PLoS one

Automatic Network Fingerprinting through Single-Node Motifs

Christoph Echtermeyer[1], Luciano da Fontoura Costa[2], Francisco A. Rodrigues[3], Marcus Kaiser[1,4,5]*

1 School of Computing Science, Newcastle University, Newcastle-upon-Tyne, United Kingdom, **2** Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, Brazil, **3** Departamento de Matemática Aplicada e Estatística, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brazil, **4** Institute of Neuroscience, The Medical School, Newcastle University, Newcastle-upon-Tyne, United Kingdom, **5** Department of Brain and Cognitive Sciences, Seoul National University, Seoul, Republic of Korea

Abstract

Complex networks have been characterised by their specific connectivity patterns (network motifs), but their building blocks can also be identified and described by node-motifs—a combination of local network features. One technique to identify single node-motifs has been presented by Costa et al. (L. D. F. Costa, F. A. Rodrigues, C. C. Hilgetag, and M. Kaiser, Europhys. Lett., **87**, 1, 2009). Here, we first suggest improvements to the method including how its parameters can be determined automatically. Such automatic routines make high-throughput studies of many networks feasible. Second, the new routines are validated in different network-series. Third, we provide an example of how the method can be used to analyse network time-series. In conclusion, we provide a robust method for systematically discovering and classifying characteristic nodes of a network. In contrast to classical motif analysis, our approach can identify individual components (here: nodes) that are specific to a network. Such special nodes, as hubs before, might be found to play critical roles in real-world networks.

Citation: Echtermeyer C, da Fontoura Costa L, Rodrigues FA, Kaiser M (2011) Automatic Network Fingerprinting through Single-Node Motifs. PLoS ONE 6(1): e15765. doi:10.1371/journal.pone.0015765

Editor: Matjaz Perc, University of Maribor, Slovenia

Received October 15, 2010; **Accepted** December 2, 2010; **Published** January 31, 2011

Funding: Marcus Kaiser and Christoph Echtermeyer were supported by the Engineering and Physical Sciences Research Council (EPSRC) (EP/G03950X/1), the CARMEN e-science project (http://www.carmen.org.uk) funded by EPSRC (EP/E002331/1), and the World Class University program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R32-10142). Luciano da F. Costa is grateful to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (301303/06-1) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (05/00587-5) for financial support. Francisco A. Rodrigues is grateful to FAPESP (2007/50633-9) for sponsorship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: m.kaiser@newcastle.ac.uk

Introduction

Networks appear in a variety of real-world systems ranging from biology to engineering [1,2]. Examples include neural [3–5], social [6–8], and computer networks [9,10] to name but a few. Networks have been used to study the emergence of cooperative behaviour [11–13]; to address epidemiological questions [14,15] especially in scale free networks [16,17]; and to investigate the causes of cascade effects [18,19] for a more complete understanding of why networks differ in robustness against error and attack [20,21]. Attempts to classify network-topologies [22] were accompanied by detailed studies of scale-free [23] and small-world networks [24,25]—properties that were identified in many real networks. Additional to investigations of concrete structures, theoretical studies of random networks collected valuable information about large classes of networks [26–28].

Mapping complex systems to networks revealed that some nodes are remarkably different from other nodes of the same network. For instance, hubs, characterized by a high number of connections (a high node degree), often play a fundamental role in protein-protein interaction networks and their removal can be lethal for an organism [29,30]. Hubs are similarly important for socio-economic systems, where defective hubs can cause cooperation to decline [31]. Also, in engineered systems like the Internet, hubs are important to maintain the communication between autonomous systems [20]. These outlier nodes have been identified since the introduction of complex network theory, e.g. in the World Wide Web [9] and the Internet [10], but hubs are outliers only in terms of their degree; other network properties can also define special nodes. For instance, Internet topology has been shown decompose onion-like into different shells around a relatively small core network [32]. The closer a node's layer is to the core, the higher is the node's shell-index (coreness) [33]. Nodes with high coreness are not necessarily hubs, which one might suspect to be the most efficient spreaders of information. Instead, the position of a node close to the network-core has more impact on successful dissemination than having a high degree [34]. In networks where hubs are not present, as in most geographical networks, nodes whose neighbours are also connected to each other are special (high local clustering coefficient). Further examples of outlier nodes can be found with different measures some of which examine more than the direct neighbourhood of a node [26,35], such that they specify rather global (network specific) than local (node specific) characteristics. Global measures, such as characteristic path length or clustering coefficient [24], summarise the whole network in a single value. Local measurements, on the other hand, analyse each node or edge individually, yielding a more fine-grained picture of the network. Nodes that express common features and outliers that are different can be identified with pattern recognition approaches, which group nodes of similar characteristics. Corresponding techniques have been proposed recently [36–38] and revealed important network properties. For example, in protein-protein interaction networks the relative number of outliers tends to decrease with the complexity of organism, i.e. proteins in more complex species show higher homogeneity in their interplay [38]. This demonstrates that, by considering multiple node-features jointly, pattern recognition based methods can point out exceptional network components.

Networks can describe complex systems whose interactivity between dynamical components changes over time. Altered connections between the elements (represented by nodes) may in turn feed back on the dynamics, such that the dynamical process and the network topology evolve in an adaptive fashion [39]. In the context of game theory, corresponding coevolution of behaviour and connectivity has been studied in socio-economic systems [12,13]. In complex scenarios like these, analysing a single network is often insufficient and several networks must be compared to gain insights. Further examples for the need of network-comparisons are families of protein-protein interaction networks, brain connectivity networks in patient- and control-populations, or time-dependent (developing or declining) networks [40]. Comparing such sets of networks requires consistent approaches, which are often non-trivial, because networks differ in size (number of nodes or edges) or they comprise a disjoint sets of nodes (some nodes occur in one network but not in others). Direct comparisons between structures may thus be ruled out. Based on outlier-detection as described above, we previously proposed motif-regions for which the relative frequencies of outliers falling into one of them yields a network specific fingerprint [41]. Relating different networks to each other has thereby become as easy as comparing bar-graphs. Nevertheless, although this methodology has been demonstrated to be suitable and accurate for outlier identification as well as for network comparisons, it suffers from several limitations, which we address in this paper.

Here, we describe a novel workflow for detecting characteristic single-node motifs and for using fingerprints for network comparison. Improvements compared to the previous approach include (a) automatic parameter determination, which facilitates high throughput analysis without user interaction, and (b) replacing the k-means clustering algorithm with a deterministic method to simplify the workflow and to improve robustness of results. In addition, we provide (c) a validation of our method and (d) an application to networks where the topology changes over time (addition or deletion of nodes or edges).

