

## DESIGN THROUGH FAILURE: A NETWORK PERSPECTIVE

C. Ellinas, M.J. Hall, A. Hultin

*Keywords: sociotechnical systems, networks, complexity, cyber-security, risk*

### 1. Introduction

Design is one of the principal processes within the domain of engineering; bettering the former will inevitable improve the latter. Designing *complex* systems is one of the most challenging tasks of our time. The need to deliver an ever-increasing number of such systems, along with increasingly challenging constraints (demands for increased functionality, limited resources, increased number of divergent stakeholders etc.) makes this a matter of great importance. As engineers, we need to understand the failure modes of such systems in order to increase our confidence in designing systems that work more often than not. Recent mega-failures (ranging from the Fukushima nuclear reactor to financial fraud examples such as the “London whale” incident) continuously serve as reminders that currently, our capacity to deliver complex systems in a sustainable manner is still in its infancy.

Organisations are increasingly reliant upon progressively more elaborate processes in order to operate. Such processes are delivered by *sociotechnical* systems, relying on interactions between social and technological (or generally, engineered) systems rather than individual instances of either. Their composing elements include social agents at various aggregated levels (i.e. people, teams, agencies, etc.), along with IT systems of varying capabilities (i.e. from PCs to servers), united by policies and various business processes. Within this context, the UK Ministry of Defence (MoD) has recognised that a key challenge lies in the process of “understanding linkages and *dependencies* between people ... and equipment” [DSTL 2012]. Such dependencies are far from being random – they are usually driven by built-in redundancies and/or required functional dependencies. Their effectiveness will inevitably lead to increased capabilities and efficient resource expenditure, resulting in sustainable systems and leading organisations towards achieving a competitive advantage.

Reducing operational risk is a necessary condition for doing so. The underlying motivation is fundamentally economic – the amount of resource put into designing any such system (i.e. capital cost, lifecycle costs, upgradability, etc.) will be proportional to the expected overall effect in terms of overall operational capacity. Situational Awareness (SA) is a key concept in maximising this utility function. To further elaborate, SA is a state whereby an organisation identifies and understands incidents or events, before assessing their likely impact on operations [Endsley 1995]. In the context of cyber-security, SA can be understood in terms of achieving a number of objectives, namely: (1) Be aware of the current situation; (2) Be aware of the impact of the attack; (3) Be aware of how the situations evolve; (4) Be aware of actor behaviour; (5) Be aware of why and how the current situation is caused; (6) Be aware of the quality and trustworthiness of the information and intelligence; (7) Assess plausible futures of the current situation [Bardord et al 2010].

It is important to appreciate that it is rather improbable (and most definitely, uneconomical) to design a fail proof system. Thus, in order to explore *resilience* in the context of sociotechnical systems, a suitably abstracted model will be constructed by utilising concepts from network science – the latter being a multidisciplinary domain that focuses on the nature of interconnectivity between discrete entities rather than their analysis. By concentrating on the first 3 objectives that define SA, artificially generated (but suitable tailored) systems will be perturbed under two extreme attack modes (i.e. no threshold of containment and thus, no capability to be restrained) in order to identify the effect of local failure at a global level along with its progression, specifically looking for early-warning signs. This was achieved through a number of numerical simulations, using MATLAB, in order to provide quantitative evidence on the fundamental mode of failure. Implications in terms of the design process of such a system will then be presented through the derived insight to improve a specific aspect of the design process – the trade-off between resilience and capacity to operate efficiently.

## 2. Model

### 2.1. Improving design by understanding failure

[Arthur 2009] suggests that engineering design is a fundamental process upon which means (let it be an artefact, process or methodology) are devised in order to utilise captured phenomena (uncovered by science) in order to perform a purpose. From an economic perspective, we also propose that a good design is one that maintains maximum functionality in a sustainable way with minimum cost. In order to do so, relevant agents need to be able to understand the potential impact of a local failure on a global scale (i.e. the entire system) the functionality of the entire system. Consequently, the purpose of this paper is to provide some insight on this aspect by providing quantitative evidence on the reaction of different system architectures on a worst-case scenario basis. Two useful measures will then be proposed to serve as proxies for monitoring the state of each system, enabling decision makers to educatedly rank local failures in terms of the potential influence on the global system and thus, improve mitigation resource allocation. We would like to emphasise the fact that improved design, within the domain of engineering, has traditionally taken place by better understanding failure mechanisms and designing accordingly (e.g. the design procedure of any structural elements begins by first understanding the failure profile of its composing material under loading, e.g. will it fail in a ductile or a brittle way? Improved resource expenditure can thus be achieved in terms of material volume used, life-cycle costs, etc.). Such will be the adopted ethos of this paper.

