ComPlexUs
NETWORK MODELLING
Complexus 2006:3:147–157 DOI: 10.1159/000094196
Published online: August 25, 2006

Lightweight Centrality Measures in Networks under Attack

Giorgos Georgiadis[a] Lefteris Kirousis[a, b]

[a]Department of Computer Engineering and Informatics, University of Patras and
[b]Research Academic Computer Technology Institute, Patras, Greece

Key Words
Scale-free networks, dynamics · Centrality · Clustering coefficient · Network attack

Abstract
In this paper we study deliberate attacks on the infrastructure of large scale-free networks. These attacks are based on the importance of individual vertices in the network in order to be successful, and the concept of centrality (originating from social science) has already been utilized in their study with success. Some measures of centrality however, as betweenness, have disadvantages that do not facilitate the research in this area. We show that with the aid of scale-free network characteristics such as the clustering coefficient we can get results that balance the current centrality measures, but also gain insight into the workings of these networks.

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com
KARGER

1424–8492/06/0033–0147
$23.50/0
Accessible online at:
www.karger.com/cpu

Giorgos Georgiadis
Department of Computer Engineering and Informatics
University of Patras, GR–26504 Patras (Greece)
Tel. +30 2610 996 943, Fax +30 2610 991 909,
E-Mail georgiad@ceid.upatras.gr

Simplexus

Networks are all around us, from the natural networks of human society and those of biological systems to the World Wide Web. The shared characteristics of such large 'scale-free' networks, natural or artificial, is a hierarchical structure in the connection of vertices in the network, which can number in their billions when considering people or web pages. Sociologists, biologists, and computer scientists hoping to understand networks need to find out how vertices are related, what effect changes to the vertices and their connections can have on the network, and how such networks evolve.

Commercial concerns and the security services also have a vested interest in understanding networks. At the functional level, a better understanding of a network should provide alternative ways to index and mine information in the network, while a robust network theory could improve protection from malware-like viruses and worms, hackers, and cyberterrorism.

In this paper, Georgiadis and Kirousis have focused on how deliberate attacks might affect the infrastructure of large scale-free networks. They explain that the success of such attacks is often based on the importance of individual vertices in a network. To understand how a network might be attacked, they considered specific mathematical characteristics of scale-free networks, such as the 'clustering coefficient'. This is a measure of how many connections there are between each vertex and its nearest and next nearest neighbours. Their results not only support earlier studies of the concept of network centrality, adopted from social science, but also provide new insights into how networks function. They explain that knowledge of which vertices in a network to protect is vital to preventing a network from being destroyed.

A network is defined as a system consisting of different objects or entities, whether hormonal glands, people, or web

1 Introduction

Recently there has been an increase of interest in many natural and artificial large-scale networks. For example, it is estimated that the network of web pages currently consists of several billions of vertices [1]. Many companies owning a search engine would like to know the specific characteristics of this network for purposes of page indexing and maybe to predict, up to a point, its future behaviour. In general, a network consists of objects, of different kinds in each area of interest, which are represented by vertices, and connections between them, represented by edges. Many theoretical results exist due to the graph theory of discrete mathematics, which handles such objects. This kind of modelling is possible for a variety of large-scale networks, both naturally occurring and artificial, such as networks of acquaintances, citation, food chains, infections, proteins [2–7] or networks of power grids, internet infrastructure, web pages, and so on [8–14].

Especially in the case of social networks of any kind, they have been studied by scientists in social sciences for quite a while, with methods such as questionnaires and personal interviews. One persistent question was that of the centrality of an individual in such a network or how well 'connected' this person is in his environment. For example, a measure of this connectivity is the degree of a vertex, that is the number of its immediate neighbours. The size of these kinds of networks is in the order of several tens or in some cases several hundreds of vertices and so the research is not directly applicable to the large networks arising today, mainly technological. This is just one of the problems scientists face with the large-scale networks.

Another problem is the nature of these networks. The most prominent model in graph theory until recently has been the random graph model introduced by Erdős and Rényi [15]. In this model, any two vertices have an equal probability to be connected by an edge. This model is very well studied and many results exist on it, but unfortunately it does not describe our observations in real-world networks. In many real-world networks there exists a percent of vertices that appear to be better connected to the rest of the vertices. Furthermore, during the network growth, they gain easier connections and certainly not with equal probability to the other vertices. One of the unique characteristics of these networks, that distinguishes them from previously studied networks, is the power law form of the distribution of the vertex degrees.

In this study we address the problem of network attack. We assume the existence of an adversary that wants to harm a network, by directly attacking and removing the vertices comprising it. He has the ability to measure some variables of the network in order to make educated guesses as to which vertex should be targeted next. In our experiments we measure the efficiency of strategies based on such measures as to the computational time needed to target a vertex on the specific network and the end result after the attack. We care about measures that produce most harm with little effort. The strategies we are using will be based on the centrality measures taken from traditional social network research.

In this paper we take into account previous similar studies and we compare our findings with theirs. We propose an attack strategy that is a trade-off between worst and best strategy so far and has significant and unique advantages. We also offer insight into the workings in power law graphs and indicate future research areas.

The paper is organized as follows. In section 2 we review the fundamental concepts needed in our study, along with a separate discussion of the most widely used measures of centrality. In section 3 we introduce the use of centrality measures as attack strategies. Our experiments can be found in section 4 along with our analysis of the results.

pages. These entities are the network's vertices and the connections between them, blood vessels, communication channels, or hyperlinks, are the 'edges', respectively. For any given network, the number of vertices is essentially only limited by resources, but the complexity and functionality of a network is due not to the number of vertices, but to the connections between them.

The human brain would be nothing more than a grey mass of nerve cells were it not for the billions upon billions of different ways the nerve cells can connect. Similarly, the web would be nothing more than inaccessible cyberjunk were it not for the hyperlink connections between its billions of isolated web pages.