Previous work

The application of single measures to complex networks has revealed important insights in many cases. However, as Newman and Leicht recognised [36], detecting exceptions is limited to network features that are quantified by the measures in use. Otherwise, if the chosen characteristics do not reflect the properties that are specific for a network or its components, important features remain unnoticed.

To solve this problem, two complementary approaches have been suggested. The first approach by Newman and Leicht groups nodes based on their connectivity without any further prior information [36]. By fitting the parameters of a mixture model (using an expectation-maximization algorithm), each node is assigned a probability of belonging to any one group that has been identified. The probabilistic nature of this approach has the advantage that nodes that can not be unambiguously categorised are not crudely assigned to one particular group, but the conflict becomes evident, such that it can be dealt with. The structure of networks can thereby be investigated without requiring any other parameter than the number of groups that are to be created. This elegant method has been examined thoroughly and improvements to it have been suggested [37,42].

Analyses with focus on only one particular aspect of a network at a time might fail to detect irregularities or similarities in structure. The second approach is to avoid single measures and to use a combination of multiple ones [41]. Instead of reducing network components down to one dimension, joint measures map it into a multi-dimensional feature space [43]; each vector-point in that space corresponds to a combination of node-characteristics and statistical methods are used to identify motif regions, such that each vertex falls into one of them: A node is either classified regular—showing features like the majority of nodes—or singular, i.e. its features deviate by following a particular single node-motif. The term motif refers to the concept of network motifs, i.e. patterns incorporating multiple nodes [44].

Each of these two approaches to identify patterns in complex networks has its drawbacks and advantages. The Newman and Leicht algorithm (NLA) does not depend on one or few network measures, but it works on network links directly. Networks are not restricted to undirected ones, but directed links and even weighted ones can be considered. The NLA requires the number of node-groups to be specified; this is also true for the approach by Costa et al. [Beyond the Average (BtA)], where the number of motif regions needs to be chosen *a priori* [41]. Unfortunately, for real-world networks this number is often unknown. The BtA-workflow requires two additional parameters to control which nodes will constitute individual motif regions. Both methods differ in their output, as BtA not only provides a grouping of nodes, but also a network-fingerprint, which can be used to compare networks from different domains. Most importantly, however, is the conceptional distinction between NLA and BtA, as they rely on local edge connectivity and local node measures, respectively. BtA will fail to pinpoint features of the network, if the chosen set of measures can not formulate a corresponding motif. Similarly NLA can fail, as it only takes into account direct connections between nodes: NLA does not consider how the neighbours of a node are connected, for example, but BtA can deal with such information (by evaluating the local clustering coefficient). Indeed, the extensibility concerning features to assess is the biggest advantage of BtA; (un-)directed and weighted links can be processed likewise and in spatial networks the location of nodes can be taken into account. In conclusion, NLA is readily applicable to a broad variety of network domains; however, considering direct connections only is a weakness. BtA can be nicely adopted to these cases, but care has to be taken at all times to ensure the set of measures is diverse enough to cover as many patterns that might occur in networks as possible.

In the next section we suggest several improvements to the BtA-workflow (Fig. 1), which can be sketched as follows: Initially, multiple local network measures are applied to each node, which yields a multi-dimensional characterisation in form of a *feature vector*. Correlation between different measures is accounted for by principal component analysis (PCA), which is used to map feature vectors of all nodes to two dimensional space [45, Chapter 8]. Next, nodes are assigned probabilities in order to distinguish nodes with common and rare features. The required probability density function (PDF) is gained by smoothing over points in the two dimensional PCA-plane (Parzen window approach [46, 47, Chapter 4.3]). Now, the least probable nodes, i.e those with uncommon features, can be identified from the PDF. These

