Supplementary Information

Profiling core-periphery network structure by random walkers

Fabio Della Rossa, Fabio Dercole & Carlo Piccardi[1]

Politecnico di Milano, DEIB - Department of Electronics, Information and Bioengineering, I-20133 Milano, Italy

1. Network datasets

In this section we provide details on network data employed in the main text.

Star, Complete, ER and BA networks. Figure 1 of the main text considers four undirected, binary networks with $n = 100$. In the star network, nodes $1, 2, ..., n-1$ are linked to node n only, whereas in the complete network each node is linked to all the others. The Erdős-Rényi (ER) network was built by linking 200 randomly selected node pairs, so to have an average degree $\langle k \rangle = 2 \times 200/n = 4$. The Barabási-Albert (BA) network was built with the standard "preferential attachment" algorithm [1]: starting from a connected node pair, the remaining 98 nodes were iteratively added one at a time, by attaching each of their 2 edges to a node of the current network, randomly selected with a probability proportional to its current degree ($\langle k \rangle = 2 \times 197/n \approx 4$).

Monkeys. Undirected, weighted network, $n = 20$. Source: Borgatti & Everett, 1999, p. 380 [2]. As described in [3], *"These data represent 3 months of interactions among a troop of monkeys, observed in the wild by Linda Wolfe as they sported by a river in Ocala, Florida. Joint presence at the river was coded as an interaction and these were summed within all pairs"*. Our results are consistent with [2], where it is acknowledged that no significant core-periphery structure exists in this social network.

Karate. Undirected, binary network, $n = 34$. Source: Zachary, 1977, p. 456-457 [4]. The network represents the social interaction among the members of a university-based karate club from 1970 to 1972. An edge exists between two nodes when *"...the two individuals being represented consistently interacted in contexts outside those of karate classes, workouts, and club meetings..."* [4].

Netscience. Undirected, weighted network, $n = 379$. Source: Newman, 2006 [5, 6]. It is the largest connected component of the network describing the collaborations (up to year 2006) among researchers in network science, the weight of the edge connecting two researchers being proportional to the number of papers they have co-authored [5].

Airports. Directed, weighted network, $n = 2868$. Source: data downloaded from Openflights.org and processed by T. Opsahl, 2011 [7, 8]. It is the largest strongly connected component of the network describing the airports and their flight connections at the worldwide level. The weight of the (directed) edge is the number of routes between the two airports.

Internet. Undirected, binary network, $n = 11745$. Source: Newman, 2006 [9, 6]. As described in [6], it is a symmetrized snapshot of the structure of the Internet (for July 22, 2006) at the level of autonomous systems, reconstructed from BGP tables posted by the University of Oregon Route Views Project at archive.routeviews.org.

Ppi. Undirected, binary network, $n = 1458$. Source: CCNR, University of Notre Dame [10]. It is the largest connected component of the protein-protein interaction network of the yeast *Saccharomyces cerevisiae*, as analyzed in [11].

Neural. Directed, weighted network, $n = 239$. Source: elaboration by Watts and Strogatz [12] from data

[1]Correspondence and requests for materials should be addressed to C.P. (carlo.piccardi@polimi.it)

by White et al., 1986 [13]; downloaded from [6]. It is the largest strongly connected component of the neural network of the nematode worm *Caenorhabditis elegans*, where an edge joins two neurons if they are connected by either a synapse or a gap junction.

Wtn. Directed, weighted network, $n = 181$. Source: elaboration by Piccardi and Tajoli [14] from data of the Direction of Trade Statistics of the International Monetary Fund [15]. This network is the largest strongly connected component of the network of bilateral trade flows recorded in 2008 by importing countries, measured in US dollars at current prices.

2. Core-periphery profile

The *core-periphery profile* of the network α_k, $k = 1, 2, \ldots, n$, has been formally introduced in the main text of the paper. We repeat here the algorithmic definition:

Step 1 : Select at random a node i among those with minimal strength ($\sigma_i \leq \sigma_j$ for all $j \in N$). Modulo a relabeling of the nodes, we can assume, without loss of generality, that the selected node is 1. Set $P_1 = \{1\}$, hence $\alpha_1 = 0$.

