Spectral Coarse Graining of Complex Networks

David Gfeller and Paolo De Los Rios

Laboratoire de Biophysique Statistique, SB/ITP, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015, Lausanne, Switzerland

Reducing the complexity of large systems described as complex networks is key to understanding them and a crucial issue is to know which properties of the initial system are preserved in the reduced one. Here we use random walks to design a coarse graining scheme for complex networks. By construction the coarse graining preserves the slow modes of the walk, while reducing significantly the size and the complexity of the network. In this sense our coarse graining allows us to approximate large networks by smaller ones, keeping most of their relevant spectral properties.

One of the most difficult hurdles in the analysis and visualization of large complex networks [1–4] is, understandably, their sheer size. Given that most algorithms used to extract information from a network topology run in times that grow polynomially with the number N of network nodes, even networks of a few thousands nodes can represent a challenge, and networks with $N > 10^5$ becomes sometimes impossible to deal with.

A promising way around this problem is to *coarse grain* the network, i.e., to reduce the number of nodes and edges by means of a mapping of the network with N nodes and E edges into a smaller network with $\tilde{N}$ nodes and $\tilde{E}$ edges. $\tilde{N}$ and $\tilde{E}$ have to be small enough to be amenable to analysis and visualization.

Several coarse graining schemes have been proposed in the literature. The k-core decomposition, which is a node decimation approach, was first proposed in [5] to isolate the central core of a network, and was shown to be extremely effective for visualization purposes [6]. Alternatively, the number of nodes can be reduced by clumping them together in clusters. A widely accepted technique is based on community detection [7]. Within this framework, groups of nodes with more edges pointing to each other then to the rest of the network are considered as one single unit. After grouping the nodes a much reduced "network of clusters" is obtained, representing the functional units of a network. Because of the importance and the complexity of finding meaningful clusters, a very large number of clustering algorithms have been developed recently [7–13]. However, there is often no clear statement on whether properties of the initial network are preserved in the network of clusters.

In this respect, the box-covering technique recently introduced by Song, Havlin, and Makse [14] and further analyzed by Goh, *et al.* [15] deserves a special mention: after covering *self-similar* networks with suitably defined boxes of a given size, the new networks obtained by substituting each box with a node preserve some of the topological features of the original ones. Song, Havlin, and Makse have thus recognized that network reduction should go hand-in-hand with the preservation of some relevant

network properties, akin to the renormalization group in statistical physics. What properties should be preserved is instead an unsettled issue: although the network topology can provide important clues about the organization of the system under scrutiny, it does not necessarily bear any insight on the internal dynamics of the system.

In this Letter we introduce a mathematical framework to coarse grain networks, based on the idea of grouping nodes together. Thus our starting point is similar to clustering approaches. However, contrary to the clustering paradigm of identifying the "correct" communities in a network, our goal is to obtain a reduced network that preserves some properties of the initial one. Here the properties of interest will be the main characteristics of random walks on networks [16].

Random walks play a central role in a large number of dynamical processes taking place on complex networks. Their evolution is described by a stochastic matrix W. If A is the adjacency matrix, then $W_{ij} = A_{ij}(\sum_l A_{lj})^{-1}$ gives the transition probability from j to i. W has several interesting properties. In particular for a connected undirected network, the Perron-Frobenius theorem states that the largest eigenvalue λ^1 is equal to 1. The right eigenvector $|p^1\rangle$ associated to λ^1 describes the stationary state and the left eigenvector $\langle u^1|$ is constant. Typically, eigenvectors with eigenvalues close to 1 capture the large-scale behavior of the random walk. Since a coarse-grained network necessarily loses the fine details of the original one, expectably, our goal is to preserve the large-scale behavior of the random walk, hence its largest eigenvectors and eigenvalues.

As a starting point, we want to ensure that two nodes in an undirected network (say nodes 1 and 2) having exactly the same neighbors are grouped together, since they cannot be distinguished from the point of view of a random walk. In terms of a left eigenvector $\langle u^\alpha|$ of W it means that $u_1^\alpha = u_2^\alpha$ for any $\lambda^\alpha \neq 0$ since column 1 and 2 of W are equal. The obvious coarse graining step is to coalesce such pair of nodes, with the resulting new node carrying the sum of the edges of the initial ones. The new network in which node 1 and 2 have been merged is characterized by a $(N-1) \times (N-1)$ adjacency matrix $\tilde{A}$, with the first line(column) of

$\tilde{A}$ being the sum of the two first lines(columns) of A. On this reduced network, the stochastic matrix $\tilde{W}$ describing a random walk is obtained by normalizing the columns of $\tilde{A}$.