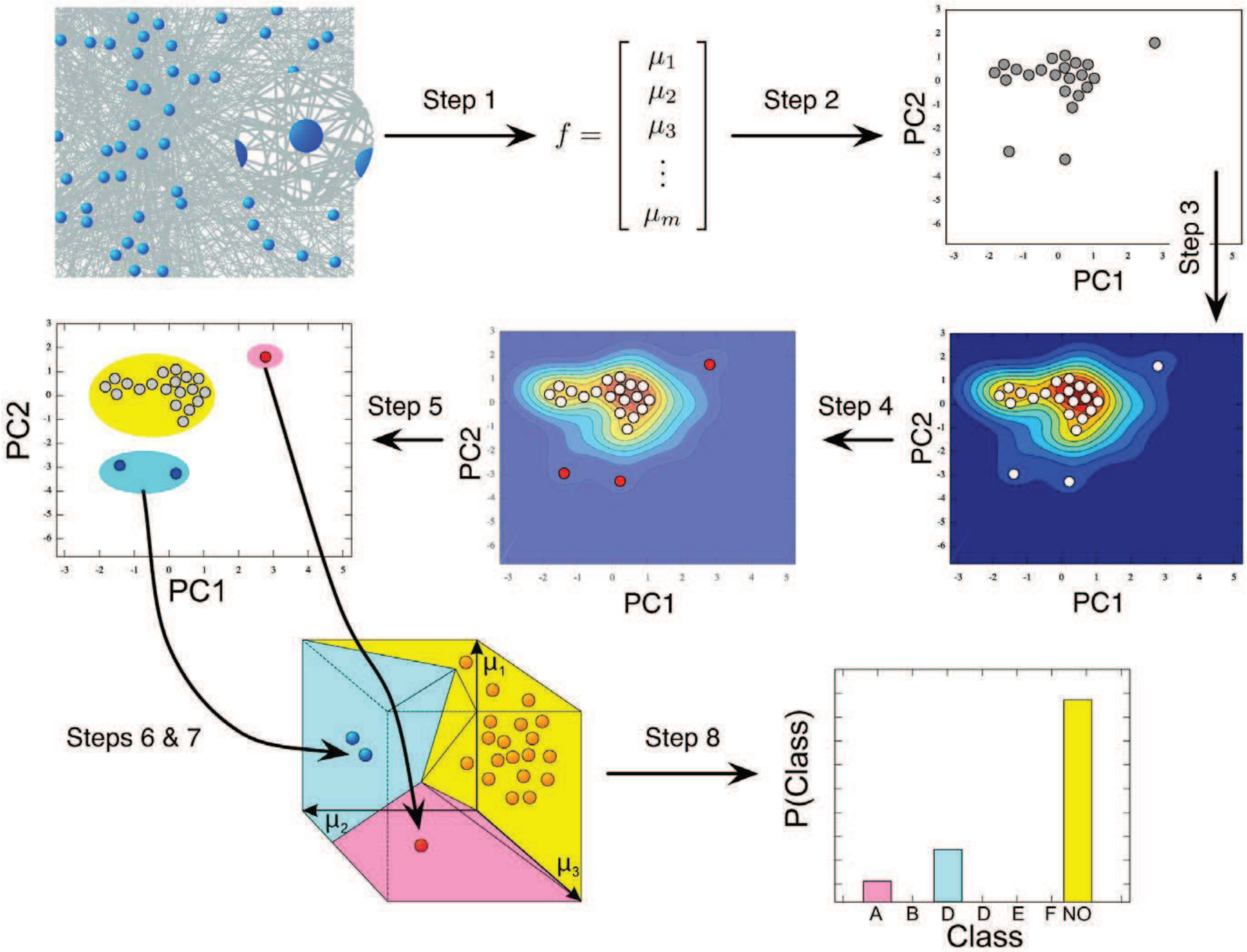

Figure 1. Analysis work-flow to identify global singular nodes from local features [41]. Step 1: Choose set of local measures to characterise network nodes [35]. Calculate local measurements for all nodes in the network (*feature vectors*). Step 2: Map each node's feature vector to lower dimensional space using principal component analysis (PCA plane) [45, Chapter 8]. Step 3: Estimate each node's probability using the Parzen window approach (PDF) [46, 47, Chapter 4.3]. Step 4: Query PDF to identify least probable nodes (*singular nodes*). Step 5: Cluster singular nodes in PCA plane using k-means (*motif groups*) [65, Chapter 20.1]. Step 6: Determine Voronoi cells for grouped nodes using a modified Mahalanobis distance (*potential motif regions*) [69]. Step 7: Join potential motif regions that are close to each other (*motif regions*). Step 8: Calculate relative frequencies of nodes falling into motif-regions (A–F) or non-motif region (NO) (*fingerprint*).
doi:10.1371/journal.pone.0015765.g001

singular nodes are then clustered in order to distinguish different *motif-groups*. Each of the two dimensional motif-groups corresponds to a higher dimensional *motif-region* into which the feature vectors split up and the distribution of feature vectors among the different motif-regions is the *fingerprint* of the network. Apart from the initial decision on which measures to use, the user needs to choose the bandwidth of the smoothing kernel, the number of singular nodes w, and the number of motif-groups k, respectively (steps 3, 4, and 5 in Fig. 1). Additionally, when comparing multiple networks, a limit must be specified below which motif regions are considered too close to each other to constitute different motifs (join threshold; Step 7). So far, these settings had to be chosen manually, but here we suggest how to determine all three parameters (bandwidth, w, and k) automatically. The last setting (join threshold), however, is not considered for automation: So far we could not identify a procedure that yields results as good as manual selection by the user. We thus concentrated our efforts on the parameters that need to be set for every network (bandwidth, w, and k), such that high-throughput applications become possible. Automating the setting of the main parameters is thus of higher benefit than for the threshold that determines Voronoi cells to be joined; this needs to be chosen only once, when all networks are compared to each other at the end.

Results

In this paper we propose how to choose all relevant parameters of the BtA-workflow automatically (see Methods section), which allows for the analysis of many networks without the need for human interaction. The only remaining limiting factor for high throughput analyses are the computational costs of the analysis, which predominantly depend on the measures that are chosen to characterise each node. Using implementations of common local measures (see File S1), the estimated run-time scales linearly to cubic with network size (Fig. S1). Costs are thus comparatively cheap considering methods that identify specific connectivity patterns by counting occurrences of particular sub-graphs (e.g. [44,48–52]); such motif-counts also scale at least linearly in network size, but they show exponentially growing costs as the size of the motif-pattern increases [50]. In practice this often means that counts can not be determined for patterns involving 10 nodes or more [53], which renders some domains computationally intractable for this approach, but eventually not for BtA. However, before processing huge networks or many different structures with BtA, we first need to verify that parameters are indeed chosen adequately, which is confirmed in the next section.

Method Verification

The first validation is on a network that is small enough to confirm BtA-results by eye: We use a family-tree from *The Simpsons* [54] to create a network with nodes representing characters and directed links pointing to their offspring (Fig. 2a). Nodes that have a sparsely connected and homogeneous neighbourhood are suitably highlighted as outliers by BtA.

With these reassuring results from a single network, we proceed by testing BtA on whole series: We generate structures with both regular components and exceptional ones, which BtA has to identify. In our first series we compose networks of two components: a regular ring lattice and a smaller Erdös-Rényi (ER) [55] random network (Fig. 2b). While the ring lattice remains unchanged, the size of the random module increases throughout the series, such that its proportion of the full network grows gradually. The ring lattice is comprised of 100 nodes, each of which is connected to its four closest neighbours (Fig. 2b). The generated ER-random networks ($n = 1, \ldots, 50$ nodes) have an average edge-density of 25%. Composed networks are analysed with BtA: Of all outlier-nodes less than 2% are missed while over 96% are classified correctly, if the random component contributes less than 25% of nodes to the network. Beyond that limit, the number of nodes in the random-part does no longer match the number of identified outliers w. But this does not imply a misclassification by BtA: The larger a random network, the more likely it is that a few nodes are connected regularly (or close to that). Quantifying these nodes with local network measures yields the same values as (or similar to) those of the ring lattice, which is why it would be incorrect to consider them singular. Additional to regular connection patterns in large random networks, other local motifs can also be frequent enough, such that they constitute a common rather than an exceptional feature of the network. Thus, network components that seem clearly separable at first may actually be very similar or—although intended to form outliers—they may contain common elements, due to random effects. Together this explains the observed deviations in numbers of outliers for growing ER-components in this test-series.

Finally, we reverse the nature of the networks: The major component is set to a random network [ER, Barabási and Albert (BA) [56], or Watts-Strogatz (WS) [24] model] in which we embed a small, but highly regular structure (Fig. 2c). The inserted structure was chosen, such that its nodes are highly clustered (both on level 1 and 2); the six outer nodes further show significant variability in their neighbours' degrees. These characteristics are rarely observed in our random networks, which is why BtA should identify these nodes (alongside with other outliers that might emerge). We confirm this in a series of ER, BA, and WS networks ($n = 100$ nodes) with varying sparseness (edge density ranging from 1% to 50% with step-size 1%). The regular structure (7 nodes) illustrated in Fig. 2c is added to each random network before BtA is applied. For these networks, the 6 outer nodes of the regular structure are classified singular in over 97% of all networks. Additionally, the inner node (with less extreme features) is regarded uncommon in 81% of all cases.