Step $k = 2, 3, \ldots, n$: Select the node attaining the minimum in:

$$\alpha_k = \min_{h \in N \setminus P_{k-1}} \frac{\sum_{i,j \in P_{k-1} \cup \{h\}} \pi_i m_{ij}}{\sum_{i \in P_{k-1} \cup \{h\}} \pi_i} = \min_{h \in N \setminus P_{k-1}} \frac{\sum_{i,j \in P_{k-1}} \pi_i m_{ij} + \sum_{i \in P_{k-1}} (\pi_i m_{ih} + \pi_h m_{hi})}{\sum_{i \in P_{k-1}} \pi_i + \pi_h}. \quad (1)$$

If it is not unique, select at random one of the nodes with minimal strength σ_h among those attaining the minimum. Without loss of generality, we can assume that the selected node is k. Set $P_k = P_{k-1} \cup \{k\} = \{1, 2, \ldots, k\}$.

We now prove that, for whatever network, the core-periphery profile is a non-decreasing sequence.

Proposition: $\alpha_{k+1} \geq \alpha_k$ for all $k = 1, 2, \ldots, n-1$.

Proof: We first note that $\alpha_2 = \min_{h \in N \setminus \{1\}} (\pi_1 m_{1h} + \pi_h m_{h1}) / (\pi_1 + \pi_h) \geq 0 = \alpha_1$. Then the proposition is proved by induction if we show that, for any $k \geq 2$, $\alpha_k \geq \alpha_{k-1}$ implies $\alpha_{k+1} \geq \alpha_k$.

We preliminary observe that, for any $a, c \geq 0$ and $b, d > 0$, the following properties hold true:

(i) $(a + c)/(b + d) \geq a/b$ if and only if $c/d \geq a/b$;

(ii) $(a + c)/(b + d) \leq c/d$ if and only if $c/d \geq a/b$;

(iii) given a set $\{c_1/d_1, c_2/d_2, \ldots\}$ with $d_h > 0$ and $c_h/d_h \geq a/b$ for all $h = 1, 2, \ldots$, and letting $c_m/d_m = \min_{h \in \{1,2,\ldots\}} c_h/d_h$, we have

$$\min_{h \in \{1,2,\ldots\}} \frac{a + c_h}{b + d_h} \leq \frac{c_m}{d_m}.$$

Properties (i) and (ii) are straightforward to check, whereas property (iii) is slightly more involved: if the min in (iii) is attained by $h = m$, then property (iii) follows from (ii). If the min in (iii) is attained by $h = H \neq m$, then assuming $(a + c_H)/(b + c_H) > c_m/d_m$ would give, thanks to property (ii), $(a + c_H)/(b + c_H) > c_m/d_m \geq (a + c_m)/(b + d_m)$, which contradicts the hypothesis that $h = H$ attains the min of $(a + c_h)/(b + d_h)$.

Now, let us define the quantities

$$a^{(k)} = \sum_{i,j \in P_k} \pi_i m_{ij}, \quad b^{(k)} = \sum_{i \in P_k} \pi_i, \quad c_h^{(k)} = \sum_{i \in P_k} (\pi_i m_{ih} + \pi_h m_{hi}), \quad d_h^{(k)} = \pi_h,$$

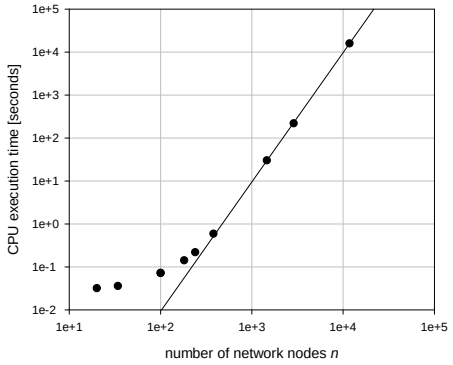

Figure 1: **Time complexity of the algorithm for deriving the core-periphery profile.** The dots correspond to the networks analyzed in the main text (see Sec. 1 of this document). The straight line with slope equal to 3 (in the log-log plane) is indicative of a time complexity $O(n^3)$ for large n. CPU execution time refers to a Matlab implementation running on a 2.9Ghz Intel Core.

so that we can write (compare with (1))

$$\alpha_k = \frac{a^{(k)}}{b^{(k)}} = \min_{h \in N \setminus P_{k-1}} \frac{a^{(k-1)} + c_h^{(k-1)}}{b^{(k-1)} + d_h^{(k-1)}}.$$