At this stage, it will be useful to write $\tilde{W}$ as a product of three matrices,

$$\tilde{W} = RWK.$$

K and R are two projectionlike operators from the N-dimensional space of the initial nodes to the $(N-1)$-dimensional space of the new nodes. In order to fulfill the definition of $\tilde{W}$ and using that $p_j^1 \propto \sum_i A_{ij}$ for undirected networks, K and R are defined as:

$$K = \begin{pmatrix} \frac{p_1^1}{p_1^1+p_2^1} & 0\dots0 \\ \frac{p_2^1}{p_1^1+p_2^1} & 0\dots0 \\ 0 & \\ \vdots & I_{N-2} \\ 0 & \end{pmatrix} \quad \text{and} \quad R = \begin{pmatrix} 1 & 1 & 0\dots0 \\ 0 & 0 & \\ \vdots & \vdots & I_{N-2} \\ 0 & 0 & \end{pmatrix},$$

(see [17,18] for a similar mathematical framework). The interesting features of $\tilde{W}$ come from the property [19] that if $u_1^\alpha = u_2^\alpha$, the vector $\langle u^\alpha|K$ is a left eigenvector of $\tilde{W}$ with eigenvalue λ^α (i.e., $\langle u^\alpha|K = \langle \tilde{u}^\alpha|$). To obtain this result one simply needs to see that $\langle u^\alpha|KR = \langle u^\alpha|$ if $u_1^\alpha = u_2^\alpha$. In the case of undirected networks, the result can be extended to the right eigenvectors. Under the same hypothesis ($u_1^\alpha = u_2^\alpha$), the vector $R|p^\alpha\rangle$ is a right eigenvector of $\tilde{W}$ with eigenvalue λ^α (for $\alpha = 1$ the result holds as well in directed networks). Moreover, we could show analytically that a perturbative approach can be carried out. If $|u_1^\alpha - u_2^\alpha| \propto \epsilon$ for a given $\lambda^\alpha \neq 0$, $\langle u^\alpha|K, R|p^\alpha\rangle(\lambda^\alpha)$ need to be corrected by vectors (a scalar) scaling as ϵ in order to become left and right eigenvectors(eigenvalue) of $\tilde{W}$.

To summarize we have introduced a mathematical framework such that grouping nodes with similar components in $\langle u^\alpha|$ has a spectral interpretation: it preserves the eigenvalue λ^α, averages the components of $\langle u^\alpha|$ and for undirected networks sums up the components of $|p^\alpha\rangle$.

For simplicity, the case where only two components of an eigenvector are equal (close to each other) has been considered. It is straightforward to generalize the grouping to all nodes having the same components (components close to each other) in $\langle u^\alpha|$. Groups are first labeled from 1 to $\tilde{N}$ and $\delta_{C,i}$ is defined as 1 if node i belongs to group C, 0 otherwise ($C = 1 \dots \tilde{N}$). Then K and R read $R_{Ci} = \delta_{C,i}$ and $K_{iC} = \frac{p_i^1}{\sum_{l \in C} p_l^1} \delta_{C,i}$, with R a $\tilde{N} \times N$ matrix and K a $N \times \tilde{N}$ matrix.

The method can be further extended to more than one eigenvector and groups are defined as the nodes having the same components (components close to each other) over a set of left eigenvectors $\{\langle u^{\alpha_i}|\}_{i=1}^S$. Choosing the S first nontrivial eigenvectors ensures to conserve the slow modes of the random walk.

Spectral properties of W have been used previously in spectral clustering techniques [19]. However, the properties derived above show that there exists a way to preserve the spectral properties of a network while reducing its size, which is the main objective of any coarse graining strategy.

