rome, Italy*
[3]*London Institute of Mathematical Sciences, 35a South St., Mayfair London W1K 2XF, United Kingdom*
[4]*Linkalab, Complex Systems Computational Laboratory, 09129 Cagliari, Italy*
[5]*Center for Polymer Studies, Boston University, Boston, Massachusetts 02215, USA*

(Received 6 March 2012; revised manuscript received 12 November 2012; published 28 February 2013)

We analyze weighted networks as randomly reinforced urn processes, in which the edge-total weights are determined by a reinforcement mechanism. We develop a statistical test and a procedure based on it to study the evolution of networks over time, detecting the "dominance" of some edges with respect to the others and then assessing if a given instance of the network is taken at its steady state or not. Distance from the steady state can be considered as a measure of the relevance of the observed properties of the network. Our results are quite general, in the sense that they are not based on a particular probability distribution or functional form of the random weights. Moreover, the proposed tool can be applied also to dense networks, which have received little attention by the network community so far, since they are often problematic. We apply our procedure in the context of the International Trade Network, determining a core of "dominant edges."

DOI: 10.1103/PhysRevE.87.020106 PACS number(s): 02.50.Le, 89.75.Hc, 89.75.Da, 89.65.Ef

In recent years complex network theory has proved to be a general-purpose and interdisciplinary tool for the analysis of a variety of systems, in different fields such as physics [1–3], economics [4–7], computer science [8], social science [9], transportation [10], and others, which can be efficiently described by a network structure, where the nodes are the system entities and the edges represent the relations between them. All the models that produce complex networks are based either on preferential attachment (or copying mechanism) or on a fitness (hidden variables) microscopic mechanism. Unfortunately, no statistical method has been developed in order to assess the relevance of both experimental data and model simulations. In this paper, we present a model of network evolution based on randomly reinforced urn (RRU) processes [11–14]. In our model we map the weight associated to a given edge with the number of balls of a given color, which are added in an urn so that at a given time step the probability of picking an edge (color) depends on the total weight associated with it until that time. At each time step we first extract an edge (color) with probability proportional to its total weight and then we associate to it a random weight (number of added balls) which increases its total weight. This results in a preferential attachment (PA) rule for edges with random weights. Hence, although our model can be considered as a particular refinement in the class of the PA mechanisms, its novelty is in the connection that we can establish between complex networks and RRU models [15]. RRU theory allows us to develop a procedure for the detection of the "dominant edges" in the evolution of a weighted network and for an evaluation of the distance from the steady state of the network, in the sense that we can assess if the structure observed at a given time is what we can expect at the steady state or not. The novelty of our methodology is also related to its applicability to dense, weighted networks (a situation often problematic both for modeling and for randomization) [16].

We consider a system with N vertices and *Lpotential* edges (directed or not). Hereafter we indicate the various edges by the index ℓ (with $\ell \in [1,L]$). Our model defines a weighted adjacency matrix $\mathbf{W}_t$ for every time step t, where the generic element $w_{t\ell} = [\mathbf{W}_t]_\ell$ is the total weight associated with the edge ℓ until time step t (the total number of added balls of color ℓ until time step t). Similarly, we define a matrix $\mathbf{K}_t$ whose elements $k_{t\ell} = [\mathbf{K}_t]_\ell$ represent the total number of extractions of edge ℓ until time step t. Note that at a given time the graph may actually be incomplete, since some weights could be zero. We can describe analytically the network dynamics. We start at time $t = 1$ by picking an edge, say ℓ^*, according to the following rule: every edge ℓ can be picked with an initial probability $Z_{0\ell} = a_\ell / \sum_{\ell=1}^{L} a_\ell$, where $a_\ell > 0$. A random weight $W_{1\ell^*}$ is associated to the picked edge ℓ^*. At time step $t + 1$, we pick another edge ℓ^* according to the probability distribution given by

$$Z_{t\ell^*} = \frac{a_{\ell^*} + \sum_{n=1}^{t} W_{n\ell^*} X_{n\ell^*}}{\sum_{\ell=1}^{L} a_\ell + \sum_{\ell=1}^{L} \sum_{n=1}^{t} W_{n\ell} X_{n\ell}}, \qquad (1)$$

where $X_{n\ell} = 1$ if at time step n we picked edge ℓ and $X_{n\ell} = 0$ otherwise, and $W_{n\ell}$ denotes the random weight associated with edge ℓ at time step n. In other words, we define (akin to the PA rule) a probability of edge extraction that takes into account the previous weights of the network. We do not assume *a priori* a specific form or probability distribution of the weights. We only require that they are positive random variables which are uniformly bounded by a constant, and each of them is independent of the previous weights and of the outcomes of the previously done extractions. The parameters a_ℓ do not explicitly appear in the tools we will present hereafter and therefore we avoid estimating them.

