New Journal of Physics
The open-access journal for physics

# Beyond centrality—classifying topological significance using backup efficiency and alternative paths*

**Yuval Shavitt**[1] **and Yaron Singer**[2,3]

[1] School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel
[2] School of Computer Science, Tel Aviv University, Tel Aviv, Israel
E-mail: yaronsin@post.tau.ac.il

*New Journal of Physics* **9** (2007) 266
Received 27 March 2007
Published 14 August 2007
Online at http://www.njp.org/
doi:10.1088/1367-2630/9/8/266

**Abstract.** In complex networks characterized by broad degree distribution, node significance is often associated with its degree or with centrality metrics which relate to its reachability and shortest paths passing through it. Such measures do not consider availability of efficient backup of the node and thus often fail to capture its contribution to the functionality and resilience of the network operation. In this paper, we suggest the quality of backup (QoB) and alternative path centrality (APC) measures as complementary methods which enable analysis of node significance in a manner which considers backup. We examine the theoretical significance of these measures and use them to classify nodes in social interaction networks and in the Internet AS (autonomous system) graph while applying the valley-free routing restrictions which reflect the economic relationships between the AS nodes in the Internet. We show that both node degree and node centrality are not necessarily evidence of its significance. In particular, we show that social structures do not necessarily depend on highly central nodes and that medium degree nodes with medium centrality measure prove to be crucial for efficient routing in the Internet AS graph.

* Partial results presented in this paper also appear in the IFIP NETWORKING 2007 Conference.
[3] Author to whom any correspondence should be addressed.


PII: S1367-2630(07)47079-3
1367-2630/07/010266+17$30.00

## Contents



## 1. Introduction

The topological study of networks appears in a wide spectrum of research areas such as physics [1], biology [2], mathematics [3, 4], computer science [5] and the social sciences [6]. The availability of data in recent years allows exploration of structural properties of networks which often significantly contributes to comprehension of their inherent traits. In systems biology, for instance, the difficulties of modeling complex cellular networks leaves the relationship between the network's topology, and function as an open question [7]–[9]. Specifically, the unavailability of mechanistic detail and kinetic parameters imposes difficulties on mathematical modeling of complex cellular networks. In contrast, the network topology is known in many cases and thus enables structural analysis which sheds light on key aspects of network properties inferred from its structure alone [10].

The topological study of networks can be conducted at three different levels of granularity: study of global behavior and properties of the network, study of inherent structures and patterns, and study of individual nodes and links and their function in the network. On the global scale, examples include discovery of properties such as the degree distribution [11] and investigation of resiliency to failures [1, 12]. Study on a global scale also includes analysis under changing physical conditions in biological networks [13] and changing conditions over time in various networks [14]. The study of structures within networks includes classic works which investigate motifs [2] and communities [15] as well as recent findings of fractals patterns [16]. Finally, on the node and link scale works range from associating central nodes with key biological functions [17], to using to topological metrics to rank web-page relevance in a web search [5].

In this work, we focus our attention on the significance of individual nodes within complex networks by considering their backup and quantifying their contribution to the networks' functionality. As network functionality is often measured by connectivity and vertex distances in the graph used as its model, measures which credit vertices connected to a relatively large number of vertices at relatively short distances are often used as significance indicators [1, 17]. However, inadequate consideration of backup by such measures often overshadows significance in context of its contribution to functionality and resilience of the network. Existence of backup raises question regarding a node's significance since failure of a

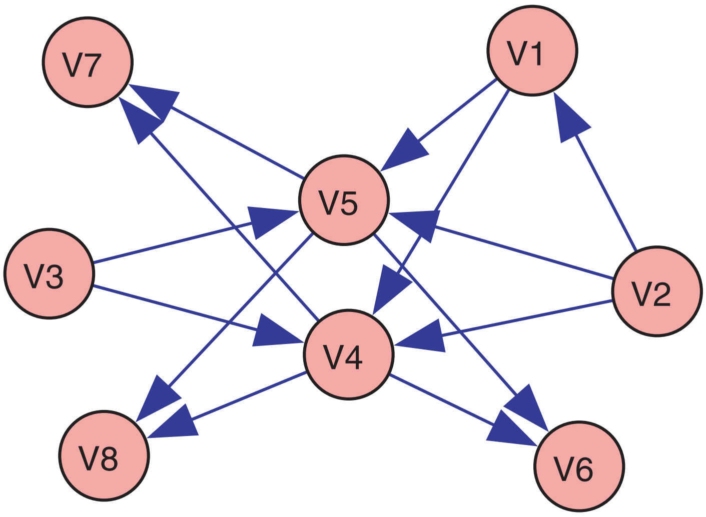


**Figure 1.** An illustration of an instance where a vertex ($v_4$) is central in the graph (connected to a relatively large number of vertices at short distances) though has a backup vertex ($v_5$) which raises questions regarding its contribution to functionality in the network modeled.

node with backup does not effect connectivity nor does it increase path lengths in the network and therefore the effect of failure in such instances is minimal (see the example in figure 1). Furthermore, existence of backup denies exclusivity of the information passing through the node in the network. Since nodes can have backups of various qualities, measures of backup efficiency and topological significance which considers backup are crucial for analysis of network functionality.

We suggest two complementary measures which capture a node's contribution to the network's functionality: the quality of backup (QoB) and the alternative path centrality (APC). The QoB measures backup quality of a vertex regardless of its centrality, enables comparison of backup efficiency between vertices in the graph as well as between vertices from different graphs, and can thus serve as a universal measure for backup. The APC considers both backup quality and centrality of vertices in graphs and therefore enables analysis of nodes' contribution to centrality in a wider context in comparison to other topological measures.

We begin with examining our methods on levels of theoretical abstraction, and then use APC and QoB to analyze social interaction networks and the Internet AS (autonomous system) graph. In our study, we use APC to identify the most significant nodes to the networks' functionality, and show that these are not necessarily the most connected individuals or the Internet core. In analysis of social networks, we show that there is some correlation between APC and the betweenness metric which was used to detect community structures in the social interaction network studied. We also show that APC reveals structural difference between communities centered around individual nodes. In accordance with properties of APC, it is not surprising that the largest Internet service providers in the core, such as UUNET and Sprint, also have very high APC values due to the large number of customers that use them as their only connection to the Internet. However, small service providers with poor backup like the French research network RENATER, and the GEANT and Abiline academic backbones which have degrees as low as 51 (RENATER) and low centrality values, have very high APC values as well.