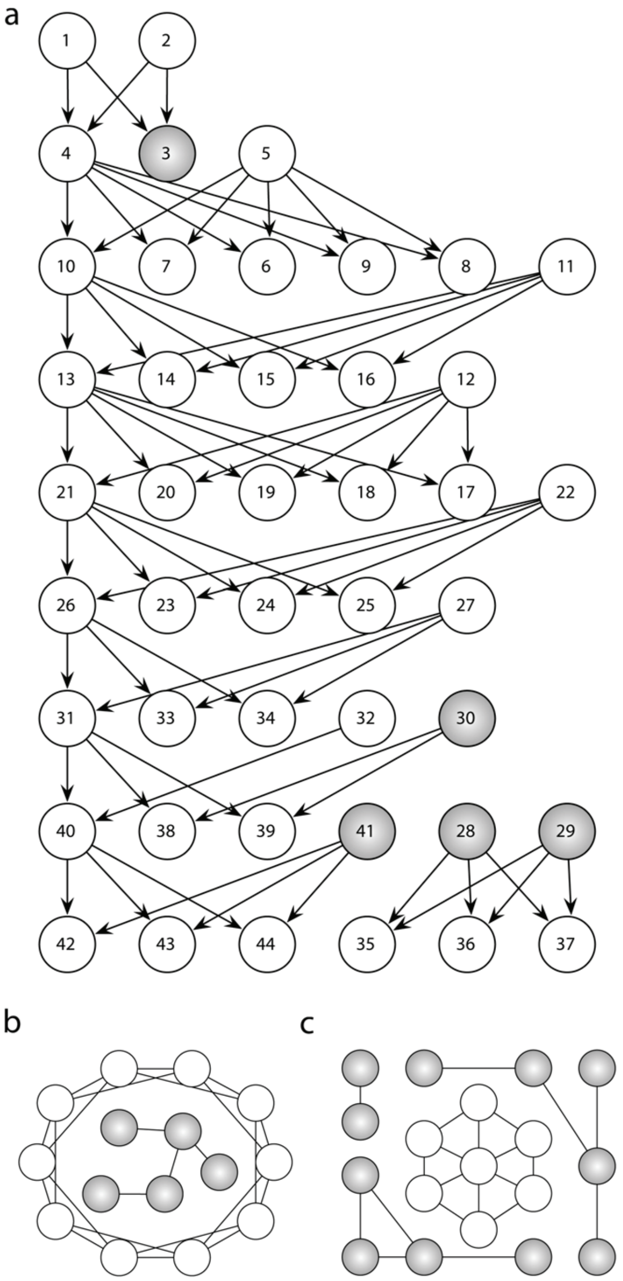

Figure 2. Network types used for testing BtA: a Network derived from *The Simpsons* family-tree [54]. Nodes in very regular parts of the network were identified singular (shaded grey) because of two characteristics: Their neighbours' degrees are comparatively low and show no variation (values r and cv significantly below average). **b** Schematic of large regular ring lattice combined with a minor ER-random component (shaded grey). **c** A small regular structure (white nodes) embedded into a large random network (ER, BA, or WS model).
doi:10.1371/journal.pone.0015765.g002

In conclusion, the automatic parameter determination gives very satisfying results, which yield confidence in BtA's ability to identify outliers in complex networks autonomously.

Network Time-Series: A Small-World Emerging

Large complex networks are challenging to analyse; time-series of such are even more so. We attempt to approach this challenge by first condensing networks to a compact representation—mapping a series of changing structures to a uniform representation benefits the identification of trends and changes of such. Therefore, all networks have to be characterised, which we do using single node-motifs. These are identified with BtA using six common local measures: (1) the normalised average degree *r*, (2) the coefficient of variation of the degrees of the immediate neighbours of a node *cv*, (3) the clustering coefficient *cc* [24,57], (4) the locality index *loc*, (5) the hierarchical clustering coefficient of level two cc_2 [58], and (6) the normalised node degree *K*. (For definitions of these measures see Methods section.) Next, we describe the time-series of 600 networks and the results found with BtA.

Similar to random graphs, small-world networks have a small characteristic path length, but at the same time they exhibit a high degree of clustering, as regular ring lattices, for example. It has been discovered early that the combination of short paths plus grouping is inherent to social networks; a phenomenon that became known as six degrees of separation [59–61]. Today it is known that small-world networks can be found in many other domains (e.g. [2,26–28]). We thus created a network-time series in which structures gradually change from a completely regular ring lattice to a small-world network (see Methods section, Fig. S2).

In total we identified 5 single node-motifs, which differ in characteristics, frequency, and time of emergence (Fig. 3): A node according to motif 1 has relatively few connections in contrast to its well connected neighbourhood. Different from that, nodes corresponding to motif 2 are signified by many connections to a rather sparsely connected neighbourhood. Motif 3-nodes have relatively few connections and nodes in their neighbourhood are similar in number of links and corresponding targets. Motif 4 describes rarely connected nodes whose neighbours have a diverse number of connections; but instead of being linked between each other, neighbours share other common targets. The final motif 5 can be best characterised by its relation to the rest of the network, which shows a higher degree of connectivity than any node involved in the motif. Neighbours of the motif-node further vary in their number of connections and do not link to each other. Motifs 2, 3 and 5 appear right from the beginning of the rewiring process; motifs 2 and 5 gradually become more common over time, whereas 3 levels out after a transient peak. The remaining motifs 1 and especially 4 only become apparent at later stages towards which both become more frequent. Together, BtA reveals the increasing irregularity in network structure and it also provides details on the characteristic connectivity patterns at different times. Both would be valuable information if real networks were analysed; here, with precise knowledge about the network-changing process, the temporally dependent motif expression levels yield another validation of the technique (detailed discussion in File S1).

Overall, results are very satisfying and we are confident that BtA could be successfully applied to real networks using the automatic parameter determination.