We assume that the mean values and the variances of the weights are constant along time and we define $\mathcal{D}$ as the set of edges such that

$$E[W_{t\ell^*}] = \mu^* > 0 \; \forall \ell^* \in \mathcal{D}, \quad E[W_{t\ell}] = \mu_\ell < \mu^* \; \forall \ell \notin \mathcal{D}. \qquad (2)$$

We set $Var[W_{t\ell}] = \sigma_\ell^2$ for each ℓ. If $\mathcal{D}$ coincides with the L edges, the above conditions mean that the weights have the same mean value for all edges. Conversely, when the number of elements in the set $\mathcal{D}$ is lower than L, the weights associated with the edges in $\mathcal{D}$ "dominate in mean" on those associated to the others. (A typical case of the first type holds when every weight $W_{t\ell}$ is equal to a same constant, i.e., the classical PA.) As $t \to +\infty$, the probability $Z_{t\ell}$ of choosing the edge ℓ converges almost surely (a.s.) to zero when $\ell \notin \mathcal{D}$, while it converges a.s. to a random variable Z_{ℓ^*} with values in (0,1) a.s. when $\ell = \ell^* \in \mathcal{D}$ and $\sum_{\ell^* \in \mathcal{D}} Z_{\ell^*} = 1$ [11,12]. Therefore the notion of "dominant edges" could provide a formalization of the empirical evidence that many real networks are rather heterogeneous in the sense that, with respect to all the possible edges, a club of edges collects the major fraction of the total weight of the network. More precisely, it has been analytically proven [11,12] that, as the number of time steps t grows, the total weight associated with the dominant edges grows according to

$$\frac{\sum_{\ell^* \in \mathcal{D}} [\mathbf{W}_t]_{\ell^*}}{t} = \frac{\sum_{\ell^* \in \mathcal{D}} \sum_{n=1}^{t} W_{n\ell^*} X_{n\ell^*}}{t} \xrightarrow{a.s.} \mu^*, \qquad (3)$$

while the same limit for the dominated edges is zero, i.e.,

$$\frac{\sum_{\ell \notin \mathcal{D}} [\mathbf{W}_t]_{\ell}}{t} = \frac{\sum_{\ell \notin \mathcal{D}} \sum_{n=1}^{t} W_{n\ell} X_{n\ell}}{t} \xrightarrow{a.s.} 0\,. \qquad (4)$$

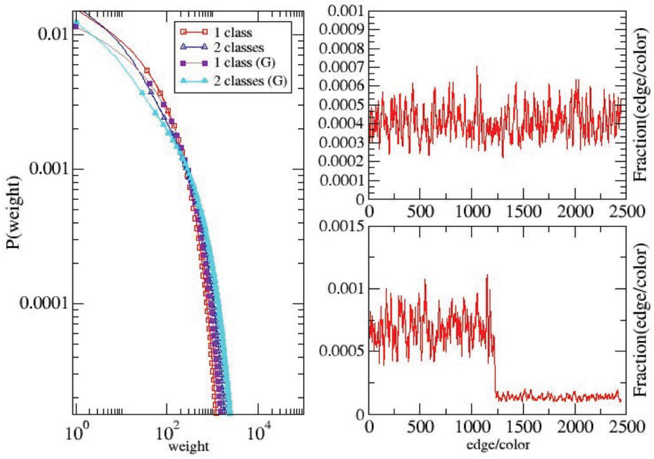

FIG. 1. (Color online) We performed some simulations of the model (with $L = 2500$ and 10^6 extractions) in both cases of no dominant set (one class) and of a dominant set (two classes). On the left we plot the frequency distributions of the weights in the case of uniform/truncated Gaussian (G) distributions of the random variables W. On the right we plot the normalized number of extractions of each edge/color in the case of no dominant set (up) and in the case of two classes with the set [1,1250] as the dominant set (below).