The rest of this paper is organized as follows. The next section discusses the concept of backup in networks and introduces the QoB as a measure of universal backup efficiency. Section 3 provides detail of the APC construction and discusses its properties. We use the APC measure to analyze a dolphin social interaction network in section 4, as well as a network which describes friendship patterns in a karate club. In order to maintain relevance in the Internet AS graph model we introduce modifications of our methods, and use the modified measures in comparison to previous works done on the Internet AS graph in section 5.

## 2. Quantifying backup efficiency in complex networks

When considering a node's contribution to the functionality of the network in terms of its effect on path lengths and connectivity, backup plays a fundamental role. Trivially, failure of a node $v$ does not effect network functionality when a backup node $b$, connected to all of $v$'s neighbors, is available. However, such instances in the network are usually rare. A more common scenario is one where $v$ has several nodes connected to a subset of its neighbors, reached at various path distances by nodes leading to $v$. It is therefore clear that backup of a node may be distributed in the network and vary in efficiency.

In our attempt to quantify backup in networks, we observed that the QoB of backup of a given vertex in the graph is determined by the number of direct children covered by a set of backup vertices, and the efficiency of reaching this backup set by the set of direct parents. For a vertex $v \in V$, we define the set of children $C_v$, set of parents $P_v$, and the backup set $B_v$, as follows:

$$C_v = \{u \in V | (v, u) \in E\}, \quad P_v = \{u \in V | (u, v) \in E\}, \quad B_v = \{w \in V | \exists u \in C_v : (w, u) \in E\}.$$

Clearly, in instances of undirected graphs $C_v \equiv P_v$, and the discussion which follows remains relevant in these instances as well.

For $u, w \in V$, we use $\delta(u, w)$ to denote the shortest path distance between $u$ and $w$ in $G$. The shortest distance, $\delta(u, w)$, can be calculated by any set of rules, e.g. based on additional annotations on the graph edges, and is not limited to minimum hop. By convention, if $u$ cannot reach $w$ through any path in $G$, then $\delta(u, w) = \infty$ and $\delta(u, u) = 0$. We also use $\delta_v(u, w)$ to represent the distance of the shortest path which bypasses $v$ from $u$ to $w$.

To discuss the properties of the methods suggested here, we formally define the intuitive concept of backup efficiency in a graph in the following manner.

**Definition 1.** *Let $G = (V, E)$ and $v \in V$ be a vertex with the set of direct children $C_v \subseteq V$, and the backup set $B_v \subseteq V$. We say that $v$ has a $d$-DISTANT BACKUP SET OF ORDER $k$, where $k = |C_v \cap (\cup_{b \in B_v} C_b)|$ and $d = \max\{\delta_v(u, w) | u \in P_v, w \in C_v\}$.*

An example of the definition is shown in figure 2. Note that since our definition above refers to the *maximal* distance between vertices from the sets $P_v$ and $C_v$, $\infty$-distant backup of a vertex is not necessarily a testament of ineffective backup. A vertex $b$ can be an $\infty$-distant backup set of order $|C_v|$ of $v$, reachable by all vertices that reach $v$, but one. Despite ambiguity that may rise from instances of $\infty$-distant backup of a vertex, this definition serves its purpose as it provides a simple indication of backup efficiency. The methods suggested here are not effected by this formal obstacle as they consider overall efficiency in terms of alternative shortest paths.

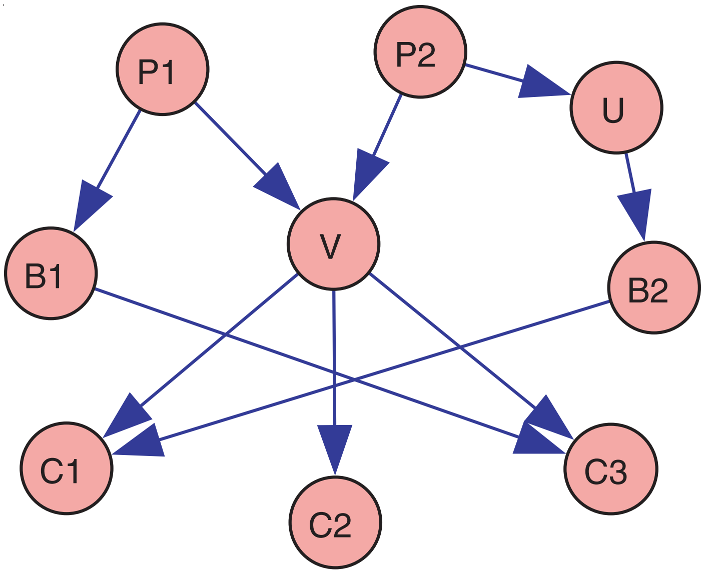


**Figure 2.** An example of a network with a node $v$ which has a three-distant backup set of order 2: $v$'s parents $p_1, p_2$ reach two of $v$'s children—$c_1, c_3$, through the backup set $b_1, b_2$.

### *2.1. Measuring QoB*

Let $G=(V, E)$ be a directed or undirected graph where $V$ is the set of vertices and $E$ is the set of edges. The QoB of $v \in V$, denoted $\rho(v)$ is:

$$\rho(v)=\frac{\sum_{u\in P_v}\sum_{w\in C_v}(\max\{\delta_v(u,w)-1,1\})^{-1}}{|P_v|\cdot|C_v|}.$$

The rational behind this measure is the following. To measure backup efficiency of a given vertex, it is enough to examine the cost of re-routing paths from its set of parents to its set of direct children. Note that

$$\max\{\delta_v(u,w)-1,1\}=\delta_v(u,w)-\delta(u,w)+1$$

for all pairs $\langle u, w\rangle$ such that $u\in P_v \wedge w\in C_v$. In instances where a parent vertex, $u$, is directly connected to a child vertex, $w$, we have $\delta_v(u,w)=\delta(u,w)=1$. We, therefore, choose the maximal value between $\delta_v(u,w)-1$ and 1 to ensure that the sum remains finite. Note that this is equivalent to a definition of $\rho(v)$ in which the sets of parents and children of $v$ used are not directly connected. This enables measuring $v$'s backup only in paths which use $v$. For a vertex $v\in V$, it is easy to see that

$$\rho(v)=1 \iff \forall\langle u,w\rangle\in P_v\times C_v \quad \exists b\in B_v : \delta(u,b)\leqslant 1\wedge\delta(b,w)=1$$

and that

$$\rho(v)=0 \iff \delta_v(x,u)=\infty \quad \forall\langle x,u\rangle\in P_v\times C_v.$$

Thus, $\rho : V \longrightarrow [0,1]$, and returns 1 for vertices with perfect backups and 0 for vertices with no backup. The following theorem shows the QoB measure indeed enables local measurement of a vertex's backup in the graph.

