

# K-Scaffold subgraphs of Complex networks

Bernat Corominas-Murtra<sup>1</sup>, Sergi Valverde<sup>1</sup>, Carlos Rodríguez-Caso<sup>1</sup> Ricard V. Solé<sup>1,2</sup>

<sup>1</sup> ICREA-Complex Systems Lab, Universitat Pompeu Fabra, Dr. Aiguader 80, 08003 Barcelona, Spain

<sup>2</sup> Santa Fe Institute, 1399 Hyde Park Road, New Mexico 87501, USA

Complex networks with high numbers of nodes or links are often difficult to analyse. However, not all elements contribute equally to their structural patterns. A small number of elements (the hubs) seem to play a particularly relevant role in organizing the overall structure around them. But other parts of the architecture (such as hub-hub connecting elements) are also important. In this letter we present a new type of substructure, to be named the  $K$ -scaffold subgraph, able to capture all the essential network components. Their key features, including the so called critical scaffold graph, are analytically derived.

*Introduction.* Networks pervade complexity [1–4]. How networks are organized at different scales is one of the main topics of complex network research [5–10]. Some approaches are based on the study of given subgraphs, from the smaller network motifs [7, 8, 11] to  $K$ -cores [6, 12], spanning trees [10] or gradient subgraphs obtained from a given internal system’s dynamics [9].

One of the most studied subgraphs is the so-called  $K$ -core, formally defined by Bollobàs in [13]. The  $K$ -core of a graph  $\mathcal{G}$ ,  $\mathcal{C}_k(\mathcal{G})$  is the largest subgraph whose vertices have, at least, degree  $k \geq K$ . The behaviour of such subgraph, and its percolation properties have been widely studied [6, 13–16].  $K$ -cores display interesting features with several implications in the study of real networks, both at the theoretical and applied level [6, 12, 14, 17].

Hubs are the center of attention of the  $K$ -core. They are responsible for the efficient communication among network units and their failure or removal can have dramatic consequences [18]. But other graph components are also relevant to understand network behavior. In particular, hubs are often related through other elements exhibiting low connectivity, the so-called *connectors*. Despite its relevance, the  $K$ -core fails in finding the hub-connector structure. This pattern is essential in highly disassortative or modular networks, where hub-hub connectors play a crucial role [2]. In such networks, robustness against failures is strongly tied to hubs, but also to the hub-connector structure. Moreover, connectors can display high *betweenness centrality* [2] despite their low connectivity, reinforcing the role of this kind of nodes in non-local organization of the global topology and dynamics.

To overcome these limitations, we introduce a subgraph definition which captures the previous traits. Specifically, we consider a subgraph that includes the most connected nodes and their connectors, if any. In doing so, we want to explore whether there is some fundamental hub-connector subgraph and its relevant properties. Such a graph, the so-called  $K$ -scaffold subgraph, was recently introduced (in qualitative terms) within the context of the human proteome [19]. This network included only transcription factors, i. e. proteins linking to DNA and thus involved in regulating gene expression (fig 1(a)). Specifically, it was shown that an appropriate choice of relevant hubs and their connectors allowed to define a

functionally meaningful subgraph. Such subgraph contained a large number of cancer-related proteins around which well-defined modules were organized as evolutionarily and functionally related subsets. Here, we define this subgraph in a rigorous way. We analytically characterize its properties and degree distributions as well as the presence of a special class of minimal scaffold graph based on a critical percolation threshold.

*K-Scaffold subgraphs.* Let us consider a graph  $\mathcal{G}(V, \Gamma)$ , where  $V$  is the set of nodes and  $\Gamma$  the set of edges connecting them. The  $K$ -Scaffold subgraph  $S_K(\mathcal{G})$  will consider the degree of nodes  $k(e_i)$ ,  $e_i \in E$  but it will take into account correlations: Specifically, if we choose a node  $e_i \in V$ , it will belong to  $S_K(\mathcal{G})$  if and only if (1)  $K \leq k_i$  or (2)  $e_i$  is connected to  $e_j \in V$ , and  $e_j$  is such that  $K \leq k_j$ . Thus, given a graph  $\mathcal{G}(V, \Gamma)$ , with its adjacency matrix  $a_{ij}$ , the  $K$ -scaffold of  $\mathcal{G}$  will be defined as:

$$S_{ij} = \begin{cases} a_{ij} & \text{iff } (K \leq k_i \vee K \leq k_j) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