### 2.2. The means to an abstraction

The *de facto* starting point of this approach is that the aim (and thus, function) of the overall system is to achieve a set number of organisational goals, which is directly linked to its capacity to undertake a set number of tasks. The latter are effectively designed outcomes of a sequence of interactions within a social system (or social *network*, hereafter S.N.) such as communication between social agents; within a technological system (similarly, a technological network, hereafter T.N.) such as information exchanges between servers; but also interaction between the two in a compounded sociotechnical system (i.e. a *network-of-networks*, hereafter NoN) such as people using PCs.

Sequentially, the functionality of the sociotechnical system (i.e. the organisation) depends on the health of the underlying T.N. as it is operated, monitored and regulated by a number of social agents (i.e. the S.N.). The essence of such abstraction lies on the fact that the capacity of the sociotechnical to fulfil its purpose (i.e. deliver the envisioned organisational goals per specified requirements) heavily relies on the ability of the T.N. and S.N. to perform their respective individual functions, i.e. their allocated tasks. The latter is only achieved if the health of the discrete entities in both T.N. and S.N. are healthy (i.e. a reductionist view), but also if the connection *within* and *between* the two systems are functioning (i.e. a holistic view). Within this paper, the latter will be adopted and specifically focus on the interactions *between* the two systems since, as the complexity and thus, resource expenditure, of such systems increases, failures tend to take place on the interfaces rather on the individual systems.

## 2.3. Constructing the model

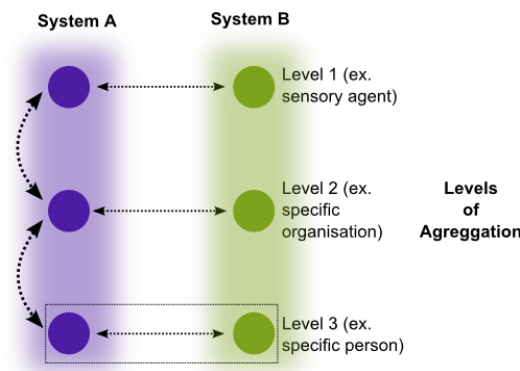
In every kind of modelling effort, two approaches can be used with respect to the data used; either utilise empirical data (prone to statistical significance concerns and sampling issues which tend to lead to result generalizability issues) or artificially-generated samples (prone to wrong parameterisation, and thus, irrelevant results).

### 2.3.1. Data generation and connectivity

For this research paper, the latter approach was chosen. In order to mitigate against the potential use of inappropriate assumptions, principal parameters that describe similar systems have been drawn through the literature (including the *topology* of and *distributions* within them). More specifically, our S.N. can be defined as a scale-free (SF) network and thus, generated using a widely used model [Barabási, Albert 1999]. Such model replicates one of the most important aspects relative to our research – the power-law degree distribution that defines the number of connections per node. It is worth noting that since the S.N. is an emergent network (i.e. no Grande architect can specify the interactions between social agents), no other network topology will be considered, as numerous equivalent networks have been found to be SF [Newman 2009]. On the other hand, the T.N. has been generated using three fundamentally different network topologies in order to account for different features. Namely, the SF network topology (again, replicating the widely noted feature of an infinitely variant degree distributions, similar to the S.N.), the Small-World (SW) topology (typically replicating the evidently high local clustering between regions of nodes, as observed in numerous real networks [Watts, Strogatz 1998] or a random network (in order to represent a homogeneous degree distribution, i.e. where each technological system is more or less connected as any other node) [Erdős, Rényi 1959]. The reasoning behind the use of three rather different models lies to the fact that the T.N. can be considered as a designed system where a designer can explicitly decide the number of connections for each subsystem. Finally, the NoN is the resulting superposition of the S.N. and T.N. – varying from a pure SF network to a SF/SW and SF/Random network. It is further reasoned that within a typical sociotechnical system of such context, executive employees have direct access to critical information (e.g. databases) and thus, do not need to interact with a large number of other social agents (assuming a vertical organisational structure). Thus, a disassortative mix is assumed i.e. highly connected nodes within the T.N. (e.g. servers) will be connected with low degree nodes of the S.N. (e.g. executive employees).