**Theorem 1.** *For $G = (V, E)$, for a vertex $v \in V$ with $P_v \neq \emptyset$ and $C_v \neq \emptyset$, $\rho(v)$ monotonically increases with respect to rise in backup efficiency.*

**Proof.** Suppose $v$ has a $d$-distant backup set of order $k$. Let $u \in P_v$ and $w \in C_v$, such that $\delta_v(u, w) = c$ in $G$, where $1 < c \leqslant \infty$, and $c = \min\{\delta_v(p, c) | p \in P_v, c \in C_v\}$. Construct $G'$ by adding some edge $e \notin E$, such that $\delta'_v(u, w) = c' < c$, where $\delta'_v(u, w)$ represents the distance between $u$ and $w$ bypassing $v$ in $G'$. Notice that it is enough to prove for decrement in $d$ only, since increase in $k$ is equivalent in the case that $c = \infty$ (by minimality of $c$). By our construction $v$ has a $d'$-distant backup set of order $k'$, where $d' \leqslant d$, and $k' > k$. We therefore have

$$(\max\{\delta'_v(u, w) - 1, 1\})^{-1} \geqslant (\max\{\delta_v(u, w) - 1, 1\})^{-1},$$

where equality holds only for $c = 2$. It thus easily follows that $\rho'(v) > \rho(v)$, where $\rho'(v)$ is the QoB measure of $v$ in $G'$. □

## 3. APC

The above section discusses backup efficiency regardless of centrality considerations. In an attempt to quantify significance, note that centrality of a node in the network (its ability to reach a relatively large number of nodes efficiently) also plays a vital role in analysis: a node which has relatively efficient backup can be crucial to the network's functionality due to its high centrality, while a node with poor backup and low centrality can have little effect on functionality in the network. The APC measure presented in this section enables quantifying topological contribution of a node to the functionality of the network as it considers both centrality and backup efficiency.

Given a graph $G = (V, E)$ as above and $u \in V$, the topological centrality measure used here, denoted $\chi$, where $\chi : V \longrightarrow R$ is:

$$\chi(u) = \sum_{w \in V \setminus \{u\}} \frac{1}{\delta(u, w)}.$$

Clearly, $0 \leqslant \chi(u) \leqslant |V| - 1, \ \forall u \in V$.

For a vertex $u \in V$, the value of $\chi(u)$ depends on the number of vertices connected to $u$ and their distances from it; $\chi$ monotonically increases with respect to both centrality and connectivity of the vertex. Thus, in relation to other vertices in the graph, high $\chi$ values are obtained for a vertex which is connected to a large number of vertices at short distances. Symmetrically, a vertex connected to a small number of vertices at large distances yields low $\chi$ values. These properties make the $\chi$ function a favorite candidate for measuring vertices' centrality in the network. In [18], the average of $\chi$ values in the graph was used to define the *efficiency* of the network. Similar topological measures have also been used in [1] and in [17] to study functionality in complex networks.

For $G = (V, E)$, the APC value of $v \in V$, denoted $\varphi(v)$ is:

$$\varphi(v) = \sum_{u \in V \setminus \{v\}} \sum_{w \in V \setminus \{u, v\}} \frac{1}{\delta(u, w)} - \frac{1}{\delta_v(u, w)}.$$

That is, we calculate the difference between the centrality values of all the vertices in the graph, except $v$, and the centrality values using shortest paths which bypass $v$. In calculating centrality, we use a slight variation of $\chi$ as defined above, since we measure the distance from a given vertex $u$ to all vertices in $w \in V \backslash \{u, v\}$, rather than all$w \in V \backslash \{u\}$, as used in the definition of $\chi$ above. In our discussion below, when referring to $\chi$ in the context of the APC metric, we consider this variation.

The rational behind APC is simple. In instances where network functionality is determined by shortest paths and connectivity, the significance of a node $v$ to the network's functionality can be measured by its effect on these criteria. Computing the difference between vertices' topological centrality using $v$, and topological centrality bypassing $v$, enables witnessing $v$'s exclusive contribution to the network's functionality. We conclude our discussion of the APC properties with the following theorem which shows that APC properly considers both centrality and backup of a vertex in the graph.

**Theorem 2.** *For $G = (V, E)$ and $v \in V, C_v \neq \emptyset, \varphi(v)$, $\varphi(v)$ monotonically increases with respect to rise in topological centrality and decrement in backup quality.*

**Proof.** To prove $\varphi(v)$ monotonically increases with respect to rise in centrality, let $\chi(v) < |V| - 1$, and $w \in V$ be a vertex for which $1 < \delta(v, w) \leqslant \infty$. Let $e / \in E$ be some edge for which $\delta'(v, w) < \delta(v, w)$, where $\delta'(v, w)$ denotes the shortest path distance in $G' = (V, E \cup \{e\})$, and $e$ does not create new alternative paths to $w$ in $G'$ (otherwise backup efficiency increases). We show that $\varphi(v) < \varphi'(v)$, where $\varphi'(v)$ denotes the APC value of $v$ in $G'$. For all $x \in V$, which reach $w$ through $v$, $\delta'(x, w) < \delta(x, w)$ and $\delta'_v(x, w) = \delta_v(x, w)$. For all such vertices, $x$, we have

$$\frac{1}{\delta'(x, w)} - \frac{1}{\delta'_v(x, w)} > \frac{1}{\delta(x, w)} - \frac{1}{\delta_v(x, w)}$$

and it therefore follows that $\varphi(v) < \varphi'(v)$.

To show monotonic increase with respect to decrement in backup quality, according to the properties of our QoB measure, it is enough to show monotonic increase with respect to monotonic decrement in $\rho(v)$. Assume that some edge $e / \in E$ has been added to $G$, such that $\rho(v)$ increases. We again denote $G' = (V, E \cup \{e\})$, and use similar notation as above. We therefore assume $\rho'(v) > \rho(v)$. Specifically, there is some pair $\langle u, w\rangle \in P_v \times C_v$ such that $\delta'_v(u, w) < \delta_v(u, w)$. For this pair we have

$$\frac{1}{\delta'(u, w)} - \frac{1}{\delta'_v(u, w)} < \frac{1}{\delta(u, w)} - \frac{1}{\delta_v(u, w)}$$

and it trivially follows that $\varphi'(u) < \varphi(u)$, which concludes our proof of the theorem. $\square$

## 4. Analysis of social networks

In this section, we use the APC measure for the benefit of social network analysis. We first analyze a social interaction network between bottlenose dolphins, and continue with analysis of a network which describes social interactions between individuals in a karate club in a US university.