An example of such  $K$ -scaffold subgraph is shown in (fig.(1b)). This allows us to define, from a given graph  $\mathcal{G}$  a nested hierarchy of subgraphs  $S_K(\mathcal{G})$  such that:

$$\dots S_{K+1}(\mathcal{G}) \subseteq S_K(\mathcal{G}) \subseteq S_{K-1}(\mathcal{G}) \dots \quad (2)$$

To clarify which elements are really relevant, we also define a *naked*  $K$ -scaffold subgraph. From the  $K$ -scaffold subgraph,  $S_K(\mathcal{G})$ , the *naked*  $K$ -scaffold,  $\gamma_K(\mathcal{G})$ , is obtained by removing all nodes having a single link (the “hair” of the graph)(fig.(1c)). Thus, from  $S_{ij}$ , it is easy to compute the adjacency matrix of the *naked*  $K$ -scaffold subgraph, namely:

$$\gamma_{ij} = S_K(a_{ij})(1 - \delta_{k_i,1})(1 - \delta_{k_j,1}) \quad (3)$$

Additionally, if two or more connectors have identical pattern of connectivity in  $\gamma_K(\mathcal{G})$  (i.e., they are connected to exactly the same hubs, understanding hubs as nodes with  $k \geq K$ ), we renormalize these sets of connectors by replacing each of them with a single node. In this way,

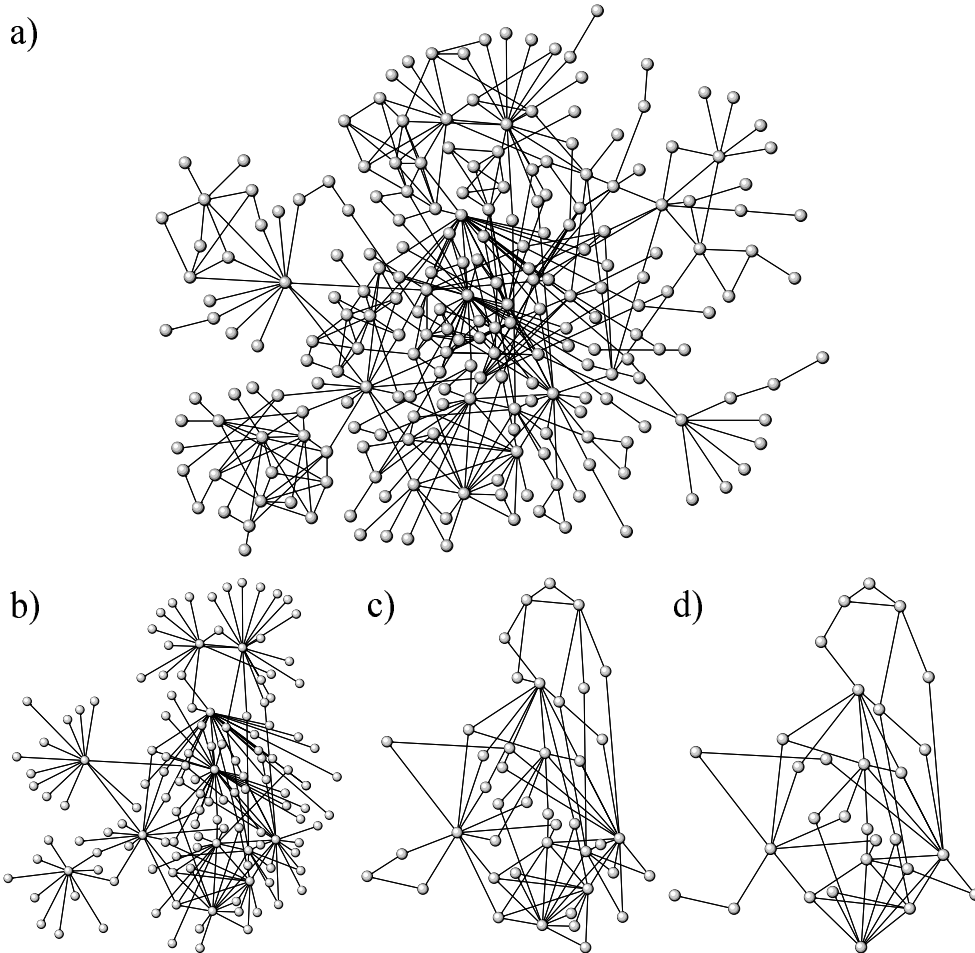


FIG. 1: (a) The Human Transcription Factor interaction Network (HTFN). (b) its 11-Scaffold subgraph. (c) The naked 11-scaffold subgraph and (d) the naked and renormalized 11-scaffold subgraph. The  $K$ -scaffold subgraph displays a fundamental hub-connector structure that organizes the general topology of the whole system. Data from [19].