Moreover, for a dominant edge ℓ^*, the total weight associated to that edge normalized by the total weight of the network asymptotically behaves as $Z_{t\ell^*}$ and so converges a.s. to the previous random variable Z_{ℓ^*}. The number of extractions of ℓ^* divided by the total number of extractions also converges a.s. to the same random variable, that is $[\mathbf{K}_t]_{\ell^*}/t = \sum_{n=1}^{t} X_{n\ell^*}/t \xrightarrow{a.s.} Z_{\ell^*}$. On the other hand, the corresponding limits for dominated edges are both equal to zero (see Fig. 1). In particular, we have $t^{1-\lambda} Z_{t\ell} \xrightarrow{a.s.} 0$ for $\ell \notin \mathcal{D}$ and each $\lambda \in (\overline{\lambda},1)$, where $\overline{\lambda} = \max_{\ell \notin \mathcal{D}} \mu_\ell/\mu^*$. The exact distribution of the limit random variable Z_{ℓ^*} is generally unknown (except in some special cases such as the trivial case of one single dominant edge ℓ^* for which we have $Z_{\ell^*} = 1$ a.s. and the case of the above-mentioned classical PA in which Z_{ℓ^*} is beta-distributed). Using the above limit relations and some asymptotic results, analytically proven in [11,12], we have developed a statistical test and a procedure, based on it, for the detection of the set $\mathcal{D}$ and for the assessment of whether a particular instance of a given network has a weight distribution that already evolved into its steady state or not. More precisely, the statistical test is the following. Assuming our model, we take as a null hypothesis H_0 the fact that the "dominant set" $\mathcal{D}$ coincides with a certain subset $\mathcal{D}^*$ of edges with card$(\mathcal{D}^*) \geqslant 2$. Then we consider a certain level α (typically $\alpha = 5\%, 10\%$), we fix ℓ^* in $\mathcal{D}^*$, and we compute

$$\frac{|C^*_{t\ell^*}|}{\sqrt{U_{t\ell^*}}} = \frac{\sqrt{t}\,|\overline{X}^*_{t\ell^*} - Z^*_{t\ell^*}|}{\sqrt{U_{t\ell^*}}}\,, \qquad (5)$$

where

$$\overline{X}^*_{t\ell^*} = \frac{\sum_{n=1}^{t} X_{n\ell^*}}{1 + \sum_{\ell \in \mathcal{D}^*} \sum_{n=1}^{t} X_{n\ell}}\,,$$
$$Z^*_{t\ell^*} = \frac{1 + \sum_{n=1}^{t} W_{n\ell^*} X_{n\ell^*}}{\text{card}(\mathcal{D}^*) + \sum_{\ell \in \mathcal{D}^*} \sum_{n=1}^{t} W_{n\ell} X_{n\ell}}, \qquad (6)$$

and $U_{t\ell^*}$ (assumed to be nonzero) is defined as

$$U_{t\ell^*} = \frac{\overline{X}_{t\ell^*}\left\{(1 - \overline{X}_{t\ell^*})^2 \widehat{\sigma}^2_{t\ell^*} + \overline{X}_{t\ell^*} \sum_{\ell \in \mathcal{D}^*, \ell \neq \ell^*} \overline{X}_{t\ell}\, \widehat{\sigma}^2_{t\ell}\right\}}{(\widehat{\mu}^*_t)^2 \left(\sum_{\ell \in \mathcal{D}^*} \overline{X}_{t\ell}\right)^4}, \qquad (7)$$

with $\overline{X}_{t\ell} = \sum_{n=1}^{t} X_{n\ell}/t$ and $\widehat{\mu}^*_t$ an estimate of the mean value μ^* and $\widehat{\sigma}^2_{t\ell}$ an estimate of the variance σ^2_ℓ:

$$\widehat{\mu}^*_t = \frac{1}{\text{card}(\mathcal{D}^*)} \sum_{\ell \in \mathcal{D}^*} \left(\frac{\sum_{n=1}^{t} W_{n\ell} X_{n\ell}}{\sum_{n=1}^{t} X_{n\ell}} \right),$$
$$\widehat{\sigma}^2_{t\ell} = \frac{\sum_{n=1}^{t} W^2_{n\ell} X_{n\ell}}{\sum_{n=1}^{t} X_{n\ell}} - \left(\frac{\sum_{n=1}^{t} W_{n\ell} X_{n\ell}}{\sum_{n=1}^{t} X_{n\ell}} \right)^2. \qquad (8)$$