For the social scientist then, it is a relatively straightforward matter to garner information about a network, albeit within the limitations of questionnaires and personal interviews. One perennially important characteristic of social networks concerns the centrality of a person – their connections, in other words. It is, for instance, likely that the professorial chair of a research department will be a more connected individual than a lowly undergraduate researcher. Finding the vertex degrees of all vertices is relatively easy, in all cases. Social science research is not applicable in large scale-free networks, however, because the methods it uses are too time consuming. This fact has no impact on studying social networks with hundreds of vertices, but makes such methods inapplicable to networks, such as the web, which have billions of nodes. At present, there is no way to map such networks, so shortcuts to understanding are essential.

According to Georgiadis and Kirousis, there have been numerous theoretical studies that have attempted to apply the graph theory of discrete mathematics to handle vertices and their associated edges. Graph theory is rather effective in modelling a variety of large scale-free networks, such as networks of family, friends, and acquaintances, food chains, the spread of in-

2 Fundamental Concepts and Definitions

In this section we define the graph-related concepts that we will use in our experiments and analysis, along with the main notions of centrality that are of interest to us.

2.1 Graph-Related Concepts

Throughout this paper we represent a network as an undirected, unweighted graph $G(V, E)$, where V is the set of vertices (i.e. computers) and E is the set of edges (i.e. communication links). Their sizes are $|V| = n$ and $|E| = m$, respectively. The degree k_v of a vertex v is the number of edges originating from or ending in vertex v. We are interested only in graphs generated by the preferential attachment procedure, first proposed by Barabási and Albert [16], which we will briefly describe here. The iterative creation process consists of 4 steps:

Step 0: Initially the graph has n_0 vertices and no edges.

Step 1: Add a new vertex v to the graph.

Step 2: Create l edges, each time connecting the new vertex v to a vertex w, with probability proportional to the degree of this vertex:

$$p_w = \frac{k_w}{\sum_{u \in V} k_u}$$

Step 3: Repeat steps 1 and 2 for $(n - n_0)$ rounds.

The end result of this procedure is a graph of n vertices and $m(n - n_0)$ edges, with vertex degree distribution $P(k)$ that follows a power law, with exponent $\gamma = 3$ ($P(k) \propto k^{-3}$). We call such a graph a Barabási-Albert network or BA network for short. Of course there are many models with creational procedures that generate graphs with power law degree distributions (like Watts 'small worlds' [17, 18]), but we feel that the classical preferential attachment model describes complex network generation in a more general way. Other than that, our specific results may differ in other models but the essence of our insights should still apply.

Other concepts that are used are that of the 1-neighbourhood and 2-neighbourhood of a vertex. Having a vertex v as a centre, its 1-neighbourhood $\Gamma_1(v)$ consists of all vertices at distance 1, i.e. its direct neighbours. Such neighbours will from now on be called first-neighbours and of course it holds that $|\Gamma_1(v)| = k_v$. Similarly, the 2-neighbourhood of a vertex v consists of all vertices at a distance of exactly 2 (from now on second-neighbours) and it holds that

$$|\Gamma_2(v)| \leq \sum_{w \in \Gamma_1(v)} k_w.$$

The inequality in the above expression stands for the fact that some first-neighbours of v may have common neighbours, thus limiting the number of (unique) vertices in $\Gamma_2(v)$. This phenomenon is called *clustering* [13, 19, 20] and is not only possible but characteristic of power law graphs. This relation between a vertex and its first- and second-neighbours leads to the emergence of several structures in the graph, the most common of which is the triangle. In a triangle, three vertices are joined by three edges, one for each pair of vertices. The existence of triangles is characteristic of a power law graph, and it is this feature that makes them so popular in different disciplines: for example, in social science, two of one's friends have a greater probability of knowing each other than two random-picked strangers.

2.2 Standard Centrality Measures

2.2.1 Degree Centrality

The degree centrality measure gives the highest score of influence to the vertex with the largest number of first-neighbours. This agrees with the intuitive way to estimate someone's influence from the size of his immediate environment. The degree centrality is traditionally defined analogous to the degree of a vertex, normalized

fection, the systems of proteins in an organism, global power grids, and the infrastructure of the internet and the web.

This is not the only issue that researchers must address to gain a better understanding of large scale-free networks. There is no reason, after all, why the seemingly badly connected undergraduate could not meet other well-connected people through the Professor's contacts, clubs, family, or indeed any of dozens of possible routes. That said, certain people, presidents, celebrities, even professors, tend to accumulate more connections to other well-connected people at a faster rate than lowly undergraduates. Similarly, the vertices in a technological network are not all created equal. Of those billions of web pages, there are perhaps a few thousand, if not just a few hundred that attract the most visitors because there are many, many edges pointing to them from countless other pages. An undergraduate's personal resume page is very unlikely to become as big a hub of interest as the Professor's departmental page. The student's friends might link to the resume page, but there will be far more students, university directory, external research sites and organizations that link to a Professor's page.

On an even larger scale, web sites such Amazon and Google, and big league university sites inevitably have many edges connected to countless pages and so are key hubs. Georgiadis and Kirousis emphasize that this clustering of edges around a small number of hubs is common to many real-world networks. The effect is magnified during network growth as the 'hubs' attract more edges simply because of their prominence, whether that is deserved or not. The same can be said of professors, of course, and other biological systems; consider the tree-like structure of blood vessels, neural pathways, and indeed trees.

In order to best protect such networks from physical or 'cyber'-attack, we need a clearer model and Georgiadis and Kirousis hope to provide just such a solution. They

over the maximum number of neighbours this vertex could have. Thus, in a network of n vertices, the degree centrality of vertex i, C_i^D, is defined as:

$$C_i^D = \frac{k_i}{n-1}$$

The normalization in the region [0, 1] is used here to make the centrality of different vertices comparable, and also independent of the size of the network.