the renormalized  $K$ -scaffold subgraph,  $\overline{\gamma}_K(\mathcal{G})$ , keeps the relevant elements without redundancies (fig.(1d)).

*Statistical Properties.* Here we derive the main statistical features of the  $K$ -scaffold subgraph from an arbitrary, uncorrelated network  $\mathcal{G}$ . First, we compute the fraction of nodes in  $\mathcal{G}$  belonging to  $S_K(\mathcal{G})$ , i.e., the probability for a random chosen node of  $\mathcal{G}$  to belong to  $S_K(\mathcal{G})$ . If we define

$$q_{<K} = \sum_{k < K} \frac{kP(k)}{\langle k \rangle_{\mathcal{G}}} \quad (4)$$

We can define  $f$  as:

$$f = 1 - \sum_{k < K} P(k) (q_{<K})^k \quad (5)$$

Where we have to read the second term of equation (5) as *the probability to find a node with  $k < K$  such that all of its  $k$  first neighbours have  $k' < K$ .*

We could consider or not links connecting connectors, i.e., nodes with  $k < K$  but connected to nodes with  $k \geq K$ . In order to simplify algorithmic procedures, it is reasonable to avoid such connector-connector links. If we avoid these class of links, the probability for a randomly chosen link to belong to the  $K$ -scaffold is, simply:

$$h = 1 - (q_{<K})^2 \quad (6)$$

However, for mathematical consistency, it is necessary to take into account such kind of links. If we do so, the probability for a randomly chosen link to belong to the  $K$ -scaffold is:

$$h = 1 - \left( \sum_{k < K} q(k) (q_{k < k})^{k-1} \right)^2 \quad (7)$$

To complete our characterization, we find the degree distribution of  $S_K(\mathcal{G})$ . To do the job, we need to define the

probability that a node with degree less than  $K$  is connected to exactly  $k$  nodes whose connectivity is equal or higher than  $K$ , namely:

$$g(k, K) = \sum_{k < i < K} P(i)(1 - q_{<K})^k (q_{<K})^{i-k} \quad (8)$$

The probability distribution for  $k$ 's above  $K$  in the  $K$ -scaffold is the same probability distribution of the substrate graph  $\mathcal{G}$ , multiplied by a normalization factor  $f$ . To see this, we can see that, from the definition of  $g(k, K)$  we have:

$$\sum_{k < K} g(k, K) + \sum_{k \geq K} P(k) = \frac{|N|_{S_K(\mathcal{G})}}{|N|_{\mathcal{G}}} = f \quad (9)$$

Thus, the normalized degree distribution of  $S_K(\mathcal{G})$  will be:

$$P_{S_K}(k) = \begin{cases} g(k, K)/f, & \text{iff } k < K \\ P(k)/f, & \text{otherwise} \end{cases} \quad (10)$$

*Minimal  $K$ -Scaffold subgraphs.* The previous definitions refer to  $K$ -dependent subgraphs, being  $K$  arbitrary. But we can ask if some specially relevant  $K$  value is involved. In other words, since larger  $K$  values support smaller scaffold graphs, we might ask what is the limit in this process and what is kept before the network is fragmented or too small. The question we are addressing concerns the existence of a characteristic scale. The *minimal* scaffold subgraph will label the minimal substructure capturing the fundamental hub-connector architecture of the net, if it exists. This subcritical subgraph will be located immediately above of the percolation threshold of  $\mathcal{G}$ , considering how the  $K$ -scaffold performs node deletion.