The dolphin social interaction network has been constructed from observations of a community of 62 bottlenose dolphins in Doubtful Sound, New Zealand, over a period of

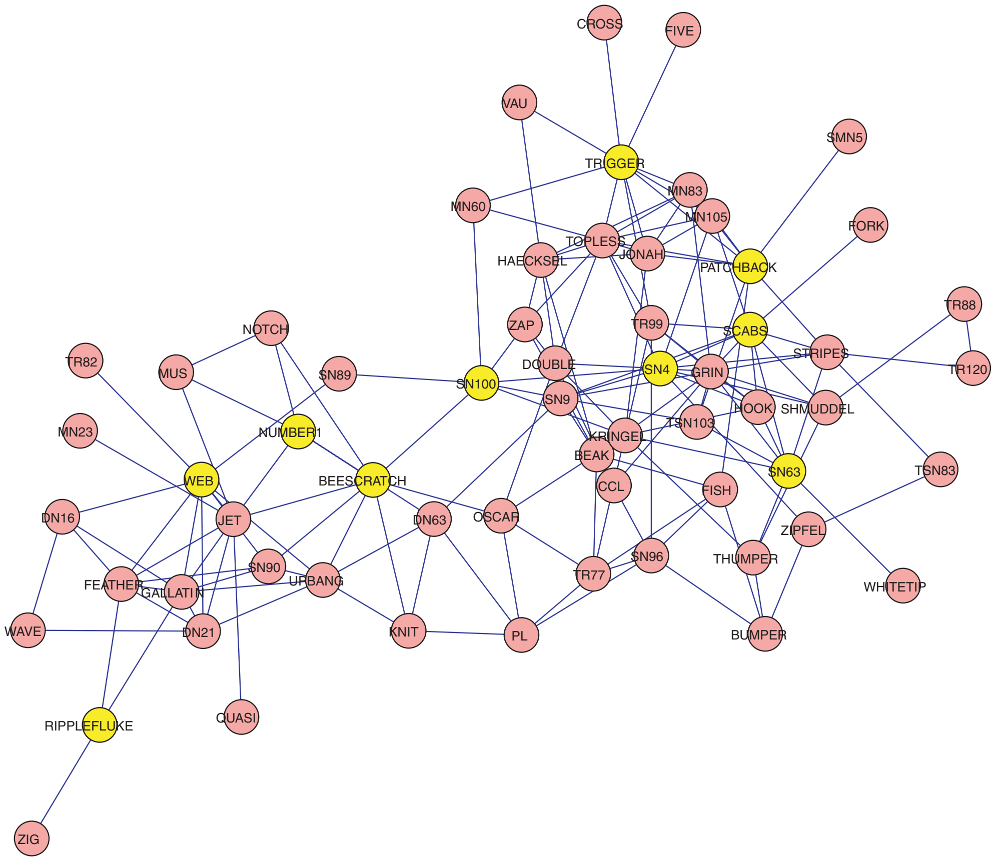


**Figure 3.** The dolphin social interaction network. Nodes in yellow are members in the top 10 APC list.

7 years (1994–2001) [19]. The nodes represent the dolphins and the links represent their social interactions which were observed and considered statistically significant.

We first applied the APC measure on the network, and categorized the nodes in the network by their descending APC values. In figure 3, we illustrate the network, and highlight the top 10 highest scoring APC nodes in the network. The top three APC nodes are TRIGGER, NUMBER1, and PATCHBACK. For further analysis, we measured the centrality of the nodes in the network (as defined by the $\chi$ function) and their degree. In order to evaluate the APC measure, we plotted the average APC values as a function of the nodes' degree, and did the same for their centrality values. The result is displayed in figure 4 (top). While centrality is highly related to the nodes' degree, it is evident that this does not apply to the APC measure; nodes with high degree do not necessarily have high APC values, and the nodes with the highest APC values are medium—high degree nodes. In comparison to the centrality metric, APC considers backup and thus can be expected to better expose a node's contribution to functionality in the network.

The dolphin network has also been studied in [20]. In their work, Lusseau and Newman focused on the community structures in the network and the role of the individual nodes in maintaining the cohesion of these structures. For this task, nodes were ranked by their betweenness values—the number of shortest paths passing through each node. Out of the nine

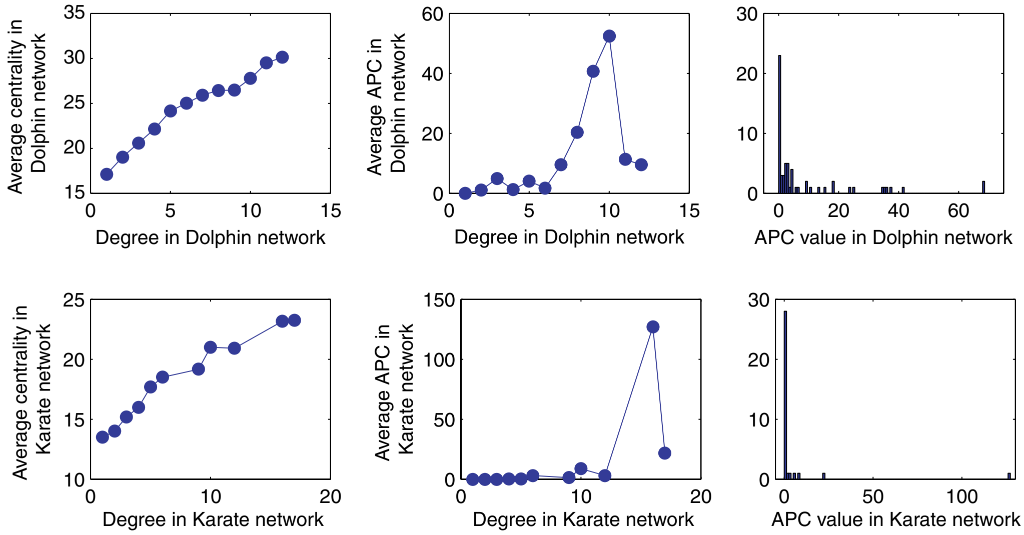


**Figure 4.** The connection between node degree, centrality, and APC in the dolphin and karate networks, and the histogram of the APC values in both networks.

nodes which Lusseau and Newman identified to be cardinal to maintaining the community structures they discovered, five are members in our top 10 APC list (TRIGGER, WEB, SN100, BEESCRATCH and SN4). The rest of the nodes which top the APC list (NUMBER1, PATCHBACK, SCABS, SN63 and RIPPLEFLUKE) are considered less significant according to the Lusseau and Newman metric. The distribution of the APC values in the network are also presented in figure 4.