2.2.2 Closeness Centrality

This notion of centrality focuses on the idea of communication between different vertices. The vertex which is 'closer' to all vertices gets the highest score. In effect, this measure indicates which one of two vertices needs fewer steps in order to communicate with some other vertex. Because this measure is defined as 'closeness', the inverse of the mean distance of a vertex from all others is used. Hence, if C_i^C is the closeness centrality, and d_{ij} the shortest distance between vertices i and j in terms of edge steps:

$$C_i^C = \frac{n-1}{\sum_{j \in V} d_{ij}}$$

Again, this measure is normalized in the region [0, 1]. Additionally, it should be stated that the distance between two disconnected vertices must be a predefined very large value and not infinite, if it is desirable to discern among low closeness scores.

2.2.3 Betweenness Centrality

Betweenness centrality refines the concept of communication, introduced in closeness centrality. Informally, betweenness centrality of a vertex can be defined as the percent of shortest paths connecting any two vertices that pass through that vertex. The normalized version divides this value with the maximum possible betweenness centrality, that is all possible shortest paths in a completely connected graph. If C_i^B is the betweenness centrality of vertex i, (u, i, v) is the set of all shortest paths between vertices u and v passing through vertex i and (u, v) is the set of all shortest paths between vertices u and v, then:

$$C_i^B = \frac{\sum_{u \in V} \sum_{v \neq u \in V} \frac{|(u,i,v)|}{|(u,v)|}}{(n-1)(n-2)}$$

This definition of centrality explores the ability of a vertex to be 'irreplaceable' in the communications of two random vertices. It is of particular interest in the study of network attacks, because at any given time the removal of the maximum betweenness vertex seems to cause maximum damage in terms of connectivity and mean distance in the network. Its main disadvantage is that the summation operator practically means that it needs global information about the network, in order to compute the betweenness of a single vertex, and that is simply not possible in many contexts. For the same reason it is expensive in computing time to compute the score of a vertex, although this disadvantage was significantly improved recently [21, 22]. The importance of betweenness centrality as an attack strategy is further discussed below.

3 Centrality Measures as Attack Strategies

It has been shown in the past that the 'random vertex hit' strategy performs poorly [23, 24], due to the hierarchical effect these networks present, i.e. a random vertex has an increased probability to be one of the less connected vertices, since there are so many of them. So it is desirable to use a strategy that achieves better results, and such strategies could be based upon a vertex measure that can profile the potential of each vertex by its value only. A number of publications exist [23–27], addressing the question of which strategy is best in achieving maximum destructive results with less vertex hits, the most extensive of which, to our knowledge, is that by Holme et al. [23]. Summarizing the remake the basic assumption that an attacker wishing to compromise a network would focus on first disabling the vertices. The aim would be to cause the most harm with the least effort, so the researchers consider two variables: the computational time needed to target a vertex and the end result of the attack. The net result of their study, in the best tradition of 'forewarned is forearmed', reveals an attack strategy that is a trade-off between the worst and best strategies seen so far. This strategy, however, has what they describe as significant and unique advantages.

The researchers looked at three novel strategies in network attack and compared them with two traditional approaches, degree and betweeness centrality. 'Betweeness' measures the centrality of a vertex very efficiently, but relies on knowing all the network's shortest paths. This is practically impossible to determine for a network such as the web. In contrast, the strategies proposed by Georgiadis and Kirousis require only local information, i.e. the 1- or 2-neighbourhood (the external/internal links of a web page), which takes far less time and effort to compute.

The search for an attack strategy fundamentally involves finding those vertices that are most important to the network and without which the network might cease to function as an holistic system. A random attack has previously been shown to have little impact on a large scale-free network, as one might expect. A research department will not close because a random student or staff member is absent. Moreover, with a random attack there is more chance of each strike hitting a lesser vertex simply because there are more of them.

Instead, attackers could use different ways to measure the 'importance' of a vertex based on its connectedness and exploit this in a more destructive approach. The researchers suggest that a different, yet equally or even more destructive, deletion of vertices can be made, in less time with less information to hand. Target the Profes-

sults, the comparison is based upon two axes: different strategies and recalculation of measures. The different strategies studied are vertex deletion based upon degree centrality scores and upon betweenness centrality scores, and it is clearly shown that betweenness produces better results. The recalculation of the involved measures refers to the recalculation of degree/betweenness centrality after each vertex deletion.

In this section we introduce a strategy that balances the advantages and disadvantages of the above-mentioned strategies. We will not study the closeness centrality, as it has the same basic flaws as betweenness and none of its advantages. Furthermore, we focus on the recalculated versions, since the distribution of these measures may vary significantly between deletions.

3.1 How to Measure Destructive Power?

We are interested in the destruction of the network under consideration. Ideally that would mean the isolation of each vertex, but it can be argued that it is enough to break the network into a sufficiently large number of connected components. We chose to examine only the size of the largest component, as a particularly small largest component would mean that the network has degenerated into many small connected components. Additionally, we can measure directly the impact of vertex deletions in the hierarchical structure by examining what happens to the largest component: a successful attack would probably target this component and shrink its size dramatically. Another reason is that this technique has been used successfully during previous studies [23], and its use will make our results directly comparable. We specifically use a normalization over the largest component size with the initial network size, in order to produce a percentage comparable between different size networks.