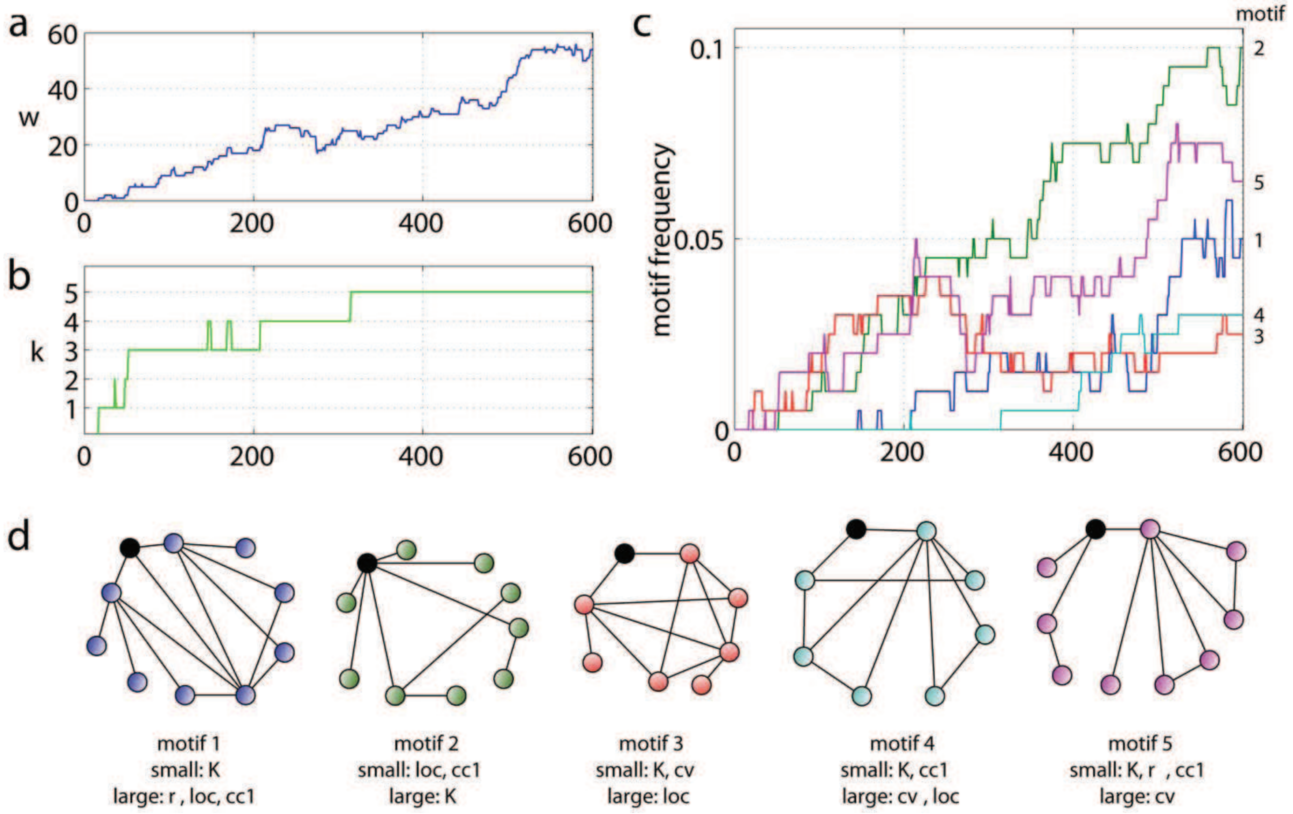

Figure 3. Single node-motifs in emerging small-world network (Fig. S2). Vertical axes in subfigures a–c correspond to number of outlier nodes *w*, number of single node-motifs *k*, and their frequencies, respectively. **a** Number of identified outliers *w* rising from 0 to 54. **b** Diversity of node-motifs *k* quickly rising during the 1st re-wiring round; less increase during 2nd round; and stable during the 3rd. **c** Proportions of nodes expressing identified motifs (*motif frequencies*). Nodes classified regular not shown. **d** Schematics of identified single node-motifs and their distinguishing characteristics.
doi:10.1371/journal.pone.0015765.g003

Discussion

In this paper we presented a method to detect single node-motifs automatically. The main parameters of the previous routine [41] —the smoothing kernel bandwidth plus the number of singular nodes and motif groups—are now selected based on the data. We further proposed a deterministic replacement for the k-means algorithm, which is used to form the different motif-groups. In contrast to k-means, our alternative approach can determine the number of motifs itself and due to the lack of random elements, clustering results are robust over multiple repetitions.

Despite our improvements to BtA certain issues and room for further advancements remain. For example, reducing feature vectors in dimension inevitably leads to a loss of information, but which has to be kept withing reasonable bounds. In other words, although 6-dimensional feature vectors were suitably represented in the 2-dimensional plane so far [41], different networks may require the use of more than just the first 2 principal components in order to ensure that network characteristics are represented properly. Thus, if the chosen number of principal components does not account for at least 80% of the variance, their number should be increased (*Kaiser's rule*). The degree to which feature vectors can be reduced thereby depends on the correlation between measured values, which is specific to the analysed network.

In cases where feature vectors can not be suitably represented in 2 dimensions, their display becomes more complicated and verifying a good fit of the estimated probability density function (PDF) is challenging. However, a good PDF estimate is needed in the BtA workflow to determine outlier nodes. Problems that might arise in these situations could possibly be circumvented by a major change to the workflow: The use of PCA to compact information offers the possibility to replace both the PDF estimation and the subsequent outlier selection with a more direct and non-parametric standard technique, which is *Hotelling's* T^2 (a generalisation of *Student's t-statistic*). This modification would allow to identify outliers without the need to estimate a PDF, but the exploration of the resulting workflow will be addressed in another publication.

Considering the BtA workflow as presented in this paper, the technique can be easily adapted by including different local network measures in the analysis. Measures that take spacial aspects of the network into account, for instance, or those including link-weights can increase quality of the analysis. Finally, interest might not only lie on motifs formed by outlier nodes, but on all single node-motifs occurring in the network. In this case regular and singular nodes are not distinguished, but all of them have to be included in the network fingerprint.

BtA-fingerprinting of many networks has so far been prevented by the need to choose parameters during the analysis manually. With the improvements presented in this paper, however, it is now possible to process large numbers of networks fully unsupervised. Identified outliers are characteristic nodes that can provide a fingerprint of a network; fingerprinting networks from numerous domains allows easy characterisation and comparisons. As already demonstrated [41], such studies can reveal important characteristics and differences between network domains. Additionally, the example on an emerging small-world network in this paper showed that BtA can also be used to analyse time-series of networks.

To encourage the use of the BtA methodology by other researchers, we provide our implementation of the workflow including the automatic parameter determination for download (http://www.biological-networks.org/). Two versions of the code exist: The first one requires Matlab (Mathworks Inc, Natick, USA) and allows the user to apply the workflow using a graphical user interface (Fig. S3). The other one is a command line utility that either requires Matlab or the free alternative Octave [62] and it can be easily used to batch process many networks.