As a consequence of a result proven in [12], if H_0 is true, the random variable $C^*_{t\ell^*}/\sqrt{U_{t\ell^*}}$ converges in distribution to the standard normal distribution $\mathcal{N}(0,1)$ as $t \to +\infty$, while this convergence does not hold when H_0 is false. Hence we compare the quantity (5) with the quantile q_α of $\mathcal{N}(0,1)$ of order $(1 - \alpha/2)$ (that is, q_α is the number such that $\mathcal{N}(0,1)(q_\alpha, +\infty) = \alpha/2$ and $q_\alpha = 1.96$ for $\alpha = 5\%$ and $q_\alpha = 1.645$ for $\alpha = 10\%$). If the computed quantity is greater than q_α, then we reject the null hypothesis at the (approximate) level α; otherwise, we cannot reject it.

Simulations have shown that if we perform the above test taking $\mathcal{D}^*$ exactly equal to the true dominant set, known *a priori*, then the percentage of indexes ℓ^* for which the test gives the rejection of the hypothesis is very low (2.28% for $\alpha = 10\%$ and 0.82% for $\alpha = 5\%$). From now on we call this percentage the "rejection percentage." If we consider a different $\mathcal{D}^*$ with the same size as the true dominant set, the rejection percentage increases (even if we change a single element): the more $\mathcal{D}^*$ and the true dominant set are different, the higher the rejection percentage is (we got values up to 93% for $\alpha = 10\%$ and 85% for $\alpha = 5\%$). However, we observed that the ability of the test of rejecting H_0 when it is false decreases with decreasing the size of $\mathcal{D}^*$. This is due to the factor $(\sum_{\ell\in\mathcal{D}^*}\overline{X}_{t\ell}) \leqslant 1$ in the denominator of (7), which may be so small for a small $\mathcal{D}^*$ as to distort the test response. As a solution to this problem, we add to the previous test a variant of it obtained by replacing the random variable $U_{t\ell^*}$ by

$$\frac{\overline{X}_{t\ell^*}\{(1-\overline{X}_{t\ell^*})^2\widehat{\sigma}^2_{t\ell^*}+\overline{X}_{t\ell^*}\sum_{\ell\in\mathcal{D}^*,\ell\neq\ell^*}\overline{X}_{t\ell}\widehat{\sigma}^2_{t\ell}\}\left(\sum_{\ell\in\mathcal{D}^*}\overline{X}_{t\ell}\right)^2}{(\widehat{\mu}^*_t)^2}. \tag{9}$$

This second test works well for an arbitrary $\mathcal{D}^*$ with a size smaller than the one of the true dominant set (the rejection percentage goes from 80% to 100%). Indeed, the above convergence in distribution to $\mathcal{N}(0,1)$ still holds under H_0, but we have eliminated the above-discussed problem since the previous factor now appears in the numerator of (9). However, for $\mathcal{D}^*$ equal to the true dominant set, the rejection percentage of the first test is lower than the one of the second test.

We can leverage the illustrated statistical test (and possibly its variant) to obtain a procedure for the detection of the dominant set of edges of a network and for an evaluation of the distance from its steady state (see Fig. 2).