Since we start with a connected network, it would take some time before it becomes disconnected, and during that time the size of the largest component would not carry significant information. Thus, in addition to the largest component size, we use the mean shortest path length of the network, and specifically its mean inverse. The mean shortest path length is the mean length of all shortest paths in the network, between all pairs of vertices. If by d_{uv} we denote the length of the shortest path between vertices u and v, then the mean shortest path length l, in a network of n vertices, is

$$l = \frac{\sum_{u \in V} \sum_{v \neq u \in V} d_{uv}}{n\,(n-1)}.$$

The mean inverse of shortest path length l^{-1} is defined as

$$l^{-1} = \frac{\sum_{u \in V} \sum_{v \neq u \in V} \frac{1}{d_{uv}}}{n\,(n-1)}$$

In practice we use the mean inverse of shortest path length because by doing that we nullify the effects of disconnected vertices and their 'infinite' distance. An increasing mean value of this measure means that average distances in the network are increasing, and this subsequently means that the attack in the network produces quantifiable, destructive results. Clearly, since we use the mean inverse of this measure, we expect it to decrease with time.

3.2 Standard Centrality Measures Explained

As already mentioned, the random vertex hit strategy has practically no effect on the network's integrity, and that is because it cannot take into consideration its hierarchical structure. This is exactly where a degree-based attack succeeds. By targeting the highest-degree vertices first, it attacks directly the global network connectivity. It must be pointed out that not all properties of Barabási-Albert networks are known.

sor's office and personal assistant though and the departmental network could be crippled.

Their first strategy uses a measure of vertex importance known as 'the edge degree'. A formal definition of edge degree does not yet exist. So the researchers have produced their own formula which defines it as the product of the number of edges at two connected vertexes. If vertex V has 30 links to other vertices and is connected to vertex Y which has 25, then the edge degree is 750. This strategy then chooses to attack vertices with the greater edge degree based on the immediate neighbourhood and is referred to as the '1-neighbourhood edge degree' strategy.

The second attack strategy defines edge degree as the product of edges within two 'neighbourhoods'. So, the approach involves attacking the edge with highest degree and then the endpoint vertex with maximum degree.

Finally, the third most sophisticated strategy uses the 2-neighbourhood edge but penalizes 'triangles' of connectivity, if vertex V is connected to Y and Y is connected to Z, which in turn is connected to Z, and then this reduces the score. To compensate for their presence the 2-neighbourhood edge degree is divided by the number of triangles plus one.

The researchers then applied their strategies to experimental networks with 1,500 vertices and found that each of the three emerged as easy to implement, low in computational cost, but efficient in causing harm to the network compared with earlier strategies.

Thankfully, a potentially devastating attack within the web is not easy, and would probably not have much impact. On the other hand, attacking the physical infrastructure of the internet might be used in times of war to isolate a specific country of the informational highway, or by terrorists to wreak havoc.

David Bradley of Sciencebase.com

Initially it was believed that high-degree vertices were connected with other high-degree vertices preferably over lower-degree vertices (assortative mixing) [28]. Recent studies [29] show that Barabási-Albert networks are rather neutral on this property, and in some cases even show the opposite behaviour (disassortative mixing), i.e. high-degree vertices prefer lower-degree vertices to connect to. We believe this observation can explain the success, albeit partial, of this strategy, as in the disassortative mixing the deletion of the highest-degree vertex would affect many vertices: but since this is not a predominant phenomenon the effectiveness of this strategy would be limited.

On the other hand, the betweenness-based strategy seems ideal, especially with the performance metrics used (mean shortest path length, largest component size). By definition, the betweenness measures the ability of a vertex to be irreplaceable in shortest paths throughout the network. So when this vertex is removed, inevitably all shortest paths that depended on it will also be removed, and equally long or longer paths would take their place. This has an obvious impact in the mean shortest path length, which is constantly non-decreasing, at least as long as a unique giant component exists. Such high betweenness vertices, which connect many others with shortest paths, would be initially located in the largest component as most vertices would be located there. Therefore, the failure of these paths also affects the largest component size, since multiple failures may produce disconnected vertices. Similar arguments can be used with the closeness centrality.

3.3 Proposed Strategies

We propose a family of strategies based, in part, on edge degree. Although a formal definition of edge degree does not exist, we experimented with several possible definitions, all based on vertex degree. Specifically, an edge's degree has some connection with the endpoint vertices of this edge. As was the case in Holme et al. [23] we settled with the edge degree being the product of the degrees of the endpoint vertices, as it followed closely our intuition on the importance of edges. If $e = (w, u)$, an edge with endpoints w and u, having degrees k_w and k_u, respectively, its edge degree $k_e^{\Gamma_1}$ is defined as:

$$k_e^{\Gamma_1} = k_w \cdot k_u$$

The first strategy which uses the edge degree to select vertices does so by first selecting the edge with maximum degree, and then the vertex of this edge with maximum (vertex) degree. In case of multiple edges/vertices with the same (maximum) degree, we chose uniformly at random. Note that this strategy examines the immediate neighbourhood of each endpoint vertex, and scores higher edges having endpoints with large 1-neighbourhoods. From now on we will refer to this strategy as '1-neighbourhood edge degree' strategy.

The second strategy defines edge degree as the product of the 2-neighbourhoods of its endpoint vertices. This 2-neighbourhood edge degree $k_e^{\Gamma_2}$ of an edge $e = (w, u)$ is defined formally as:

$$k_e^{\Gamma_2} = \sum_{i \in \Gamma_1(w)} k_i \cdot \sum_{j \in \Gamma_1(u)} k_j$$

The vertex selection is exactly the same as before: choose the edge with maximum degree and then the endpoint vertex with maximum degree. We will refer to this strategy as '2-neighbourhood edge degree' strategy.