In conclusion, we provide a robust method for systematically discovering and classifying characteristic nodes of a network. The distribution of node-classes results in a fingerprint, which in turn can give a classification of whole networks, as for network motifs of multiple nodes [63]. In contrast to classical motif analysis, our approach can identify the individual components that are specific to a network. Such special nodes, as hubs before, might be found to play critical roles in real-world networks.

Methods

Local Network Measures

Network nodes were characterised with six common local measures whose definitions are given in the following. Therefore, let $A = (a_{ij})$ denote the adjacency matrix of the network, i.e. $a_{ij} = 1$, if a link from node i to node j exists, and otherwise $a_{ij} = 0$. Row- and column-sums of A correspond to the *in-* and *out-degrees* of nodes, respectively. In undirected networks, in- and out-degree are equal and either of them can be used as a node's *degree*. If links are directed, the degree is the sum of in- and out-degree. Dividing a node's degree by the number of all links in the network yields the *normalised node degree* K. The *normalised average degree* r_i of a node i is the average over all its neighbours' degrees. (Nodes that are directly linked to node i are called *neighbours*.) Likewise, the *coefficient of variation* cv of the degrees of the immediate neighbours of a node can be calculated. The neighbours' connectivity with each other is quantified by the *clustering coefficient* cc_i, which is the proportion of existing connections between node i's neighbours to the number of all possible links between them [24,57]. The clustering coefficient thus reflects the relative number of triangle-shaped paths a node has—a concept that is extended to connections between neighbours' neighbours (further away node node i) by the *hierarchical clustering coefficient* of level two cc_2 [58]. Whereas the cluster coefficients quantify connectivity within a node's neighbourhood, the *locality index* loc_i, which is based on the matching index (e.g. [64]), is the fraction of neighbours' links that connect to the same node (not necessarily a neighbour of node i). Further details and measures can be found in the literature [26–28,35].

In the following sections we describe how appropriate settings for the parameters of the BtA-workflow can be found automatically. Kernel-bandwidth, the number of singular nodes w, and the number of motif regions k are discussed separately below.

Kernel-Bandwidth

In step 3 of the workflow (Fig. 1), the Parzen window approach is used to estimate a probability density function (PDF) over all nodes [46, 47, Chapter 4.3]. This is achieved by smoothing the overall arrangement of reduced feature vectors, which were obtained using principal component analysis (PCA) [45, Chapter 8] in the previous step 2. The dimensions of the smoothing kernel, i.e. the width and breadth of the Gaussian function $\mathcal{N}_2(\mu,\Sigma)$ can be controlled through its covariance matrix $\Sigma = (\sigma_{ij})$. (Mean vectors μ are fixed to equal the data-points.) The original publication made use of the fact that the absolute covariance values ($\forall k : \sigma_{kk} = 0.05$) do not matter for the estimated PDF. However, their values relative to each other do matter and we therefore scale them according to the standard deviation along each principal component (PC) axis. Variability-based re-shaping of the kernel function improves the overall fit of the PDF to the points (Fig. S4). A further refinement would be to tilt the Gaussian in order to account for correlation between axes (Fig. S5); however, the PCs are expected to show weak correlation only, which is why we chose un-tilted kernels (for which the covariance matrix Σ is zero except for the variances on the diagonal).

Number of Singular Nodes w

After assigning probabilities to all nodes (Step 3), nodes with an exceptionally low probability come into focus: These outliers correspond to points in the PCA-plane that are spatially separated from larger clusters; and this separation corresponds to abnormalities of measured features. Due to their uncommon characteristics, these nodes are considered singular. For humans it is usually straightforward to identify these non-regular nodes, if interactive visual aids are provided; we therefore implemented a graphical user interface for the whole workflow (Fig. S3). In the following, however, we discuss how the number of singular nodes w can be adjusted without interaction.

To determine singular nodes, automated methods can query the PDF that has been estimated earlier (Step 3). For example, for a fixed number w of singular nodes, the w least probable ones can be selected easily. Alternatively, a probability cut-off can be set, e.g. at 1% or 5%, to separate nodes into regular and singular ones. Both these simple methods involve constants, but which have to be chosen depending on network size to yield sensible results. Choosing one fixed number of singular nodes w for differently sized networks can render the majority of nodes non-regular in comparatively small networks; vice versa, w may be too small compared to the number of exceptional nodes in large networks. A fixed probability cut-off does not circumvent this problem, because the nodes' absolute probability values are dependent on network size. In the following, we therefore propose a flexible probability-threshold: The cut-off does not occur at a fixed pre-defined level, but where it yields the best separation between singular and regular nodes.

A necessary condition for a node being considered singular is a sufficiently low probability compared to other nodes. Additionally, it is desirable that singular nodes appear somewhat separated from the regular ones, which renders their classification non-arbitrary. We therefore suggest to set the borderline between regular and singular nodes where the steepest increase in probability among the low probability nodes appears. Nodes with a probability below mean $\bar{p}$ minus one standard deviation $\sigma(p)$ of all nodes' probabilities $p=(p_k)_{k=1,\dots,n}$ are potentially singular. Given that the probabilities $p=(p_k)_{k=1,\dots,n}$ are sorted increasingly, the number of singular nodes w is then chosen as

$$w=\arg\max_{k\,:\,p_k<\bar{p}-\sigma(p)} p_{k+1}-p_k, \qquad (1)$$

or $w=0$, if probabilities undershoot the mean only minimally (i.e. $\nexists\, k : p_k<\bar{p}-\sigma(p)$).

Number of Motif Groups k

Once nodes are classified as either regular or singular (Step 4), clusters of singular nodes (*motif-groups*) are identified using k-means [65, Chapter 20.1]. The k-means clustering algorithm requires the number of clusters k to be chosen *a priori*; the actual procedure then determines k centroids and assigns each node to the closest one of them. Choosing k too low results in clustering errors, because multiple motif-groups are falsely considered as one. Conversely, too many clusters split motif-groups into non-existing sub-groups. Determining a suitable k is thus crucial for automating the workflow and we come back to this issue later. Even if k is chosen adequately, clustering results are not guaranteed to be satisfactory when using k-means: The algorithm initially chooses the cluster-centroids at random, but their actual distribution impacts on the quality of clustering results [66]. Attempts to optimise the centroid initialisation have been made (e.g. the k++$-$algorithm [67]), but random effects still remain; we therefore suggest a deterministic replacement for k-means.