Assuming that $\alpha_k \geq \alpha_{k-1}$ is equivalent to assuming (due to property (i)) that $c_h^{(k-1)}/d_h^{(k-1)} \geq a^{(k-1)}/b^{(k-1)}$ for all $h \in N \setminus P_{k-1}$, which also guarantees (property (iii)) that $\alpha_k \leq \min_{h \in N \setminus P_{k-1}} c_h^{(k-1)}/d_h^{(k-1)}$. At the next step, we have

$$\alpha_{k+1} = \min_{h \in N \setminus P_k} \frac{a^{(k)} + c_h^{(k)}}{b^{(k)} + d_h^{(k)}},$$

where $a^{(k)}/b^{(k)} \leq \min_{h \in N \setminus P_{k-1}} c_h^{(k-1)}/d_h^{(k-1)} \leq \min_{h \in N \setminus P_k} c_h^{(k-1)}/d_h^{(k-1)}$ because $(N \setminus P_k) \subset (N \setminus P_{k-1})$, $c_h^{(k)} \geq c_h^{(k-1)}$ because $P_k \supset P_{k-1}$, and $d_h^{(k)} = d_h^{(k-1)}$. Thus, we have $\alpha_k = a^{(k)}/b^{(k)} \leq \min_{h \in N \setminus P_k} c_h^{(k)}/d_h^{(k)}$, which is obviously less than or equal to $c_h^{(k)}/d_h^{(k)}$ for all $h \in N \setminus P_k$. Property (i) then guarantees that $\alpha_{k+1} \geq \alpha_k$. □

3. Computational issues

Computational complexity. The straightforward implementation of the above algorithm, with the exhaustive search of the node h attaining the minimum in (1), has time complexity $O(n^3)$. This is confirmed by numerical tests performed on the same pool of artificial and real-world networks considered in the main text (see Sec. 1 of this document), and summarized in Fig. 1 (the Matlab code can be requested to the corresponding author). It is presumable that the time requirement could be considerably reduced, perhaps at the price of a (mild) suboptimality. For instance, when examining the set of p-nodes (the periphery in the strict sense) for which the core-periphery profile is 0, one could stop the min search in (1) as soon as a node k is found such that α_k is still 0. Given that many networks have a large periphery, this would imply a dramatic decrease in the time requirement. Research on this and other possible numerical improvements is in progress.

Robustness to randomicity in the core-periphery profile algorithm. In the **Methods** section of the main text, it is pointed out that the above algorithm may have some randomicity when, at step 1, many nodes share the minimum strength σ_i and when, at step k, the minimum of α_k is equivalently attained by

network	$\min(C)$	$\max(C)$	$\text{mean}(C)$	$\text{std}(C)/\text{mean}(C)$
monkeys	0.261	0.265	0.263	8.25×10^{-3}
karate	0.709	0.713	0.711	3.26×10^{-3}
netscience	0.644	0.645	0.644	1.11×10^{-4}
ppi	0.767	0.768	0.768	4.25×10^{-4}
airports	0.823	0.824	0.824	8.58×10^{-5}
internet	0.942	0.942	0.942	4.31×10^{-5}
neural	0.940	0.940	0.940	0
wtn (binary)	0.349	0.349	0.349	0
wtn (weighted)	0.819	0.819	0.819	0

Table 1: **Results of the randomization of the core-periphery profile algorithm.** For each network, the algorithm is run 100 times and the corresponding cp-centralization C is computed.

many nodes having the same strength. To assess the impact of this randomicity, we run the algorithm 100 times for each of the real-world networks considered in the paper, and we computed the statistics of the corresponding cp-centralization C. The results are reported in Table 1. The computation of C appears to be very robust, as the ratio $\text{std}(C)/\text{mean}(C)$ ranges from 0 to 8.25×10^{-3}, with a tendency of being even smaller for larger n.

Exact vs. approximate α-periphery. In the main text, the sets yielded by the core-periphery profile algorithm are proposed as heuristic approximations of the α-periphery, which is, by definition, the largest subnetwork S with $\alpha_S \leq \alpha$. More precisely, we take the largest P_k such that $\alpha_k \leq \alpha$ as our approximation of the α-periphery. It is not possible, in general, to assess the quality of such an approximation, since the problem of finding the α-periphery falls in a class known to be computationally untractable [16]. But we can do it on very small networks, where the exact α-periphery can be computed by exhaustively enumerating all the subnetworks.

Figure 2 reports the results of the analysis for three networks: the toy-network discussed in Fig. 4 of the main text ($n = 16$), an ER network with $n = 20$, and a BA network also with $n = 20$ (see Sec. 1 of this document for details). For each network, we compute the persistence probabilities of all the 2^n possible subnetworks (they are more than 10^6 when $n = 20$), and we put a dot at coordinates (k, α_S) if α_S is the persistence probability of a k-node subnetwork. For a given α, we obtain an exact α-periphery by taking one of the rightmost dots (i.e., one of those with largest k) falling not above the horizontal line $\alpha_S = \alpha$ (black dots in Fig. 2). Conversely, if we denote by (k, α_k^*) the coordinates of the lowest point having abscissa k, we can say that the subnetwork corresponding to (k, α_k^*) is an α-periphery for all $\alpha_k^* \leq \alpha < \alpha_{k+1}^*$. Therefore, assessing the quality of the approximation boils down to measuring the difference between the curves α_k^* (i.e., the "exact" core-periphery profile) and α_k (i.e., our "approximate" core-periphery profile proposed in the main text). Figure 2 points out that, in the cases here considered, the two curves are very close or even coincident.