The third proposed strategy is based on the '2-neighbourhood edge degree', as defined above. The main difference is that it penalizes the existence of triangles in which the edge is present. Specifically, it divides the above computed edge degree by the number of triangles that this edge participates in plus one to avoid division by zero. Thus, if T is the number of triangles involving the edge in question as a side of the triangle, the formal definition of the alternative edge degree is:

$$K_e^{\Gamma_2} = \frac{\sum_{i \in \Gamma_2(w)} k_i \cdot \sum_{j \in \Gamma_2(u)} k_j}{T + 1}$$

We refer to this strategy as '2-neighbourhood edge degree with penalty'.

4 Experiments

For the experiments we used networks of 1,500 vertices, created by the BA procedure mentioned in section 2. The parameters of importance are the size of the initial network (before the procedure starts adding vertices) and the degree of each added vertex. We used degree 5 for each new vertex and we kept the initial network small, consisting of 5 vertices connected with random edges. Each edge between two vertices had 0.5 probability of existing, so as to differentiate the vertices for the growing procedure. We intentionally kept the initial network small because larger (initial) networks create larger gaps between high-degree and low-degree vertices during the network growth. As a result, highly central vertices are fewer and more easily recognizable by any targeting strategy and are diminished quickly, leaving no time for the various strategies to produce different results.

The results are shown in figures 1–4. The inverse mean shortest path length, the size of the largest component and the clustering coefficient are measured after each vertex deletion and shown in separate illustrations. For each of these parameters, five different data sets exist, corresponding to the five strategies under study (highest degree, betweenness, 1-neighbourhood edge degree, 2-neighbourhood edge degree and 2-neighbourhood edge degree with penalty). Their values at each deletion step are the average of 50 experiments with different networks of 1,500 vertices. Of the monitoring parameters,

the easiest to read is the size of the largest component and its transition is shown in magnification in figure 3. It is easy to see the relation between the various strategies, as each one, having done preliminary work, performs better or worse than the others during the transition.

4.1 Drilling into Experimental Results

The inverse mean shortest path length initially declines, meaning that distances inside the network begin to grow in general. At some point this trend is reversed because during the deletion process the connected components become quite small and the distances inside them are even smaller than in the initial network. Thus, the inverse length continues to increase as connected components are cut into smaller pieces and this continues until they stop breaking up. At this point the inverse length is at its maximum value and almost all significant vertices are gone, as subsequent deletions leave the components at roughly the same size. From this maximum point on, the network continues to shrink at an almost constant rate. The five strategies differ mainly in their ability to break the already small connected components into even smaller ones, leading to higher maximum points, as shown in figure 1.

The size of the largest component is a more straightforward measure. After each deletion the size of a component is reduced by one and, at least initially, the deleted vertex is selected from the largest component. As more 'central' vertices are deleted, critical paths collapse and the largest component breaks into smaller pieces. Figure 2 shows clearly that there is an early stage where the strategies built up tensions by deleting important vertices, a transition phase where very important vertices are gone and each deletion breaks the largest component into small pieces, followed by a slow shrinking of the largest component. The transition phase is where the various

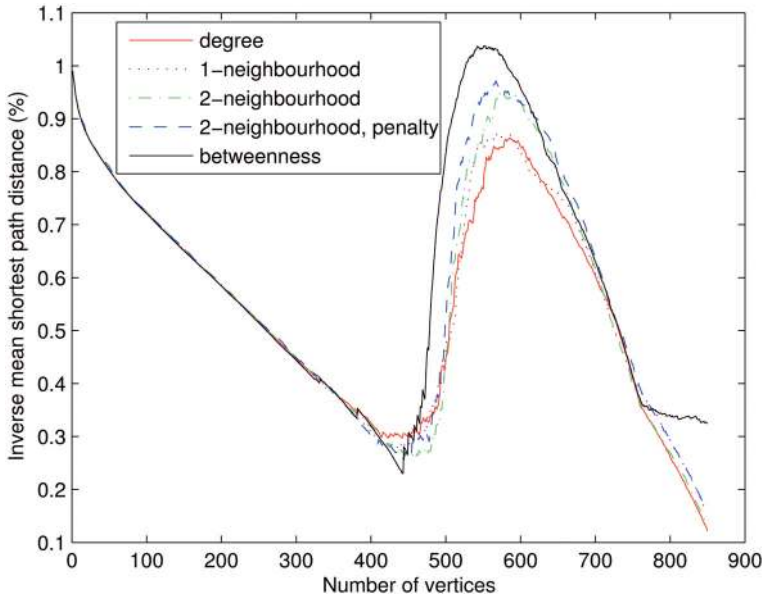

Fig. 1. Inverse mean shortest path length shown as percent of the initial length, as vertices are sequentially deleted. The results are the average of 50 experiments with networks of 1,500 vertices.

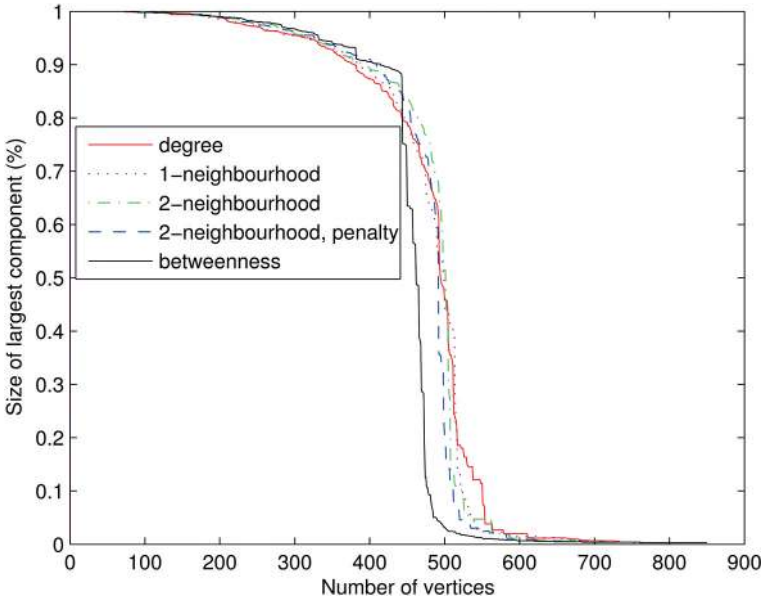

Fig. 2. Size of largest connected component shown as percent of the initial size, as vertices are sequentially deleted. The results are the average of 50 experiments with networks of 1,500 vertices