Following the configuration model [20], we consider very large random graphs with an arbitrary degree distribution [21]. Thus, given a graph  $\mathcal{G}$  we compute  $S_K(\mathcal{G})$  by increasing  $K$  until it breaks down into many unconnected components. The probability for a node to belong to the  $K$ -scaffold subgraph will be a function of its connectivity  $k$ . We will refer to this function as  $f_k$ :

$$f_k = \begin{cases} 1, & \text{iff } k \geq K \\ 1 - (q_{<K})^k, & \text{otherwise} \end{cases} \quad (11)$$

Clearly, a node with degree  $k < K$  has a probability  $1 - (q_{<K})^k$  to be connected with at least to one node with degree higher than  $K$ .

Now we define the generating functions [6, 21–23], taking in account that to compute the  $K$ -scaffold implies deleting a given fraction nodes [23], namely:

$$F_0(x) = \sum_k P(k) f_k x^k \quad (12)$$

Let us define  $F_1(x)$  as:

$$F_1(x) = \frac{1}{\langle k \rangle_{\mathcal{G}}} \sum_k k P(k) f_k x^{k-1} = \frac{1}{\langle k \rangle_{\mathcal{G}}} \frac{dF_0(x)}{dx} \quad (13)$$

Using previous theoretical results [21–23], we compute the generating functions for the probability distribution of component sizes other than the giant component, if any.  $H_1(x)$  will be defined as the generating function for the probability that one end of a randomly chosen edge on the network  $\mathcal{G}$  -when computing its  $K$ -scaffold- leads to a connected component of a given number of nodes. This includes the probability that such component will contain zero nodes, because of the deletion of nodes of  $\mathcal{G}$  when computing the  $K$ -scaffold. As we discussed above, this will happen with a probability  $1 - f = 1 - F_1(1)$ . The end of the edge can be occupied by a node with  $k$  outgoing edges, distributed along  $F_1(x)$  [23]. Thus, it leads us to the self-consistency equation [21–23] -for a clear and detailed derivation of these results, see [22, 24]:

$$H_1(x) = 1 - F_1(1) + x F_1(H_1(x)) \quad (14)$$

And the generating function for the size of the component to which a random chosen node belongs will be [21–23]:

$$H_0(x) = 1 - F_0(1) + x F_0(H_1(x)) \quad (15)$$

The size of the giant component  $\mathcal{S}_K$  that we will further identify with the  $K$ -scaffold subgraph will be  $\mathcal{S}_K = F_0(1) - F_0(v)$  where  $v$  is the first non trivial, physically relevant solution of  $v = 1 - F_1(1) + F_1(v)$ .

We can now look for a singularity in the average size of components. Immediately above of this point we define the minimal  $S_K(\mathcal{G})$ . Knowing that  $\langle s \rangle = [dH_0(x)/dx]_{x=1}$ , after some algebra [21–23]), we find:

$$\langle s \rangle = F_0(1) + \left. \frac{dF_0(x)}{dx} \right|_{x=1} \times \frac{F_1(1)}{1 - \left. \frac{dF_1(x)}{dx} \right|_{x=1}} \quad (16)$$

With a singularity when  $[dF_1(x)/dx]_{x=1} = 1$ . Now using:

$$\left. \frac{dF_1(x)}{dx} \right|_{x=1} = \frac{1}{\langle k \rangle_{\mathcal{G}}} \sum_k k(k-1) f_k P(k) \quad (17)$$

we can derive the percolation condition for a  $K$ -scaffold graph from a given substrate graph  $\mathcal{G}(V, \Gamma)$  with given average connectivity  $\langle k \rangle_{\mathcal{G}}$  and degree distribution  $P(k)$ . We compute such condition taking in account the above critical condition (17). (Recall that computations are performed considering the successive pruning of  $\mathcal{G}$  by increasing  $K$ ). Knowing that:

$$\sum_k \frac{k P(k)}{\langle k \rangle_{\mathcal{G}}} = 1 \quad (18)$$

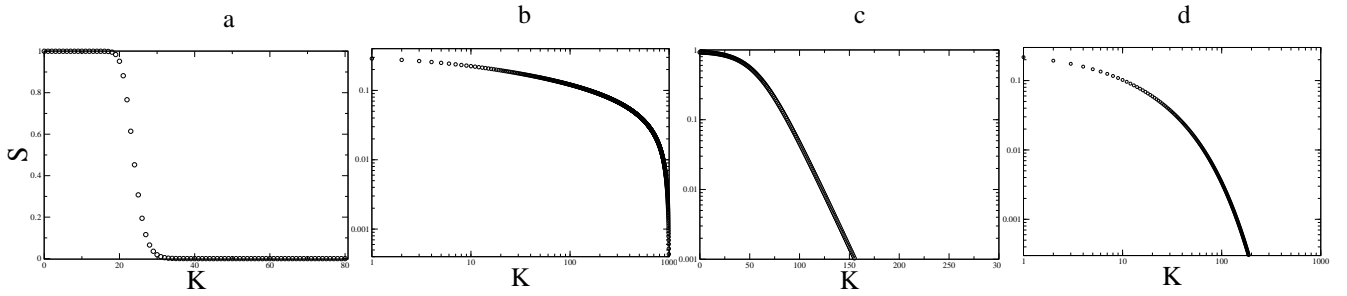


FIG. 2: Numerical simulations of the relative size of the giant component  $S_K$  of the  $K$ -scaffold subgraphs of model degree distributions against  $K$ .  $S_K$  is computed as  $S_K = F(0) - F(v)$ , where  $v$  is the numerical approach of the first non-trivial, physically relevant solution of the equation  $v = 1 - F_1(1) + F_1(v)$  (See text). The plots display  $S_K$  for (a) Erdős-Rényi graph, which  $\langle k \rangle = 15$ . (b) A Scale-free network with  $P(k) \propto k^{-\alpha}$ ,  $\alpha = 2.3$ . No  $K_c$  can be properly identified (see text). The presence of the cut-off can be due to the finite size effects of the simulation; (c) Exponential net with  $P(k) \propto e^{-k/\mathcal{K}}$ ,  $\mathcal{K} = 13$  (d) Power Law with exponential cut off net with  $P(k) \propto k^{-\alpha} e^{-k/\mathcal{K}}$ , for  $\alpha = 2.5$  and  $\mathcal{K} = 52$ .

and since equation (17) is equivalent to:

$$\sum_k k(kf_k - f_k - 1)P(k) = 0 \quad (19)$$

the percolation condition for a  $K$ -scaffold,  $S_K$ , is:

$$\sum_k k(kf_k - f_k - 1)P(k) > 0 \quad (20)$$

We can easily see such a condition as the extension of the Molloy and Reed criteria [25] for the  $K$ -scaffold,  $S_K(\mathcal{G})$ :

$$\sum_k k(k-2)P(k) > \sum_{k < K} k(k-1)q_{<K}^k P(k) \quad (21)$$

Note that the right-hand side of the equation (21) is always finite [26], whereas the left-hand side could not be finite. It is straightforward that:

$$\sum_k k(k-2)P(k) = \langle k^2 \rangle - 2\langle k \rangle \quad (22)$$

Thus, a finite  $K$  for the critical scaffold will exist *if and only if* the degree distribution of the network has a finite second moment  $\langle k^2 \rangle$ . An Erdős-Rényi graph, for example, will display a critical  $K$ -scaffold, provided that  $\langle k^2 \rangle_{ER} = \langle k \rangle^2$ . But for arbitrary large scale-free networks with realistic exponents ( $2 < \alpha < 3$ ), we find that there is not such minimal  $K$ -scaffold. This is due to the divergence of the second moment. Thus, condition (21) always applies for all  $K$ 's. Implications should be studied, provided that it implies that the hub-conector structure appears at all scales.

Numerical simulations (Fig. (2)) show that if we introduce a cut-off in the degree distribution, a characteristic scale  $K_c$  is present. In E-R graphs, the size of the  $K$ -scaffold displays an abrupt decay beyond  $\langle k \rangle$ . Finally,

the size of  $S_K(\mathcal{G})$  displays a critical  $K$  in exponential networks. Note that by its definition, if  $S_K(\mathcal{G})$  percolates, also do the corresponding naked and renormalized counterparts ( $\gamma_K(\mathcal{G})$  and  $\overline{\gamma}_K(\mathcal{G})$ ).

*Discussion.*  $K$ -Scaffold subgraphs can be easily measured on any arbitrary network and can be useful to detect both key elements and the topological components that glue them. If topological organization is linked with functionality, particularly in relation to hubs, the scaffold of a complex network should be able to capture the relevant subsystem. For the human transcription factor network [19] it was found that for  $K = 11$ , a small set of proteins having relevant cellular functions (including oncogenes, tumor suppressor genes and the TATA-binding protein) was obtained, being all of them related through intermediate connector proteins. This is in agreement with the disassortative character of cellular networks. Since each hub was associated to a group of functionally related TFs, the connectors were actually relating different parts of the protein machinery. Other real systems have also been analysed and provided further confirmation of the relevance of the scaffold approach (Corominas Murtra et al, in preparation). Further extensions and properties of this subgraph, together with the analysis of finite size effects associated to real systems will be presented elsewhere.