The 'karate club' network is a famous social network which is often used to test clustering algorithms [21]. The network models friendship patterns between 34 members in a karate club at a US university and has been introduced by Zachary [22]. In [21] the application of the eigenvector-based community detection algorithm has shown that the network is divided into two communities. This finding correlates to the split of the club into two due to an internal dispute, shortly after the network was measured.

In our analysis, we find again that high degree is not necessarily an indicator to a node's contribution to functionality, and the behavior of average centrality and average APC as a function of node degree are similar to those measured on the dolphin network as shown in figure 4. Inspection of the distribution of APC values in the network also in figure 4, shows that the karate network includes only one node with a significantly high APC value. This is in contrast to the dolphin network where several individuals were found to have high APC values. By inspection of the graph in figure 5 one can witness the two communities which are discussed in [21], characterized by the density around the colored nodes. Further analysis, however, indicates that the two communities are inherently different in structure; while one community is centralized around one node (marked in red), the second community is centralized around two nodes (marked in yellow), which symmetrically serve as each other's backup. The node marked in red is the top scoring APC node, and the only one with significant APC value as can be seen from the APC distribution in figure 4.

The karate network proves to be an interesting case study when applying the APC metric. While there are four nodes with high centrality values (values above 20) in the network, only

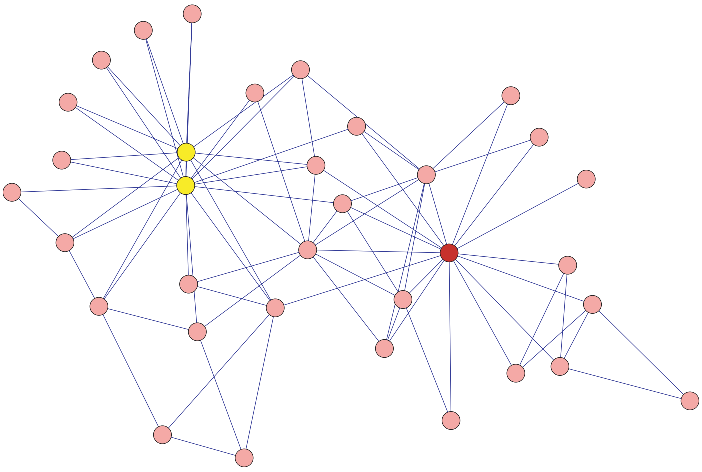

**Figure 5.** The 'karate club' network. The node marked in red is the one with the highest APC value in the network. The two nodes in yellow are both with high centrality and low APC, due to their symmetrical backup of each other.

one of these nodes has a significantly high APC value (the second highest central node), due to the backup of the rest of the highly central nodes. In study of the community structure, the APC measure shows the difference between a community which revolves around highly central nodes, that can each fail without affecting the community and the network, and a community which depends on a single node for its structural survival.

## 5. Analysis of the Internet AS graph

In research of the Internet, node significance classification has received attention in past studies [23]–[25] and was treated in two different contexts: study of the Internet resiliency against attacks and failures [1, 12], [26]–[29] and identification of the Internet core nodes [23, 25]. Both study threads were conducted at the level of the Internet AS graph which models the interconnections of the Internet's service providers.

Several attempts have been made in the past to characterize the core of the Internet AS graph. In [23] the most connected node was used as the natural starting point for defining the Internet's core. Other autonomous systems (ASes) were also classified to four shells and tendrils that hang from the shells, where ASes in shells with a small index are considered more important than ones in higher indices. Further work has dealt with classification of nodes into few shells with decreasing importance [30, 31]. In a recent study [25] *k*-shell graph decomposition was used to classify nodes by importance to roughly 40 layers of hierarchial significance. The *k*-shell classification, based on the node's connectivity, identified over 80 ASes as the Internet core, some of which with medium degrees. Almost exclusively, an attempt to rank ASes by metrics other than node degree was done by CAIDA [32], where the 'cone' of the node was used to determine its importance, namely the number of direct and indirect customers of the AS. In this section, we use the APC and QoB measures to analyze the Internet AS graph. Since failure of a node on the AS level is possible [12] though highly unlikely, applying APC and QoB on the Internet AS graph allows a unique insight to the Internet structure as opposed to quantifying effects of failures. On the AS level, centrality which considers backup reveals significance in context of potential information which exclusively passes through a node, and its backup quantifies the dependency of its customers on its transit services. We begin with a brief description of the AS graph model.

### 5.1. The Internet AS graph

The Internet today consists of tens of thousands of networks, each with its own administrative management, called ASes. Each such AS uses an interior routing protocol (such as OSPF or RIP) inside its managed network, and communicates with neighboring ASes using an exterior routing protocol, called BGP. The graph which models inter-connection between ASes in the Internet is referred to as the Internet AS graph. Since the ASes in the Internet are bound by commercial agreements, restrictions are imposed on the paths which may be explored. The commercial agreements between the ASes are characterized by customer–provider, provider–customer and peer-to-peer relations. A customer pays its provider for transit services, thus the provider transits packets to and from its customers. The customer, however, will not transit packets for its provider. Specifically, a customer will not transit packets between two of its providers, or between its provider and its peers. Peers are two ASes that agree to provide transit information between their respective customers.

In a pioneering work, Gao [30] has deduced that a legal AS path may either be an *up hill* path, followed by a *down hill* path, or an *up hill* path, followed by a peering link, followed by a *down hill* path. An *up hill* path is a sequential set, possibly empty, of customer–provider links, and a *down hill* path is a sequential set, possibly empty, of provider–customer links. Therefore a legal route between ASes can be described as a *valley free* path. A peering link can be traversed only once in each such path, and if it exists in the path it marks the turning point for a *down hill* path.

### 5.2. Adaptation of QoB and APC for the directed AS graph

We have adjusted our measures to conform to the model of the AS graph and specifically to the routing restriction which it imposes. Since transitivity is not immediate in the AS graph, the QoB requires two cardinal adjustments to maintain relevance. Consider the AS graph $G = (V, E)$ and some $v \in V$, for which we wish to obtain $\rho(v)$ in $G$. Let $u \in P_v$ and $w \in C_v$. The first adjustment is to calculate $\delta(u, v)$ using valley free routing. Since the BFS algorithm for discovery of shortest paths in unweighted graphs does not consider the *up*, *down* and *peer* labels, *valley free* paths are not exclusively discovered, and it cannot be used to measure minimum-hop distances in the AS graph. For this, we use the ASBFS algorithm [33] which discovers *valley free* shortest paths from a source vertex in the unweighted AS graph in linear time.