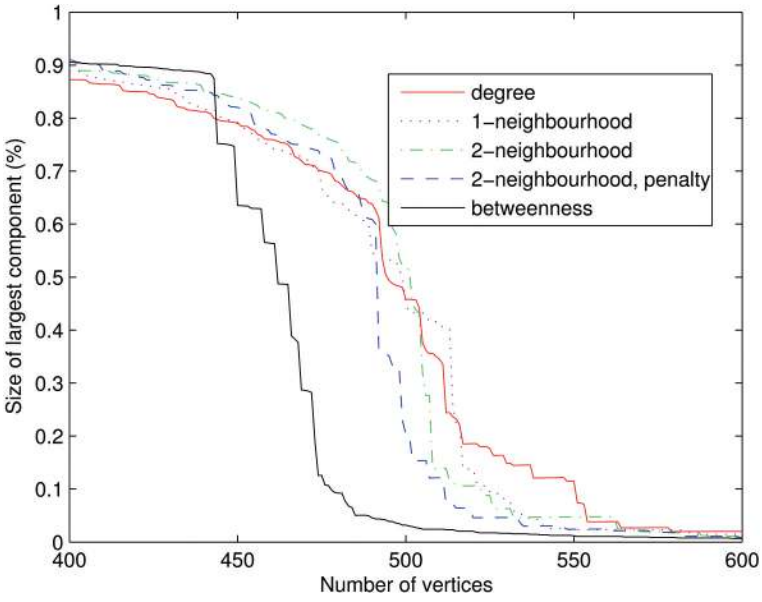

Fig. 3. Size of largest connected component shown as percent of the initial size, as vertices are sequentially deleted. Detail of the transition.

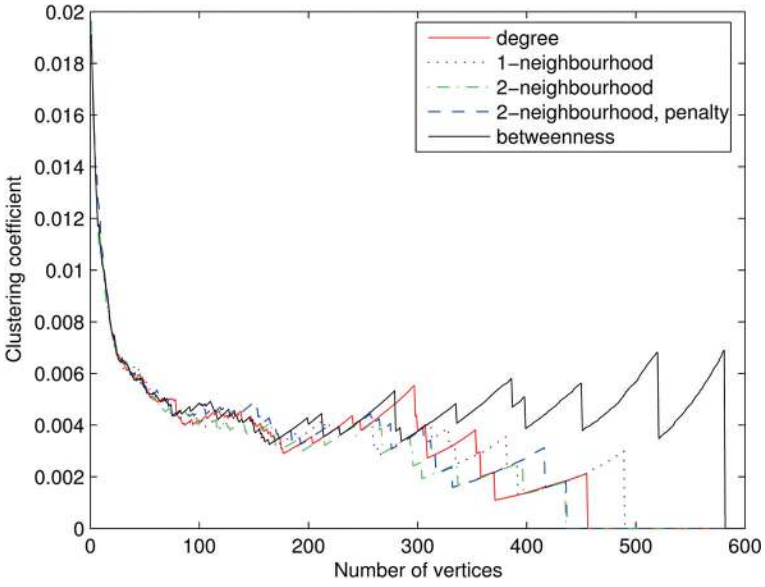

Fig. 4. Clustering coefficient of network, as vertices are sequentially deleted. The results are the average of 50 experiments with networks of 1,500 vertices.

strategies compete, and betweenness is most successful in making the transition in fewer deletions. However, comparable to the degree strategy which performs poorly, our proposed strategies bridge the gap with betweenness by up to 23, 29 and 55% for 1-neighbourhood, 2-neighbourhood and 2-neighbourhood with penalty edge degree, respectively.

The clustering coefficient during the early stage is decreasing by orders of magnitude, meaning that the deleted vertices, tagged as central by the various strategies, contribute greatly to the global clustering coefficient (fig. 4). During the transition phase it appears to be fluctuating due to the shrinking of the largest component and the increase in the number of components, and in the last phase it is completely wiped out as triangles do practically not exist. The betweenness stands out, since during the transition phase it creates a seesaw effect on the clustering coefficient, never destroying all triangles in the connected components. Although we have no solid evidence, we feel that this observation is the key to understanding the role and the success of betweenness, and to replicate its behaviour in other measures.

In order to understand why the proposed strategies work as they do, we focus on a high-degree edge and examine its specific characteristics (fig. 5). Just by looking at the high-degree edge alone, one can argue that it connects high-degree vertices; therefore, it is important for the communication of $(k_w - 1)$ vertices (at the one endpoint) with another $(k_u - 1)$ vertices (at the other endpoint). So its deletion alone would probably affect many vertices and the distances between them. As for the highest-degree endpoint (which will eventually be deleted), one must keep in mind that high-degree vertices do not usually connect to other high-degree vertices. On one hand, deleting high-degree vertices is a strategy successful enough on its own (see degree centrality strategy), but with

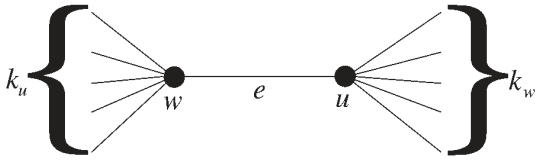

Fig. 5. 1-neighbourhoods of two connected vertices.