References

[1] Barabási, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

[2] Borgatti, S. & Everett, M. Models of core/periphery structures. *Soc. Networks* **21**, 375–395 (1999).

[3] Borgatti, E. M., S.P. & Freeman, L. UCINET software (accessed July 10, 2012). URL `https://sites.google.com/site/ucinetsoftware/`.

[4] Zachary, W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977).

[5] Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74**, 036104 (2006).

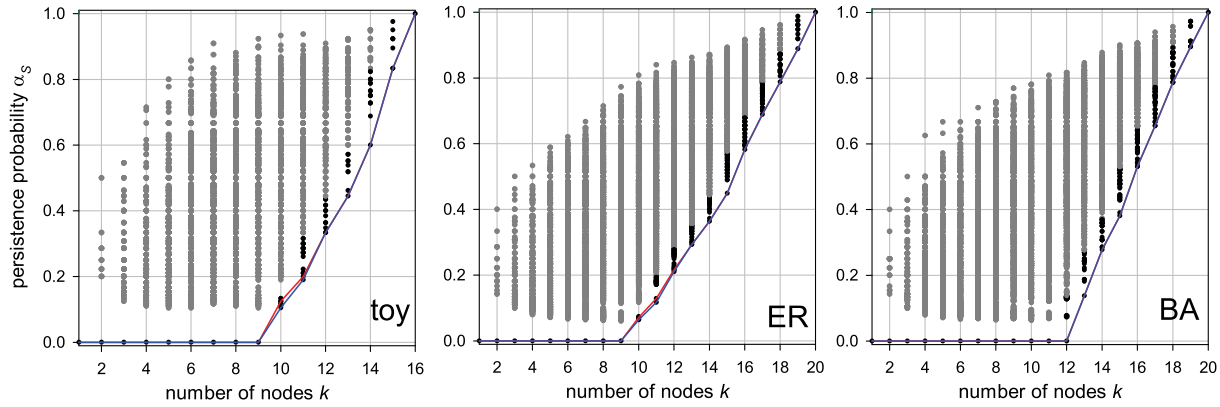

Figure 2: **Exact vs. approximate α-periphery.** Each dot (k, α_S) corresponds to one of the 2^n possible subnetworks, having k nodes and persistence probability α_S. Black dots correspond to subnetwork which are α-periphery for some α (opposite to grey ones). The blue curve connects the subnetworks having, for each k, the smallest α_S ("exact" core-periphery profile); the red curve is our "approximate" core-periphery profile (in the rightmost panel the two curves are coincident). The three panels refer, respectively, to the toy-network discussed in Fig. 4 of the main text, an ER network, and a BA network.

[6] Newman, M. E. J. Network data (accessed June 6, 2012). URL `http://www-personal.umich.edu/~mejn/netdata/`.

[7] Opsahl, T. Why Anchorage is not (that) important: binary ties and sample selection (2011). URL `http://toreopsahl.com/2011/08/12/`.

[8] Openflights.org. Airport, airline and route data (accessed August 12, 2011). URL `http://openflights.org/data.html`.

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

[10] Interdisciplinary Center for Network Science and Applications, University of Notre Dame (accessed July 13, 2012). URL `http://www.nd.edu/~networks/resources.htm`.

[11] Jeong, H., Mason, S., Barabasi, A. & Oltvai, Z. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).

[12] Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).

[13] White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. The structure of the nervous-system of the nematode Caenorhabditis-elegans. *Philos. Trans. R. Soc. Lond. Ser. B - Biol. Sci.* **314**, 1–340 (1986).

[14] Piccardi, C. & Tajoli, L. Existence and significance of communities in the world trade web. *Phys. Rev. E* **85**, 066119 (2012).

[15] International Monetary Fund. Direction of Trade Statistics (accessed Autumn 2010). URL `http://www.imf.org/external/data.htm/`.

[16] Sima, J. & Schaeffer, S. On the NP-completeness of some graph cluster measures. In Wiedermann, J and Tel, G and Pokorny, J and Bielikova, M and Stuller, J (ed.) *SOFSEM 2006: Theory and Practice of Computer Science, Proceedings*, vol. 3831 of *Lecture Notes in Computer Science*, 530–537 (2006).