In order to provide motivation for the second adjustment required, we present the following theoretical example. Consider the graph illustrated in figure 6. In quantifying the QoB of $v \in V$. Suppose a vertex $u \in P_v$ has reached a vertex $w \in C_v$ through an *up hill* path through $v$, though by using the alternative path through the vertex $y \in B_v$, $u$ now reaches $w$ through a *down hill* path. All vertices in $C_w$ which are reached through an *up hill* path ($x$ in this example), are now unreachable to $u$ as this creates an illegal AS path. Therefore, to factor this into the QoB measure in the AS graph, we use the following strategy. For all vertices $w \in C_v$, we scan for vertices $x \in C_w$ which are reachable from $v$ through legal AS paths, and consider the pairs $\langle u, x\rangle \in P_v \times C_w$ as well. Since these instances occur when a customer AS, $u$, uses a provider AS, $w$, which is a provider another one of $u$'s providers, $v$, we only use these two levels in our

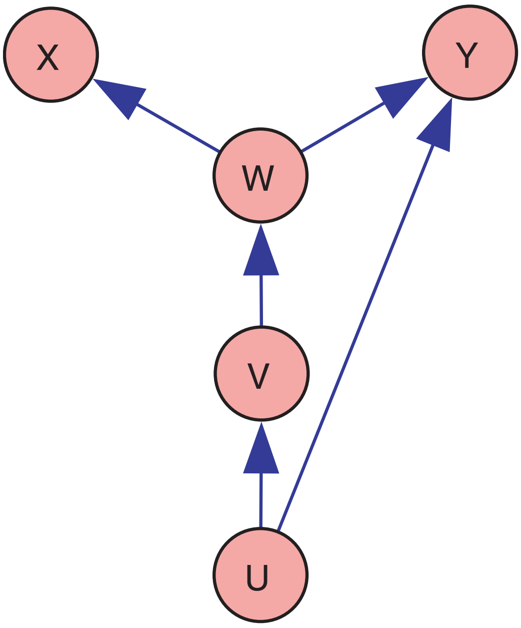


**Figure 6.** Illustration of an instance in an AS graph where a direct child can be reached through a backup vertex, though its paths cannot be used. Direction of an edge implies it is an *up* edge, and for each *up* edge a *down* edge in the opposite direction exists (not portrayed). Here, $y$ serves as a backup for $v$. In accordance with the *valley free* restrictions, $u$ can reach $w$, but cannot reach $x$ through $y$.

adjustment. Indeed, in such backup paths are usually limited to four hops in the Internet AS graph [23].

After adjusted accordingly, the QoB remains faithful to the principles of measuring backup efficiency in the AS graph. For $v \in V$, as reachable children are scanned in two levels, we are guaranteed that $\rho(v) = 1$ if and only if $v$ has a perfect backup which does not disqualify legal AS paths.

Substituting the BFS algorithm with its analogous for the AS graph, ASBFS, applying APC on the AS graph is immediate. The calculation of a shortest path, $\delta$, is done while considering the *valley free* routing and all the properties discussed in section 3 hold.

### *5.3. Results on the Internet AS graph*

We used the combined data from the DIMES [34] and RouteViews [35] projects for week 11 of 2006. The AS graph is comprised of 20 103 ASes and 57 272 AS links. We approximate the AS relationship by comparing the $k$-core index [25] of two ASes and taking the one with the highest $k$-core index as the provider of the other. If the $k$-core indices of two ASes are equal, the ASes are treated as peers. While we are aware that our approximation involves some inaccuracies, there is no known error free algorithm for this task. Since the majority of the interesting ASes are within the range of AS numbers 1–22 000, we present results of these 11 407 ASes along with the results of ASes with degree higher than 40 of the rest of AS graph.

We first show that while centrality is closely related to the node degree in the AS graph, our APC criteria captures significance which is not necessarily associated with high degree. Figure 7 shows the centrality values of AS nodes averaged by their degree on a log–log scale.

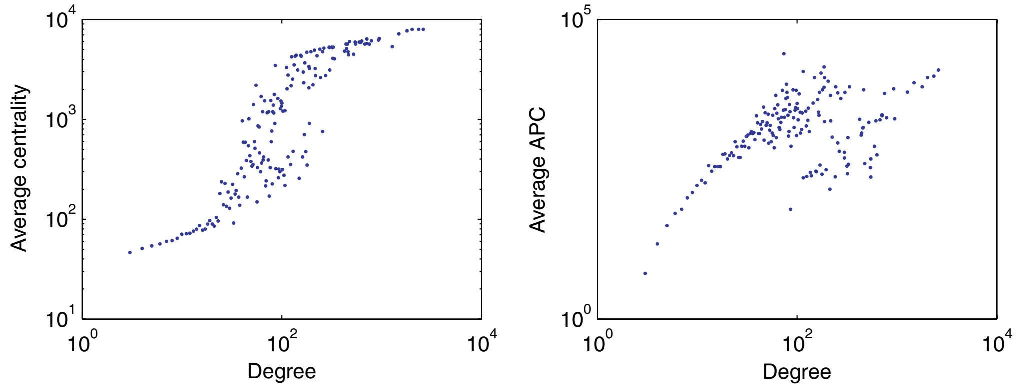


**Figure 7.** Centrality and APC in the Internet AS graph. Average centrality as a function of its degree (left) and average APC as a function of its degree, as measured on the Internet AS graph.

There is almost a monotonic increase in centrality for nodes of degree above 300, and the close relationship between centrality and degree is evident. Since high degree is an excellent predictor of the node centrality, centrality does not add new insight into the significance of a network node. On the other hand, figure 7 shows that even nodes with medium degree may have high APC values.

To display the relationship between high centrality and high APC we plot the degree and APC values of the nodes with the highest centrality (figure 8 left) and the degree and centrality of the nodes with the highest APC values (figure 8 right). The five ASes with the highest degree, 701 (UUNET), 7018 (AT&T), 1239 (Sprint), 3356 (Level 3) and 174 (Cogent), are also the five ASes with the highest centrality. These are the largest tier-1 providers. In contrast, only UUNET is in the top ten APC list; Sprint and Cogent also have high APC values. These three tier-1 providers support many stub ASes but have relatively low backup measure (0.7–0.75) which explains their high APC values. Level 3, which has high centrality, has low APC value because it has a rather high QoB around 0.82. This means that although Level 3 (3356) plays a central role in Internet routing, it may be replaced through alternative routes and thus is not as important as the previous three nodes. The next nodes with high centrality are 3549 (GBLX), 2914 (Verio), 7132 (SBC), 6461 (Abovenet) and 12956 (Telefonica). These are all tier-1 providers or major providers in Europe.