As an application to real data, we consider the international trade network (ITN), also known in complex network literature as the world-trade web [17]. ITN is defined as the network of import-export relationships between world countries in a given period (usually a year). Many efforts have been devoted to analyze the structure and the dynamics of the ITN from an empirical and theoretical modeling perspective (see, for instance, [18–26]). In particular, since it is a dense weighted network, it is rather difficult to define a tractable reference case against which one can measure the specific features of the real system. Our model can contribute to fill this gap. In the context of the ITN, the nodes represent the countries and the edges represent the trade relationships among them. With regard to the weights [27], there are different possibilities. The most natural choice is to define the weight of a certain edge $\ell = i,j$ in terms of the value of the flow from i to j. As a real example, we illustrate here a short analysis, based on the first test, of the data of trades between countries in the years 1948–2000 reconstructed from COMTRADE data [28]. Our aim here is to briefly show the potentialities of the introduced model and tools. We computed for each year and for each couple of countries $\ell = i,j$ the amount of dollars for the total exports from i to j in that year. When it is nonzero, we interpreted this fact as an extraction of that edge (color) where the associated weight (number of added balls) is equal to this amount. Hence, for each edge we set the total edge weight equal to the sum of the weights associated to that edge during the considered years. We fixed $\mathcal{D}^*$ equal to the subset of the 2000 edges with the largest total edge weight (the "top 2000 edges"), and we performed the first test for $\mathcal{D}^*$ with $\alpha = 5\%$, picking up $\ell^* \in \mathcal{D}^*$ in descending order starting from the one with the largest total edge weight. Making a plot of the number of no-rejections along the whole set of ℓ^* in $\mathcal{D}^*$ (see Fig. 2, lower panel), we found that the number of no-rejections grows linearly with constant slope but at a certain point the curve starts bending. After this bending, it saturates and reaches a plateau where ℓ^* always gives a rejection, pointing out the presence of a core subset of dominant edges. For simulated data, we made the plot for different values of t in order to point out that the degree of the curvature around the turning point gives information regarding the distance from the steady state. (For simulated data, we used a very small α, around 0.05%, in order to stress the behavior of the corresponding curve.) Remarkably, performing the first test as above but for various sizes of $\mathcal{D}^*$ and also for a randomly chosen set $\mathcal{D}^*$ of the same size (see Fig. 2, upper panel), we found an "optimal" size of $\mathcal{D}^*$ for which the difference between the rejection percentages of the two cases (the case of the randomly chosen edges and the case of the "top edges") is maximal. This maximum point coincides with the turning point of the previous curve.

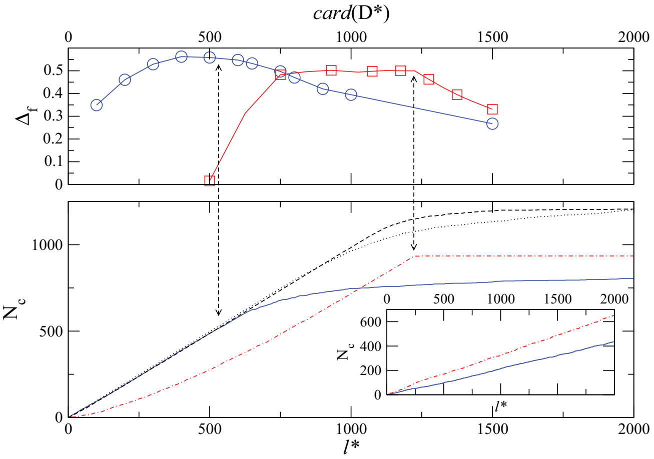

FIG. 2. (Color online) In the lower panel we plot the number of no-rejections N_c for the COMTRADE data (solid blue line) and for the simulated data of an urn with colored balls in the case of uniform distributions ($t = 10^6$ dotted and dashed red line, $t = 250\,000$ dashed black line, and $t = 125\,000$ dotted black line). For both cases we ordered the edges/colors in descending order according to the total edge-weight/number of added balls and we considered as $\mathcal{D}^*$ the set of the top 2000 edges/colors. We then executed the test taking ℓ^* in $\mathcal{D}^*$ running from the highest to the lowest value and accumulating the number of no-rejections in the y axis. After a constant no-rejection rate, the curves associated to simulations start bending, exactly in correspondence of the true dominant set (the top 1250 edges), known *a priori*. The turning point is always the same, but the higher t is, the sharper the turning point is. The curve associated to COMTRADE has a similar trend with a turning point around 500, which reveals the presence of a core subset of dominant edges (the top 500 edges). In the inset we performed the same procedure for a randomly chosen set $\mathcal{D}^*$ with size 2000 for the two collections of data. In the upper panel, we calculated for various sizes of $\mathcal{D}^*$ the difference Δ_f between the rejection percentage obtained for the randomly chosen set (averaged over 10 realizations) and the one obtained for the set of the top edges (red square for simulations and blue circle for COMTRADE), and we found a maximum where the two curves of the lower panel start bending.