To illustrate our coarse graining scheme, we applied it to the di-alanine folding network studied in [20], considered as undirected. The network was built from molecular dynamics simulations of a di-alanine peptide and consists of 1832 nodes [Fig. 1(a)]. A node in Fig. 1(a) accounts for a configuration sampled during the simulation and edges represent transitions between configurations [21]. The weight of an edge between two configurations is equal to the total number of transitions sampled during the simulation. In a previous work [20], the di-alanine folding net-
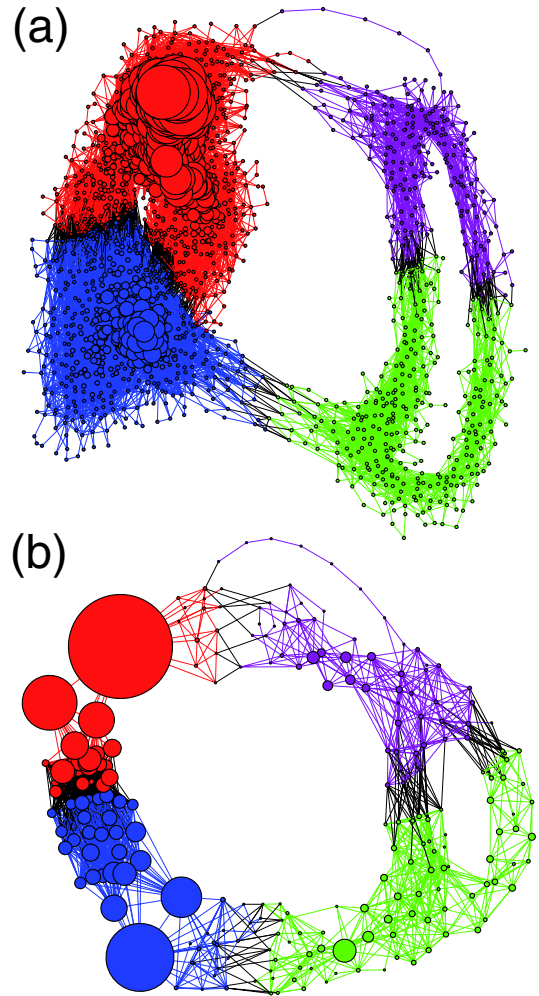

FIG. 1 (color online). (a) Di-alanine folding network ($N = 1832$). Node size is proportional to their weight (i.e., the number of times nodes have been visited in the simulation). The four different colors correspond to the clusters found in [20]. (b) Coarse-grained network ($\tilde{N} = 227$) according to $\langle u^\alpha|$, $\alpha = 2, 3, 4$. Node size is proportional to the total weights of the groups. Colors correspond to the clusters in which the nodes of each group have been classified.

work was shown to consist of four main clusters (colors in Fig. 1), corresponding to the four main energy basins of the underlying free-energy landscape. Random walks on such networks are representative of peptide dynamics since the elements of W correspond to the effective transition probabilities, as observed along the simulation.

To coarse grain the network, we have used the first three nontrivial left eigenvectors $\langle u^2|$, $\langle u^3|$ and $\langle u^4|$ of W. Along each eigenvector, $I = 60$ intervals of equal size have been defined between the highest and the lowest component. Nodes have been grouped together if they belonged to the same interval along the three eigenvectors. In this way 227 nonempty groups have been found. The coarse-grained network is shown in Fig. 1(b). Colors were set according to the clusters of the nodes in each group. Clearly the coarse-grained network is not equivalent to the network of clusters. Although the nodes of a group do not necessarily belong to the same cluster, this situation happened only for 4 groups (representing 15 nodes) out of the 227. We also applied on the coarse-grained network the same clustering algorithm [22] used to identify the clusters in Fig. 1(a). Exactly 4 clusters were obtained corresponding to more than 98% of the initial nodes correctly classified. Thus, even if the aim of our coarse graining approach is different than the usual clustering, the results are indeed consistent with the global features revealed by the cluster structure of the network. Moreover, the cluster structure is robust under coarse graining.

As expected from our perturbative derivation, the first eigenvalues are preserved in the coarse-grained network with high accuracy (Table I A, columns 2 and 3). Moreover, the normalized scalar product (Table I A, columns 4 and 5) shows that the projected eigenvectors $\langle u^\alpha|K$ and $R|p^\alpha\rangle$ are almost equal the corresponding eigenvectors of $\tilde{W}$. Similar results have been obtained considering the giant component of an Erdös-Rényi network [23] ($N = 5626$, $\langle k \rangle = 2$, Table I B) and a Barabási-Albert network

[24] ($N = 6005$, $m = 1$, Table I C), always considering the three first nontrivial left eigenvectors $\langle u^2|$, $\langle u^3|$ and $\langle u^4|$ and $I = 60$. The general agreement indicates that our perturbative approach is robust for various kinds of networks even if components in $\langle u^\alpha|$ are not equal but close to each other within the groups.