Optimal groupings of singular nodes consider well separated nodes to be in different clusters, whereas relatively close ones are grouped together. The standard deviations along each PC-axis can serve as a threshold for *closeness* and we consider each of the singular nodes to occupy a certain volume in the PCA-plane, i.e. an ellipse-shaped area centred on it. All ellipses have the same dimensions, which equal the standard deviations along the two axes. Nodes are then assigned to the same motif-group if all their ellipses constitute a connected area (Fig. 4). Practically, this idea can be implemented in 3 steps:

1. Similar to an adjacency matrix, create a binary *overlap-matrix* $O=(o_{ij})$ in which nodes are connected if their ellipses overlap; otherwise they are not. For two nodes i and j let $x=(x_1,\,x_2)$ and $y=(y_1,\,y_2)$ denote their corresponding points on the PCA-plane, i.e. the centres of their ellipses with dimensions σ_1 and σ_2. Using the rescaled centres $c_x=(x_1/\sigma_1,\,x_2/\sigma_2)$ and $c_y=(y_1/\sigma_1,\,y_2/\sigma_2)$ the entry of the overlap-matrix is defined by

$$o_{ij}=\begin{cases}1, & d_2\left(c_x,c_y\right)<1,\\ 0, & \text{otherwise},\end{cases} \qquad (2)$$

where $d_2(\cdot,\cdot)$ is the Euclidean distance.

2. Determine a corresponding *clique-matrix* $C=(c_{ij})$ that specifies whether a path—a connected area of ellipses—between any two nodes exists or not. Paths or cliques can be determined through powers $O^k=\left(o_{ij}^{(k)}\right)$ of the overlap-matrix O via

$$c_{ij}=\begin{cases}1, & \exists\, k\in\mathbb{N} : o_{ij}^{(k)}>0,\\ 0, & \text{otherwise}.\end{cases} \qquad (3)$$

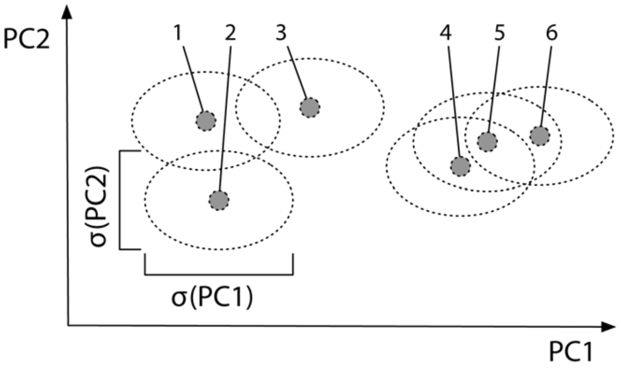

Figure 4. Example of 2 clusters (left, right) with 3 points each (1–3, 4–6). Ellipses are centred on each point with dimensions corresponding to standard deviations σ along PC-axes. A set of points is considered a *clique*, if the area of all their ellipses is connected (e.g. {1}, {1, 2}, or {1, 2, 3}; but not {2,3}). A maximal clique is called a *cluster* (i.e. {1,2,3} or {4,5,6}) and is used to define a distinct motif-group.
doi:10.1371/journal.pone.0015765.g004

3. Colour all cliques differently, which finally yields the motif-groups.

Note that this procedure has no parameter controlling the number of motif-groups, but these are identified automatically. Instead of using this method to actually group nodes it might also serve as a pre-processing step in order to determine the number of clusters k for k-means. The drawback of this simple approach is that long elongated clusters can result when nodes are widely distributed, but connected by a chain of nodes that are just less than one standard deviation apart from each other. However, we have not observed such formation in practical applications.

Generation of Small-World Networks

The prevalence of small-world networks has risen questions about their generating mechanisms and different explanatory models have been proposed [24,68]. We use one of them here in order to generate a series of networks: Watts and Strogatz described a rewiring procedure by which a regular ring-lattice is randomly rewired by which it becomes a small-world network [24]. This is a step-wise process, which allows to sample a network at each intermediate stage. Starting with a completely regular structure, over time, networks become increasingly perturbed (Fig. S2). In total, we sampled 600 networks (à 200 nodes), which were then analysed with BtA, to determine the single node-motifs that evolve over time.

Supporting Information

File S1 Supplementary figures, notes on software implementation, notes on run-time complexity, and detailed discussion of the small-world network results.
(PDF)

Author Contributions

Conceived and designed the experiments: CE LdFC FAR MK. Performed the experiments: CE. Analyzed the data: CE. Wrote the paper: CE MK.