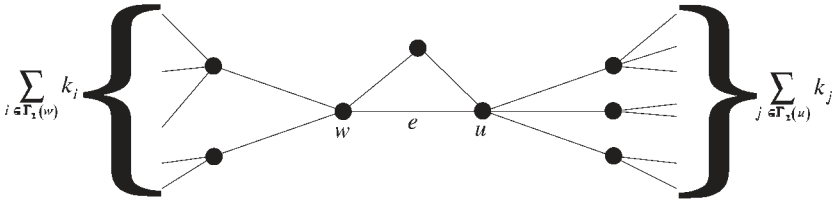

Fig. 6. 2-neighbourhoods of two connected vertices.

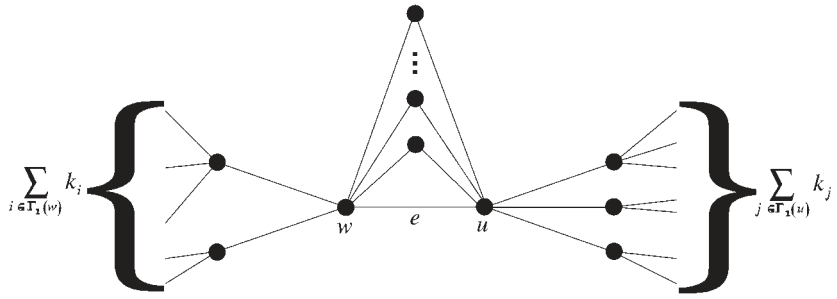

Fig. 7. 2-neighbourhoods of two connected vertices with triangles.

maximum edge degree we ensure that the high-degree vertex to be deleted will be connected to the highest possible degree vertex (given it is not a common phenomenon) and the deletion will affect a greater number of vertices. On the other hand, since high-degree vertices do not connect often, this filter adequately differentiates otherwise equal vertices (i.e. when degree centrality is used).

The 2-neighbourhood edge degree strategy operates in a similar way. The same arguments as above are still valid here, i.e. a high-degree edge connects more vertices than a low-degree one. The main difference is that we are now talking about vertices that are part of the 2-neighbourhood of the one endpoint vertex which connect with the vertices of the 2-neighbourhood of the other endpoint vertex. This may be more reliable than the 1-neighbourhood of the endpoints since a light disassortative mixing seems to exist. This means that high-degree vertices connect to lower-degree ones, thus their influence dies out quickly as we move further from their centre. By using 2-neighbourhoods we favour vertices whose influence two steps away is still strong. The downside of this strategy is its slightly larger computational load compared to the 1-neighbourhood edge degree, but this is still far from that of betweenness centrality. Furthermore, it uses semilocal information for its computation, which, we estimate, should not be a problem in most practical uses.

The mechanism behind the alternative 2-neighbourhood edge degree strategy is somewhat different. Obviously the same arguments of the two previous strategies are still valid here. An instance that is handled differently is shown in figure 6. Normally the 2-neighbourhood edge degree of this edge would be

$$\sum_{i \in \Gamma_2(w)} k_i \cdot \sum_{j \in \Gamma_2(u)} k_j$$

but since it participates in a triangle due to the common neighbour of both endpoint vertices, this edge degree is divided by 2. Thus edges that connect two 'smaller' vertices in terms of 2-neighbourhoods can have a larger edge degree and be selected instead. The situation is even worse if the edge participates in more triangles, as in figure 7, for its edge degree would be even smaller. The edge degree can gradually increase, if vertices comprising the triangles get selected for the deletion process, and the triangles collapse.

This edge degree with penalty measures the size of two 2-neighbourhood connecting through one edge, as was the case in the previous 2-neighbourhood edge degree. But it also considers the importance of alternative paths between these two 2-neighbourhoods. There is no doubt that selecting one endpoint of an edge participating in many triangles will also destroy these triangles, but vertices connecting same size 2-neighbourhoods with no triangles are more important to the whole network, and this is expressed by this measure and verified by our results.

4.2 Algorithm Complexities

Theorem 1. The worst case time complexities of the proposed strategies are $O(m)$, $O(m\sqrt{n})$ *and* $O(mn)$ for 1-neighbourhood, 2-neighbourhood and 2-neighbourhood with penalty edge degrees, respectively. Furthermore the aver-

age case time complexity is $O(m)$ for all strategies.

Proof. The 1-neighbourhood edge degree just multiplies two integers, namely the vertex degrees of the edge endpoints, for all edges. The query of the degree of a vertice is an $O(1)$ operation in the LEDA environment we are using [30], so the total cost of computing 1-neighbourhood edge degree for all edges is $O(m)$ in any case.

The 2-neighbourhood edge degree queries for each endpoint vertex, the vertex degree of all its neighbours and sums it up, multiplying the two endpoint sums, and does this for all edges. The iteration of all neighbouring vertices of a vertex has guaranteed asymptotic complexity on the number of actual neighbours. So in the worst case, this computation has $O(mk_{max})$ complexity, where k_{max} is the maximum degree in the network. Specifically for the Barabási-Albert network there exists an analytic solution [31] for the degree $k_i(t)$ of a vertex i at time step t, as

$$k_i(t) = l\sqrt{\frac{t}{t_i}} \qquad (4)$$

where l is the number of edges per new vertex and t_i is the time step when vertex i was added to the network. After n time steps the maximum degree in the network is

$$k_{max} = O\left(\sqrt{n}\right)$$

And so the worst case time complexity is

$$O\left(m\sqrt{n}\right).$$

But since the mean degree in the network is $\overline{k} = 2l$ (as can easily be seen), the average case time complexity is $O(m)$.

The third strategy is computed as above, but for one endpoint of the edge, we scan its neighbour's neighbour lists to find the other endpoint (indicating the existence of triangles). So its worst case complexity is $O(mk_{max}{}^2)$ and thus $O(mn)$. Similarly its average case complexity is $O(m)$.