Figure 1 hints that the global architecture of the coarse-grained network is representative of the original one. For instance, most of the nodes buried in the center of the red cluster form one single group, while the nodes lying along the few paths connecting the red and violet clusters, and therefore critical for the network global connectivity, are well preserved. A more stringent test is done by comparing the mean first passage time (MFPT) from node j to node i, T_{ij}. In the context of transport phenomena or search on a network, MFPT is an important characteristic of random walks [16,25]. To compute it exactly, one usually considers node i as a sink and uses the stochastic matrix $\hat{W}$ with the ith column set to 0 ($T_{ij} = \sum_{t=0}^{\infty} t(\hat{W}^t)_{ij} = \sum_{\alpha=1}^{N} \hat{u}_i^\alpha \hat{p}_j^\alpha \frac{\hat{\lambda}^\alpha}{(1-\hat{\lambda}^\alpha)^2}$). To compare the MFPTs, we used the coarse graining shown in Fig. 1(b), defining the sink node i as a single group. Figure 2 shows with black circles (○) the average MFPT to node i for the original network. The MFPT to the group consisting of node i in the coarse-grained network is displayed with red lines. The excellent overlap indicates that the MFPT is extremely well preserved, whereas this is not the case in the network of clusters (see insets in Fig. 2). Hence the coarse-grained network is representative of the general features of the diffusion process in the initial network. Moreover, this finding was shown to be robust if other eigenvectors are included, as long as the size of the intervals is kept small enough. In this respect the value $I \propto \epsilon^{-1}$ tunes the degree of precision: increasing I improves the agreement between the initial and the

TABLE I. Columns 2 and 3: the three largest (nontrivial) eigenvalues of the stochastic matrices W and $\tilde{W}$. Column 4: Scalar product between $\langle u^\alpha|K$ and $\langle \tilde{u}^\alpha|$ for the three left eigenvectors used in the coarse graining procedure. Column 5: Scalar product between $R|p^\alpha\rangle$ and $|\tilde{p}^\alpha\rangle$ for the three right eigenvectors. Box A: Di-alanine network shown in Figs. 1(a) and 1(b). Box B: Erdös-Rényi network. Box C: Barabási-Albert network.

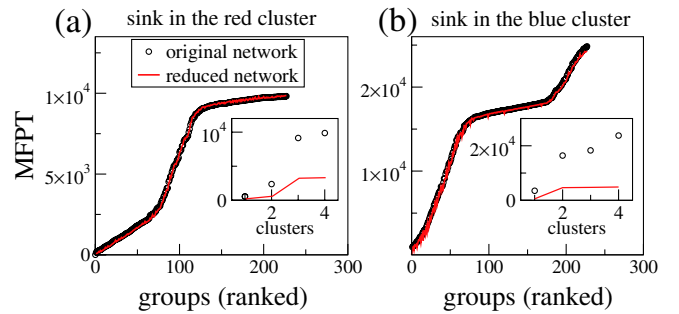|	α	λ^α	$\tilde{\lambda}^\alpha$	$\frac{\langle u^\alpha|K|\tilde{u}^\alpha\rangle}{\|\langle u^\alpha|K\|\cdot\|\tilde{u}^\alpha\|}$	$\frac{\langle \tilde{p}^\alpha|R|p^\alpha\rangle}{\|R|p^\alpha\rangle\|\cdot\|\tilde{p}^\alpha\|}$
A	2	0.999 87	0.999 87	0.9999	0.9999
	3	0.999 47	0.999 44	0.9998	0.9999
	4	0.997 85	0.997 80	0.9999	0.9999
B	2	0.989 55	0.989 22	0.9985	0.9941
	3	0.989 01	0.988 61	0.9989	0.9924
	4	0.987 79	0.987 41	0.9885	0.9686
C	2	0.999 71	0.999 71	0.9999	0.9999
	3	0.999 34	0.999 33	0.9994	0.9988
	4	0.999 17	0.999 16	0.9998	0.9997

FIG. 2 (color online). Comparison of the MFPT. The circles (○) represent the average MFPT ranked for each group in the original network (variances are not shown since they are always smaller than the size of the circles). The MFPT of the corresponding nodes in the coarse-grained network is displayed with red lines. (a) di-alanine folding network with the sink i as the heaviest node of the red cluster. (b) di-alanine folding network with the sink i as the heaviest node of the blue cluster. Insets: Comparison of the MFPT between the original network (○) and the network of clusters (red line).