References

1. Bornholdt S, Schuster HG, eds (2003) Handbook of graphs and networks: from the Genome to the Internet. John Wiley and Sons, 1st edition.
2. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D (2006) Complex networks: Structure and dynamics. Physics Reports 424: 175–308.
3. Sporns O, Chialvo DR, Kaiser M, Hilgetag CC (2004) Organization, development and function of complex brain networks. Trends in Cognitive Sciences 8: 418–25.
4. Kaiser M (2007) Brain architecture: a design for natural computation. Philosophical Transactions of the Royal Society A 365: 3033–45.
5. Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience 10: 186–98.
6. Lazer D, Pentland A, Adamic L, Aral S, Barabási AL, et al. (2009) Social science. Computational social science. Science 323: 721–3.
7. Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323: 892–5.
8. Centola D (2010) The Spread of Behavior in an Online Social Network Experiment. Science 329: 1194–7.
9. Albert R, Jeong H, Barabási AL (1999) Diameter of the World-Wide Web. Nature 401: 130–1.
10. Faloutsos M, Faloutsos P, Faloutsos C (1999) On Power-Law Relationships of the Internet Topology. In: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication. ACM. pp 251–62.
11. Nowak MA (2006) Five rules for the evolution of cooperation. Science (New York, NY) 314: 1560–3.
12. Szabo G, Fath G (2007) Evolutionary games on graphs. Physics Reports 446: 97–216.
13. Perc M, Szolnoki A (2010) Coevolutionary games–a mini review. BioSystems 99: 109–25.
14. Barthélemy M, Barrat A, Pastor-Satorras R, Vespignani A (2005) Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. Journal of theoretical biology 235: 275–88.
15. Funk S, Salathe M, Jansen VAA (2010) Modelling the influence of human behaviour on the spread of infectious diseases: a review. Journal of The Royal Society Interface.
16. Pastor-Satorras R, Vespignani A (2001) Epidemic Spreading in Scale-Free Networks. Physical Review Letters 86: 3200–3.
17. Meloni S, Arenas A, Moreno Y (2009) Traffic-driven epidemic spreading in finite-size scale-free networks. Proceedings of the National Academy of Sciences of the United States of America 106: 16897–902.
18. Watts DJ (2002) A simple model of global cascades on random networks. Proceedings of the National Academy of Sciences of the United States of America 99: 5766–71.
19. Buldyrev SV, Parshani R, Paul G, Stanley HE, Havlin S (2010) Catastrophic cascade of failures in interdependent networks. Nature 464: 1025–8.
20. Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. Nature 406: 378–82.
21. Kaiser M, Martin R, Andras P, Young MP (2007) Simulation of robustness against lesions of cortical networks. European Journal of Neuroscience 25: 3185–92.
22. Estrada E (2007) Topological structural classes of complex networks. Physical Review E 75: 1–12.
23. Barabási AL (2009) Scale-free networks: a decade and beyond. Science 325: 412–3.
24. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440–2.
25. Schnettler S (2009) A structured overview of 50 years of small-world research. Social Networks 31: 165–78.
26. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Reviews of modern physics 74: 47–97.
27. Newman MEJ (2003) The Structure and Function of Complex Networks. SIAM Review 45: 167–256.
28. Newman MEJ, Barabási AL, Watts DJ (2006) The Structure and Dynamics of Networks. NJ: Princeton University Press.
29. Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411: 41–2.
30. Rodrigues FA, Costa LDF (2009) Protein lethality investigated in terms of long range dynamical interactions. Molecular BioSystems 5: 385–90.
31. Perc M (2009) Evolution of cooperation on scale-free networks subject to error and attack. New Journal of Physics 11: 033027.
32. Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E (2007) A model of Internet topology using k-shell decomposition. Proceedings of the National Academy of Sciences of the United States of America 104: 11150–4.
33. Seidman SB (1983) Network Structure and Minimum Degree. Social Networks 5.
34. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, et al. (2010) Identification of influential spreaders in complex networks. Nature Physics 6: 888–93.
35. Costa LDF, Rodrigues FA, Travieso G, Boas PRV (2007) Characterization of complex networks: A survey of measurements. Advances in Physics 56: 167–242.
36. Newman MEJ, Leicht EA (2007) Mixture models and exploratory analysis in networks. Proceedings of the National Academy of Sciences of the United States of America 104: 9564–9.
37. Wang J, Lai CH (2008) Detecting groups of similar components in complex networks. New Journal of Physics 10: 123023.
38. Costa LDF, Rodrigues FA (2009) Seeking for simplicity in complex networks. Europhysics Letters 85: 48001.
39. Gross T, Blasius B (2008) Adaptive coevolutionary networks: a review. Journal of the Royal Society, Interface 5: 259–71.
40. Saavedra S, Reed-Tsochas F, Uzzi B (2008) Asymmetric disassembly and robustness in declining networks. Proceedings of the National Academy of Sciences of the United States of America 105: 16466–71.
41. Costa LDF, Rodrigues FA, Hilgetag CC, Kaiser M (2009) Beyond the average: Detecting global singular nodes from local features in complex networks. Europhysics Letters 87: 18008.
42. Ramasco JJ, Mungan M (2008) Inversion method for content-based networks. Physical Review E 77: 036122.
43. Costa LDF, Kaiser M, Hilgetag CC (2007) Predicting the connectivity of primate cortical networks from topological and spatial node properties. BMC Systems Biology 1: 16.
44. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network motifs: simple building blocks of complex networks. Science 298: 824–7.
45. Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis. Prentice Hall Englewood Cliffs, NJ, USA, 6th edition, 800 p.
46. Parzen E (1962) On Estimation of a Probability Density Function and Mode. The annals of mathematical statistics 33: 1065–76.
47. Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley Interscience, 2nd edition, 680 p.
48. Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences). Cambridge University Press.

49. Kuramochi M, Karypis G (2001) Frequent Subgraph Discovery. In: Proceedings of the 2001 IEEE International Conference on Data Mining IEEE Computer Society. pp 313–20.
50. Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics 20: 1746–58.
51. Middendorf M, Ziv E, Wiggins CH (2005) Inferring network mechanisms: the Drosophila melanogaster protein interaction network. Proceedings of the National Academy of Sciences of the United States of America 102: 3192–7.
52. Bordino I, Donato D, Gionis A, Leonardi S (2008) Mining Large Networks with Subgraph Counting. In: 8th IEEE International Conference on Data Mining (ICDM). IEEE.
53. Ribeiro P, Silva F, Kaiser M (2009) Strategies for Network Motifs Discovery. In: 5th IEEE International Conference on e-Science. IEEE. pp 80–7.
54. Groening M (2005) The Simpsons Uncensored Family Album. HarperCollins Entertainment.
55. Erdös P, Rényi A (1959) On Random Graphs I. Publ Math (Debrecen) 6: 290–7.
56. Barabási AL, Albert R (1999) Emergence of Scaling in Random Networks. Science 286: 509–12.
57. Kaiser M, Goerner M, Hilgetag CC (2007) Criticality of spreading dynamics in hierarchical cluster networks without inhibition. New Journal of Physics 9: 110.
58. Costa LDF, Silva FN (2006) Hierarchical Characterization of Complex Networks. Journal of Statistical Physics 125: 841–76.
59. Milgram S (1967) The small world problem. Psychology today 2: 60–7.
60. Kochen M, ed (1989) The Small World. Ablex Publishing, Norwood, NJ.
61. Guare J (1994) Six degrees of separation: a play. Vintage Books, 2 edition.
62. Eaton JW (2002) GNU Octave Manual. Limited, Network Theory.
63. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, et al. (2004) Superfamilies of evolved and designed networks. Science 303: 1538–42.
64. Kaiser M, Hilgetag CC (2004) Edge vulnerability in neural and metabolic networks. Biological cybernetics 90: 311–7.
65. MacKay DJ (2003) Information theory, inference, and learning algorithms. Cambridge University Press, 1st edition.
66. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Computing Surveys 31: 264–323.
67. Arthur D, Vassilvitskii S (2007) k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics. pp 1027–35.
68. Ozik J, Hunt BR, Ott E (2004) Growing networks with geographical attachment preference: Emergence of small worlds. Physical Review E 69: 026108(5).
69. Mahalanobis PC (1936) On the generalized distance in statistics. Proceedings of the National Institute of Science of India 2: 49–55.