## Acknowledgments

The authors thank the members of the Complex Systems. This work has been supported by grants FIS2004-0542, IST-FET ECAGENTS, project of the European Community founded under EU R&D contract 01194, by the EU within the 6th Framework Program under contract 001907 (DELIS), by NIH 113004, FIS2004-05422 and by the Santa Fe Institute.

- 
- [1] Albert, R. and Barabasi, A. (2001) *Rev. Mod. Phys.* 74, 47.
- [2] Newman, M. E. J. (2003) *SIAM Review* 45, 167.
- [3] Dorogovtsev, S. N. and Mendes, J. F. F. (2002) *Adv. Phys.* 51, 1079-1187.
- [4] Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M. and Hwang, D.-U. (2006) *Physics Reports* 424, 175.
- [5] Palla, G.; Derenyi, I.; Farkas, I. and Vicsek (2005). *Nature* 435, 814.
- [6] S.N. Dorogovtsev, A.V. Goltsev, and J.F.F. Mendes. (2006) *Phys. Rev. Lett.* 96, 040601 1-4.
- [7] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv and U. Alon (2003) *Phys. Rev. E* 68, 026127.
- [8] Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D. and Alon U. (2002) *Science* 298, 824-827.
- [9] Z. Toroczkai, B. Kozma, K. E. Bassler, N. W. Hengartner, and G. Korniss (2004). *arXiv:cond-mat/0408262*.
- [10] Kim, D. H.; Noh, J. D.; Jeong, H. (2004) *Phys. Rev. Lett.* 96, 018701.
- [11] Valverde, S. and Solé, R.V. (2005) *Phys. Rev. E* 72, 026107.
- [12] J.I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat and A. Vespignani, *cs.NI/0504107*; *cs.NI/0511007*
- [13] Bollobás, B. (1984) *Graph Theory and Combinatorics: Proc. Cambridge Combinatorial Conference in honour of Paul Erdős.* (B. Bollobás, ed) Academic Press, New York.
- [14] Fernholz, D. and Ramachandran, V. (2004) Technical Report TR04-13, University of Texas at Austin. G/A.
- [15] Molloy, M. and Reed, B. (1999). *Electronic J. Comb.* 6 , R35.
- [16] Molloy, M. (1996) *Random Structures and Algorithms* 8, 159-160.
- [17] S. Wuchty and E. Almaas, (2005) *BMC Evol Biol.* 5, 24
- [18] Albert, R., Jeong, H., Barabasi, A. (2000) *Nature* 407, 651.
- [19] Rodriguez-Caso C., Medina M. A., Solé R. V.(2005) *FEBS Journal* 272, 6423.
- [20] A. Bekessy, P. Bekessy, J. Komlos, (1972) *Stud. Sci. Math. Hungar.* 7, 343; E.A. Bender, E.R. Canfield, (1978) *J. Combinatorial Theory A* 24, 296; B. Bollobás, (1980) *Eur. J. Comb.* 1, 311 ; N.C. Wormald, (1981) *J. Combinatorial Theory B* 31, 156,168.
- [21] Newman, M. E. J.; Strogatz, S. H.; Watts, D. J. (2001) *Phys. Rev. E* 64 026118.
- [22] Moore C., Newman M. E. J. (2000) *Phys. Rev. E*, 62, 7059-7064.
- [23] Callaway, D. S., Newman, M. E. J., Strogatz, S. E and Watts, D. J. (2000) *Phys. Rev. Lett.* 85, 5468.
- [24] Newman, M. E. J. (2002) *arXiv:cond-mat/0202208 v1*
- [25] Molloy, M. and Reed, B. (1995) *Random Structures and Algorithms* 6, 161-180.
- [26] Note that  $q_{<K} \leq 1$ , thus,  $(q_{<K})^k \leq 1$ . Futhermore,  $P(k) \leq 1$ . Thus, the sum  $\sum_{k < K} k(k-1)q_{<K}^k P(k)$  is always finite provided that  $k$  is bounded.