5 Conclusions

We have studied three novel strategies in network attack and compared them with two traditional approaches, degree and betweeness centrality, both with its own merits and flaws. These strategies have proven to be simple enough to implement, with low computational cost, and yet efficient compared to the best strategy. In addition to their value as attack strategies, they can help to shed light on the inner workings of a power law network. One of the great difficulties in their study is our ignorance as to what measures are important to the behaviour of these networks. Our experiments link the degree-degree correlations among vertices with their centrality in the network. Furthermore, to the extent of our knowledge, it is the first time that the clustering effect is linked to the centrality of a vertex. Although we know this is responsible for the 'denseness' of power law networks, its exact role remains unclear. Our third strategy indicates that it plays a major role in conjunction with other phenomena, such as the degree-degree correlations. It would be of interest to study several models of networks, other than the BA model, that show documented assortative or disassortative behaviour and models that have known clustering coefficient distributions, in order to explore further these effects of our strategies. Furthermore, it is the subject of future research to determine whether the utilization of other network structures, similar to the triangles we are using in this study, will help bridge the gap between local strategies and global ones, as is betweenness. This development will not only help us to study larger networks but will also reveal the role of individuals in such a vast network.

Acknowledgement

This research was partially supported by the EU within the 6th Framework Programme under contract 001907 'Dynamically Evolving, Large Scale Information Systems' (DELIS) and was also partially supported by the European Social Fund (ESF), Operational Program for Educational and Vocational Training II (EPEAEK II), and particularly the Program *Pythagoras*.

References

1 Baldi P, Frasconi P, Smyth P: Modeling the Internet and the Web: Probabilistic Methods and Algorithms. Chichester, Wiley, 2003.

2 Jeong H, Tombor B, Albert R, Oltvai N, Barabási A: The large-scale organization of metabolic networks. Nature 2000; 407: 651–654.

3 Liljeros F, Edling R, Amaral N, Stanley E, Aberg Y: The web of human sexual contacts. Nature 2001; 411: 907–908.

4 Mariolis P: Interlocking directorates and control of corporations: the theory of bank control. Soc Sci Q 1975; 56: 425–439.

5 Pimm L: Food Webs, ed 2. Chicago, University of Chicago Press, 2002.

6 Podani J, Oltvai N, Jeong H, Tombor B, Barabási A, Szathmáry E: Comparable system-level organization of archaea and eukaryotes. Nat Genet 2001; 29: 54–56.

7 Jones J, Handcock M: An assessment of preferential attachment as a mechanism for human sexual network formation. Proc R Soc Lond B Biol Sci 2003; 270: 1123–1128.

8 Albert R, Jeong H, Barabási A: Diameter of the world wide web. Nature 1999; 401: 130–131.

9 Chen Q, Chang H, Govindan R, Jamin S, Shenker S, Willinge W: The origin of power laws in internet topologies revisited. IEEE Infocom 2002, 2002.

10 Faloutsos M, Faloutsos P, Faloutsos C: On power-law relationships of the internet topology. SIGCOMM, 1999, pp 251–262.

11 Huberman A: The Laws of the Web: Patterns in the Ecology of Information. Cambridge, MIT Press, 2001.

12 Lawrence S, Giles L: Searching the world wide Web. Science 1998; 280: 98–100.

13 Ravasz E, Barabási A: Hierarchical organization in complex networks. Phys Rev E Stat Nonlin Soft Matter Phys 2003; 67: 026112.

14 Watts J: A simple model of global cascades on random networks. Proc Natl Acad Sci USA 2002; 99: £5766–5771.

15 Erdős P, Rényi P: On random graphs. Publ Math Debrecen 1959; 6: 290–291.

16 Barabási A, Albert R: Emergence of scaling in random networks. Science 1999; 286: 509–512.

17 Watts J, Strogatz H: Collective dynamics of 'small-world' networks. Nature 1998; 393: 440–442.

18 Watts J: Small Worlds. Princeton, Princeton University Press, 1999.

19 Newman J: The structure and function of complex networks. SIAM Rev 2003; 45: 167–256.

20 Bornholdt S, Schuster G: Handbook of Graphs and Networks. Berlin, Wiley-VCH, 2002.
21 Brandes U: A faster algorithm for betweenness centrality. J Math Sociol 2001; 25: 163–177.
22 Newman J: Scientific collaboration networks. ii. Shortest paths, weighted networks, and centrality. Phys Rev E 2001; 64: 016132.
23 Holme P, Kim J, Yoon No, Han K: Attack vulnerability of complex networks. Phys Rev E 2002; 65: 056109.
24 Cohen R, Erez K, Ben-Avraham D, Havlin S: Resilience of the internet to random breakdowns. Phys Rev Lett 2000; 85: 4626–4628.
25 Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J: Graph structure in the web. Comput Networks 2000; 33: 309.
26 Callaway S, Newman J, Strogatz H, Watts J: Network robustness and fragility: percolation on random graphs. Phys Rev Lett 2000; 85: 5468–5471.
27 Cohen R, Erez K, Ben-Avraham D, Havlin S: Breakdown of the internet under intentional attack. Phys Rev Lett 2001; 86: 3682–3685.
28 Krapivsky L, Redner S: Organization of growing random networks. Phys Rev E 2001; 63: 066123.
29 Zhuang-Xiong H, Xin-Ran W, Han Z: Pair correlations in scale-free networks. Chin Phys 2004; 13: 273–278.
30 Mehlhorn K, Nahe S: LEDA: a Platform for Combinatorial and Geometric Computing. Cambridge, Cambridge University Press, 2000.
31 Barabási A, Albert R, Jeong H: Mean-field theory for scale-free random networks. Physica A 1999; 272: 173–187.