### 2.3.2. Discrete entities

In terms of the nature of the entities themselves, a high level, abstract representation have been chosen in order to strive for generalizability and extend the applicability of the model. Specifically, the entities composing the S.N. are coined as regulatory agents (which can be further decomposed as organisations, companies, teams, individuals, etc.) Similarly, a high-level representation of the information system can be considered to be composed of technological entities (as such, they may include PCs, routers, servers, etc.) – see Figure 1.



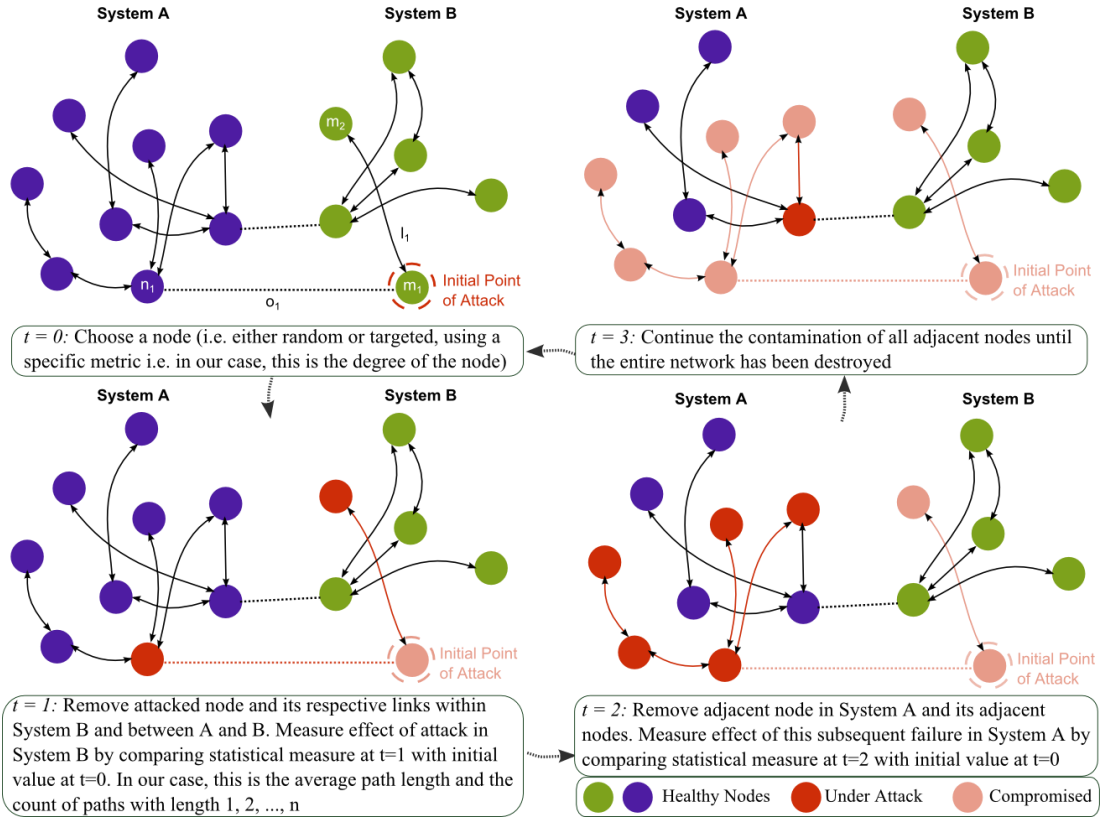
**Figure 1: Entity aggregation in order to contribute towards a generalizable model. This paper focuses on the interaction between two systems (see boxed area) rather than within a system.**

### 2.3.3. Attack mode

Two different attack modes will be explored; a random attack (i.e. select a node at random and remove it) and a targeted attack (i.e. attack the most connected node in the overall NoN). Figure 2 presents the propagation process diagrammatically. As an example, let as first attack node  $m_1$ ; node  $n_1$  and  $m_2$  will be subsequently affected (due to link  $o_1$  and  $l_1$  respectively) and a cascade of effects will be resonate within the entire NoN until the latter reaches a non-functioning state due to the lack of links.