In summary, we present here a model of weighted-network evolution based on a PA principle for edges [29] with random weights. We provide a theoretical framework, which accounts for the empirical evidence that many real networks grow in a heterogeneous way, generating a subset of dominant edges that controls a major share of the total weight of the network. Although the proposed methodology is suitable for weighted networks evolving according to a PA mechanism for edges, our approach is quite general and flexible in the sense that it does not require a particular probability distribution or functional form of the weights. These features and the initial parameters of the model affect the outcomes of the extractions and so the observed data implicitly depend on them, but the asymptotic results on RRU processes have allowed us to develop some tools for the detection of the set of dominant edges which do not require to exploit them. Further, the proposed procedure can serve to evaluate if the network has reached its steady state or not, a problem often encountered in assessing the relevance of the observations in complex networks. It is worthwhile to note the applicability of our method to dense networks. Finally, our model produces uncorrelated weighted networks to be used as benchmarks in order to understand the specific features of the considered networked system.

The authors acknowledge support from FET Open, Project 255987 FOC, and CNR PNR, Project "CRISIS Lab" also support from FET Project MULTIPLEX 317532.

[1] R. Albert and A. L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[2] G. Caldarelli, *Scale-Free Networks* (Oxford University Press, New York, 2007).

[3] F. Radicchi, J. J. Ramasco, A. Barrat, and S. Fortunato, Phys. Rev. Lett. **101**, 148701 (2008).

[4] D. Garlaschelli, S. Battiston, M. Castri, V. D. P. Servedio, and G. Caldarelli, Physica A **350**, 491 (2005).

[5] M. Kitsak, M. Riccaboni, S. Havlin, F. Pammolli, and H. E. Stanley, Phys. Rev. E **81**, 036117 (2010).

[6] M. Riccaboni and F. Pammolli, Research Policy **31**, 1405 (2002).

[7] S. Galluccio, G. Caldarelli, M. Marsili, and Y. C. Zhang, Physica A **245**, 423 (1997).

[8] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet* (Cambridge University Press, Cambridge, United Kingdom, 2004).

[9] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, New York, 2010).

[10] C. Roth, S. M. Kang, M. Batty, and M. Barthélemy, PLoS ONE **6**, e15923 (2011).

[11] P. Berti, I. Crimaldi, L. Pratelli, and P. Rigo, J. Appl. Probability **48**, 527 (2011).

[12] P. Berti, I. Crimaldi, L. Pratelli, and P. Rigo, Stochastic Processes and their Applications **120**, 1473 (2010).

[13] I. Crimaldi, International Mathematical Forum **4**, 1139 (2009).

[14] C. May and N. Flournoy, Annals of Statistics **37**, 1058 (2009).

[15] R. Pemantle, A., Probability Surveys **4**, 1 (2007).

[16] V. Zlatić, G. Bianconi, A. Díaz-Guilera, D. Garlaschelli, F. Rao, and G. Caldarelli, Eur. Phys. J. B **67**, 271 (2009).

[17] M. A. Serrano and M. Boguñá, Phys. Rev. E **68**, 015101(R) (2003).

[18] D. Garlaschelli and M. I. Loffredo, Phys. Rev. Lett. **93**, 188701 (2004).

[19] E. Helpman, M. Melitz, and Y. Rubinstein, NBER working paper series 12927 (2007).

[20] K. Bhattacharya, G. Mukherjee, J. Saramäki, K. Kaski, and S. S. Manna, J. Stat. Mech. (2008) P02002.

[21] G. Fagiolo, J. Reyes, and S. Schiavo, Phys. Rev. E **79**, 036115 (2009).

[22] M. Riccaboni and S. Schiavo, New J. Phys. **12**, 023003 (2010).

[23] D. Garlaschelli, A. Capocci, and G. Caldarelli, Nat. Phys. **3**, 813 (2007).

[24] D. Garlaschelli and M. I. Loffredo, Physica A **355**, 138 (2005).

[25] K. Head, Gravity for beginners (2003), available at http://economics.ca/keith/gravity.pdf.

[26] J. Tinbergen, *Shaping the World Economy: Suggestions for a International Economic Policy* (The Twentieth Century Fund, New York, 1962).

[27] A. Barrat, M. Barthélemy, and A. Vespignani, Phys. Rev. Lett. **92**, 228701 (2004).

[28] United Nations Commodity Trade Statistics Database http://comtrade.un.org/.

[29] G. Bianconi, Europhys. Lett. **71**, 1029 (2005).