For the nodes with the highest APC values the picture is different: while UUNET (701) has the fourth largest APC value, many of the high locations in the list are captured by medium-sized ASes with poor (and sometimes extremely poor) backup. Through study of the QoB distribution in the AS graph we have learned that there is a large concentration around 1, which is a testament of perfect backup. The median QoB value is 0.9799, and a large majority of the nodes have QoB values above 0.95. The nodes ranked first, third and eighth in the top APC list are educational networks: GEANT (20965) in Europe, ENA (11686) in the USA, and RENATER (2200) in France (Abiline the US research network was ranked eleventh). The other group of nodes is of medium-size providers, France Telecom (3215), YIPES (6517), Ukraine Telecom (6849) and ServerCentral (23352), each appears to have high APC values due to a different reason. France Telecom, YIPES and UKR Telecom have extremely low QoB, while ServerCentral connects

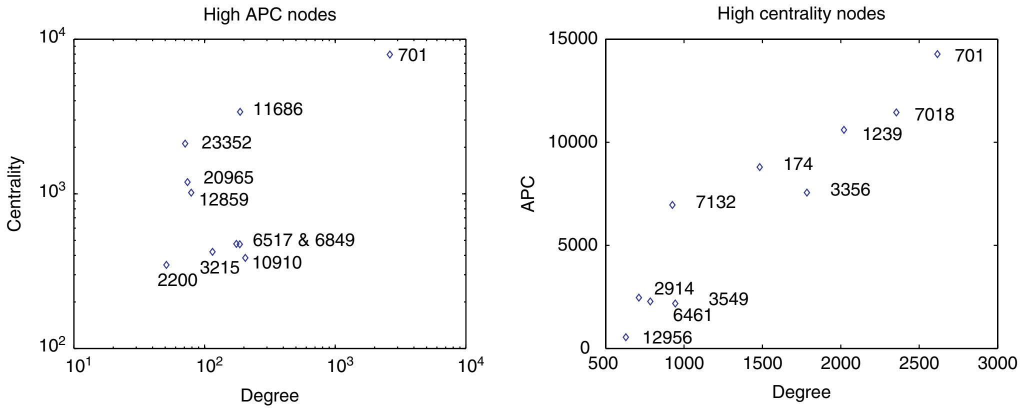


**Figure 8.** The degree and centrality of the nodes with the highest APC values (left), and degree and APC of the nodes with the highest centrality values.

**Table 1.** Statistics of AS nodes with highest APC values (top) and statistics of AS nodes with highest CAIDA significance rankings (bottom).

| AS no. | deg. | cent. | QoB | APC |
|---|---|---|---|---|
| 20965 | 74 | 1190 | 0.79 | 26628 |
| 10910 | 205 | 385 | 0.59 | 16298 |
| 11686 | 187 | 3389 | 0.92 | 16042 |
| 701 | 2616 | 7956 | 0.72 | 14276 |
| 3215 | 115 | 422 | 0.80 | 13493 |
| 6517 | 175 | 474 | 0.83 | 12851 |
| 6849 | 186 | 472 | 0.56 | 12765 |
| 2200 | 51 | 347 | 0.50 | 12549 |
| 12859 | 79 | 1017 | 0.94 | 12396 |
| 23352 | 71 | 2113 | 0.94 | 12065 |
| 3356 | 1784 | 7690 | 0.82 | 7559 |
| 209 | 1272 | 5381 | 0.72 | 6113 |
| 7018 | 2354 | 7992 | 0.74 | 11448 |
| 1239 | 2020 | 8022 | 0.74 | 10604 |
| 701 | 2616 | 7956 | 0.72 | 14276 |
| 3561 | 708 | 5762 | 0.79 | 2579 |
| 174 | 1483 | 7144 | 0.76 | 8797 |
| 703 | 216 | 1441 | 0.86 | 10539 |
| 19262 | 188 | 905 | 0.75 | 10763 |
| 702 | 680 | 5672 | 0.77 | 2101 |

remote locations that may not have efficient alternative paths. Statistics of nodes with highest APC values are displayed in table 1.

Figure 9 (left) shows the distribution of the APC values in the AS graph (note the truncation of the first column). The APC distribution is shown to have a long but narrow tail with only a few nodes with very high APC values, these nodes are scattered almost over the entire degree range, starting with nodes with degree of just above 50 (see figure 8 and table 1).

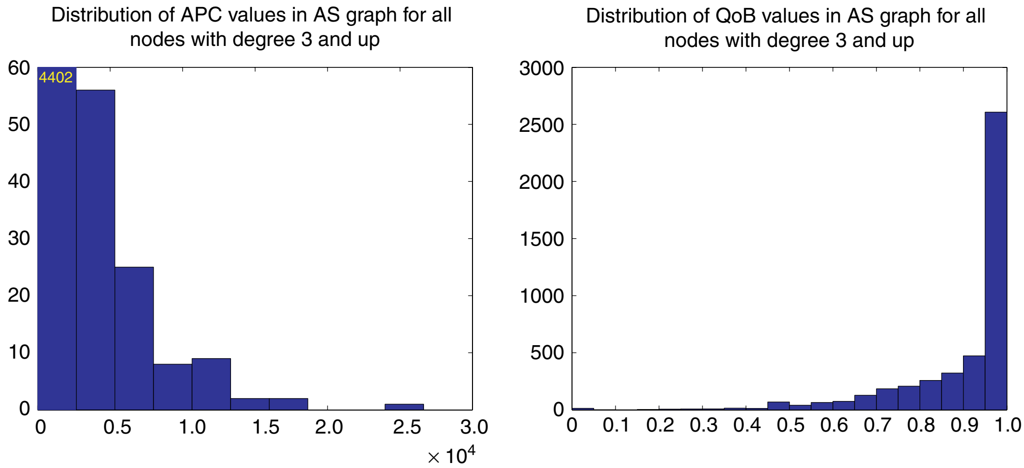


**Figure 9.** APC and QoB distribution in the AS graph. On the left is a histogram of the APC values for nodes of degree greater than 2. The first bin holds 4402 ASes and was truncated. On the right is a histogram of the backup values for nodes of degree greater than 2.

The QoB distribution shown in figure 9 (right) has a large concentration around 1, which is a testament of perfect backup. The median value is 0.9799, and as the histogram shows a large majority of the nodes have QoB values above 0.95.