The attack is assumed to be in the form of a contagion without a threshold, i.e. the progression of the infection has a probability of 1 to progress and thus, will continue to spread and remove functional nodes until all nodes have been removed. This is an extreme case and is rather simplified as mitigation measures are expected to allow for nodes to recover. Nevertheless, since the purpose of the paper is to stress test a sociotechnical system under a worst-case scenario, such mechanism can serve as a reasonable methodology for comparing the resilience of varying networks topologies. Similarly, another assumption is that the attacks in the model are based on whether a node is available or not (i.e. it is of binary nature). Further improvement could introduce a simple continuous function in order to accommodate for a number of realistic situations such as a node (e.g. a web server) operating at half capacity due to the attack.

For the purpose of monitoring the effect of the attack on the functional capacity of both S.N., T.N. and NoN, a suitable proxy needs to be chosen. Since the fundamental characteristic of the network is the capacity for nodes to interact with each other, a natural measure is the use of “distance” (i.e. how many links need to be traversed in order to reach any one node from any other node) between nodes as a measure of its capacity to deliver its designated function. If one was to measure the rate of change of this measure during an attack, then it could serve as a reasonable proxy on the overall performance of the system.



**Figure 2: Illustration of the attack process adopted in our model. Notice that the nature of the attack mode only matters at the initial set-up; the subsequent steps are identical**

Such measure is the path-length – it will be used in two forms; the raw path length (i.e. the number of links needed to connect a pair of nodes) and the *average* path length. The latter can be defined as the shortest distance ( $d$ ) between every pair of nodes  $i$  and  $j$ , averaged for the entire set of nodes ( $n$ ) within a network; mathematically defined as:

$$l_{average} = \frac{1}{n \times (n-1)} \times \sum_{i \neq j} d(v_i, v_j) \quad (1)$$

The average shortest path is a commonly used proxy for understanding the rate of which a given quantity (let it be information flow, a malicious virus, regulatory signals, etc.) can transverse a network [Newman 2009]. Although it is rather obvious that one should strive for as small average path length as possible in order to utilise less resources in moving the aforementioned quantity, the immediate drawback is that any malicious inclusion within a network can spread much faster within the network and thus, its resilience reduces.

### 3. Simulation results

#### 3.1. Average Path Length

The first output of the simulation focuses on the change of the average path length (averaged over 5 simulation runs) as the two aforementioned attack modes (i.e. targeted and random) progress. The description of the results will be divided in two sections; (1) revolving around the influence of the nature of the underlying topology and attack model; (2) on the effect that the network size has upon the rate of change of its average path length. Relevant model output can be seen in Figure 3. It should be noted that three variants of each network are used in order to introduce the size of the network (in terms of node count) as another variable as it will inevitably affects the average path length.