P H Y S I C A L R E V I E W L E T T E R S

coarse-grained network, but in the same time results in a larger $\tilde{N}$.

In general the large eigenvalues and eigenvectors of W represent the large-scale behavior of random walks. However, in some cases eigenvectors are directly associated with useful quantities. This is the case with the PAGERANK matrix [26]. PAGERANK is defined via a stochastic process on the World Wide Web, where for each step a "random surfer" either follows with probability d one of the existing outgoing links or jumps at random to another site with probability $(1 - d)$. The PAGERANK of the nodes corresponds to the stationary state of the process (i.e., to $|p^1\rangle$). Thus under coarse graining the PAGERANK in the reduced network is the sum over the PAGERANK in the initial network. The only effect of the directed nature of the WWW, compared to undirected network, is that nothing ensures that $|p^\alpha\rangle$ is preserved when coarse graining along $\langle u|^\alpha$ for $\alpha > 1$.

As a second example, we consider the exit probability on a network with two nodes (say node 1 and N) in which the random walk is trapped. One can show that the exit probability in node 1, respectively, N, starting at j can be expressed as the left eigenvector $\langle u_j^1|$, respectively, $\langle u_j^2|$, of the stochastic matrix describing the transition probabilities and including the two traps. Hence coarse graining the network according to the eigenvector $\langle u^1|(\langle u^2|)$, is equivalent to preserving the exit probabilities in node 1, (N). In the case of a network describing the dynamics of a peptide, as the example studied in this Letter, the exit probability can be associated with the p fold [27], defined as the probability to reach the native state before the denaturated state. If the two traps are chosen as representatives of the native and denaturated states (for instance the heaviest nodes of the two main clusters of Fig. 1), our method allows us to coarse grain the network in such a way that the p fold is perfectly preserved for every configuration (see [28] for a related coarse graining framework considering a continuous diffusion process).

In conclusion, we have defined a mathematical framework for coarse graining complex networks based on random walks. This operation has the intrinsic property of preserving the first eigenvalues and the corresponding eigenvectors. In this sense it can be regarded as a decimation of the fast modes, without altering the slow modes, akin to k-space coarse graining, and eventually coming back to a real-space coarse-grained network. Moreover, we have shown for a network on which random walks have a physical interpretation that the coarse graining provides a highly representative approximation of the initial network, giving rise to a way to circumvent the large size of complex networks for their analysis and visualization. Finally, from a computational point of view, the first eigenvectors are fast to calculate with the existing optimized routines for sparse matrices. Therefore our method can be easily utilized on large networks.

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[3] S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).
[5] B. Bollobás, in *Graph Theory and Combinatorics* (Academic, London, 1984), p. 35.
[6] I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani, arXiv:cs.ni/0511007.
[7] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[8] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).
[9] L. Donetti and M. A. Muñoz, J. Stat. Mech. (2004) P10012.
[10] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).
[11] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[12] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).
[13] A. Capocci, V. D. P. Servedio, G. Caldarelli, and F. Colaiori, Physica (Amsterdam) **352D**, 669 (2005).
[14] C. Song, S. Havlin, and H. A. Makse, Nature (London) **433**, 392 (2005).
[15] K.-I. Goh, G. Salvi, B. Kahng, and D. Kim, Phys. Rev. Lett. **96**, 018701 (2006).
[16] J. D. Noh and H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004).
[17] S. Lafon and A. B. Lee, IEEE Trans. Pattern Analysis Machine Intelligence **28**, 1393 (2006).
[18] B. Gaveau and L. S. Schulman, Bull. des Sci. Math. **129**, 631 (2005).
[19] M. Meila and J. Shi, in *AI and Statistics (AISTATS), Key West, Florida, 2001* (unpublished).
[20] D. Gfeller, P. De Los Rios, A. Caflisch, and F. Rao, Proc. Natl. Acad. Sci. U.S.A. **104**, 1817 (2007).
[21] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).
[22] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, Nucleic Acids Res. **30**, 1575 (2002).
[23] P. Erdös and A. Rényi, Publ. Math. Debrecen **6**, 290 (1959).
[24] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[25] A. Baronchelli and V. Loreto, Phys. Rev. E **73**, 026103 (2006).
[26] S. Brin and L. Page, in *Proceedings of the 7th International WWW Conference* (Elsvier Science Publisher B. V., Amsterdam, 1998), p. 107.
[27] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[28] Y. M. Rhee and V. S. Pande, J. Phys. Chem. B **109**, 6780 (2005).