To discuss our results in comparison to other measures of node significance, we refer to table 1 (bottom) which shows the top ten nodes in the CAIDA ranking [32] based on the number of customers a node has. The list is dominated by high degree nodes; the two medium degree nodes in the list have also rather high APC values; in general all the nodes have relatively high APC values and eight of them are in the top 38 APC list. All the nodes in the list have poor QoB values, possibly due to relatively large stub ASes connecting to them. It is noticeable that the centrality of the nodes in the CAIDA list is much larger than on our APC list. While all the nodes identified as important in the CAIDA list have high APC values, the opposite analogy does not apply. Several of the nodes in our top 10 list are ranked below 200 in the CAIDA list.

## 6. Conclusions

We have shed light on the contribution of backup efficiency for the node significance classification problem. Given our theoretical analysis, we believe this contribution has merit in classification of network nodes in other research areas.

In the Internet AS graph, we are aware that our results are not accurate for several reasons. Firstly, as we stated in the main text, our AS relationship approximation is not accurate. Secondly, although we used the most detailed Internet map available through the DIMES project, the graph itself is still missing many links which can effect the calculation of all the measures, as well as the AS relationship deduction.

In the future, we intend to broaden this research to study the effect of node failure on the point of presence (PoP) level as well as study relationship of sets of nodes in complex networks in the context of backup and functionality. On the theoretical level, we intend to study the robustness of the APC and QoB measures to error in measurements, as well as further formal analysis of their properties.

## References


[1] Albert R, Jeong H and Barabási A-L 2000 Error and attack tolerance of complex networks *Nature* **406** 378–82

[2] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovski D and Alon U 2002 Network motifs: simple building blocks of complex networks *Science* **298** 824–7

[3] Erdös P and Rényi A 1961 On the strength of connectedness of a random graph *Acta Math. Sci. Hung.* **12** 261–7

[4] Bollobàs B 2001 *Random Graphs* 2nd edn (New York: Academic)

[5] Kleinberg J M 1998 Authoritative sources in a hyperlinked environment *SODA* 668–77

[6] Scott J 2000 *Social Network Analysis: A Handbook* 2nd edn (London: Sage)

[7] Dandekar T, Schuster S, Snel B, Huynen M and Bork P 1999 Pathway alignment: application to the comparative analysis of glycolytic enzymes *Biochem. J.* **343** 115–24

[8] Edwards J S and Palsson B O 2000 The *Escherichia coli* MG1655 in silico metabolic penotype: its definition, characteristics and capabilities *Proc. Natl Acad. Sci. USA* **97** 5528–33

[9] Cornish-Bowden A and Cardenas M L 2001 Complex networks of interactions connect genes to phenotypes *Trends Biochem. Sci.* **26** 463–5

[10] Stelling J, Klamt S, Bettenbrock K, Schuster S and Gilles E D 2002 Metabolic network structure determines key aspects of functionality and regulation *Nature* **420** 190–3

[11] Chen Q, Chang H, Govindan R, Jamin S, Shenker S and Willinger W 2002 The origin of power-laws in Internet topologies revisited *IEEE INFOCOM 2002* (*New York, USA*)

[12] Dolev D, Jamin S, Mokryn O and Shavitt Y 2006 Internet resiliency to attacks and failures under BGP policy routing *Comput. Netw.* **50** 3183–96

[13] Luscombe N M, Madan Babu M, Yu H, Snyder M, Teichmann S A and Gerstein M 2004 Genomic analysis of regulatory network dynamics reveals large topological changes *Nature* **431** 308–12

[14] Leskovec J, Kleinberg J M and Faloutsos C 2005 Graphs over time: densification laws, shrinking diameters and possible explanations *KDD* 177–87

[15] Girvan M and Newman M E J 2002 Community structure in social and biological networks *Proc. Natl Acad. Sci. USA* **99** 7821–26

[16] Song C M, Havlin S and Makse H A 2006 Origins of fractality in the growth of complex networks *Nature Phys.* **2** 275–81

[17] Jeong H, Mason S P, Barabási A L and Oltvai Z N 2001 Lethality and centrality in protein networks *Nature* **411** 41–2

[18] Latora V and Marchiori M 2001 Efficient behavior of small-world networks *Phys. Rev. Lett.* **87** 198701

[19] Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E and Dawson S M 2003 The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations—can geographic isolation explain this unique trait? *Behav. Ecol. Sociobiol.* **54** 396–405

[20] Lusseau D and Newman M E J 2004 Identifying the role that animals play in their social networks *Proc. R. Soc. Lond.—Biol. Sci.* **271** (Suppl 6)

[21] Newman M E J 2006 Modularity and community structure in networks *Proc. Natl. Acad. Sci. USA* **103** 8577–82

[22] Zachary W W 1977 An information flow model for conflict and fission in small groups *J. Anthropol. Res.* **33** 452–73

[23] Tauro L, Palmer C, Siganos G and Faloutsos M 2001 A simple conceptual model for the Internet topology *Global Internet* (*November*)

[24] Bar S, Gonen M and Wool A 2004 An incremental super-linear preferential internet topology model *PAM '04 (Antibes Juan-les-Pin, France, April 2004)*

[25] Carmi S, Havlin S, Kirkpatrick S, Shavitt Y and Shir E 2007 A model of Internet topology using *k*-shell decomposition *Proc. Natl Acad. Sci.* **104** 11150–4

[26] Callaway D S, Newman M E J, Strogatz S H and Watts D J 2000 Network robustness and fragility: Percolation on random graphs *Phys. Rev. Lett.* **85** 5468

[27] Cohen R, Erez K, ben Avraham D and Havlin S 2000 Resilience of the Internet to random breakdowns *Phys. Rev. Lett.* **85** 4626–8

[28] Magoni D 2003 Tearing down the internet *IEEE J. Sel. Areas Commun.* **21** 949–60

[29] Park S-T, Khrabrov A, Pennock D M, Lawrence S, Lee Giles C and Ungar L H 2003 Static and dynamic analysis of the internet's susceptibility to faults and attacks *IEEE INFOCOM 2003* (*San-Francisco, CA, USA*)

[30] Gao L 2001 On inferring automonous system relationships in the Internet *IEEE/ACM Trans. Netw.* **9** 733–45

[31] Subramanian L, Agarwal S, Rexford J and Katz R H 2002 Characterizing the Internet hierarchy from multiple vantage points *IEEE INFOCOM 2002* (*New York, USA*)

[32] CAIDA AS Ranking online at http://as-rank.caida.org/

[33] Shavitt Y and Singer Y 2007 A linear time shortest paths algorithm for the Internet AS graph *Technical Report* EE102, Tel Aviv University

[34] Shavitt Y and Shir E 2005 DIMES: Let the internet measure itself *ACM SIGCOMM Comput. Commun. Rev.* **35**

[35] University of Oregon RouteViews Project online at http://www.antc.uoregon.edu/route-views/