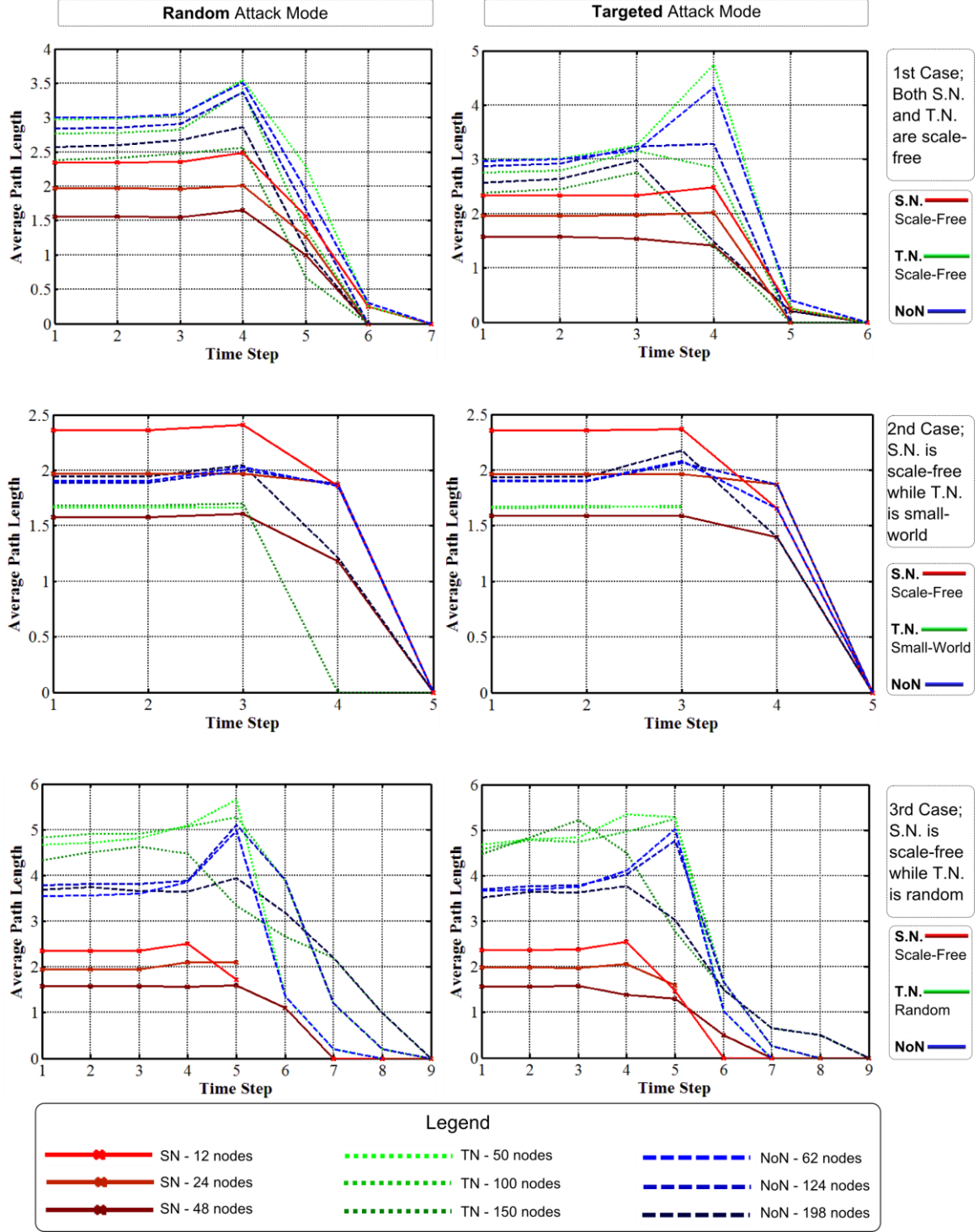
##### 3.1.1. On the influence of network topology/attack mode in terms of average path length

By carefully examining the two columns of results, it appears that the effect of the different attack mode heavily depends on the nature of the underlying topology. To begin with, it appears that under a targeted attack, more steps are required to completely destroy the network when the underlying topology of the T.N. is either SW or Random (2nd and 3rd row respectively); less so when it is a SF network (1st row). The underlying reason can be found on the degree distributions of these different topologies [Albert, Jeong, Barabási 2000] [Demetrius, Manke 2005]. More specifically, both SW and Random are homogeneous since their degree distributions follow a normal distribution (i.e. the majority of nodes are equally connected and thus, the difference between random removal and targeted removal is small). On the other hand, a SF topology implies a highly heterogeneous degree distribution (i.e. most nodes are poorly connected while some are highly connected) thus the removal of such highly connected nodes (which effectively act as “distribution hubs” in terms of transferring any set quantity within the network) has a major effect in its underlying connectivity.

##### 3.1.2. On the influence of network size in terms of average path length

One can first notice that depending on the scale of each underlying network, the initial average path length is higher – this is intuitively expected as larger networks need, on average, more steps to transfer a quantity between any two randomly chosen nodes.

A significant feature that is present in the entirety of cases is the sudden phase change observed. More specifically, there appears to be a sudden difference where small, incremental changes in the average path length (usually decreasing, though instances exist where a notable increase is first observed) immediately lead to a large decrease in the next. As a typical example, consider sub-plot (1, 1) in Figure 3, in which small changes take place up to the time step 4 (upon which the user of any of the three networks would not notice any significant changes in the functionality of the network), followed by a radical decline after only one time step, concluding with the complete deterioration of the largest NoN at time step 7.



**Figure 3. Change in average path length for S.N., T.N. and NoN under the two attack modes. For each case, S.N. is defined by a SF topology while T.N. alternates from SF (1<sup>st</sup> case) to SW (2<sup>nd</sup> case) to random (3<sup>rd</sup> case). Note that for each network, three different sizes are used per case.**

### 3.2. Frequency of paths of various lengths

We will now shift our focus on the frequency of path of various lengths found within each network, as an attack progresses, again averaged over 5 runs. For the sake of simplicity, the targeted attack will be used as the attack mode of interest as it is expected to be the largest concern for any organisation that strives for SA as it will be an attack mode which is out of the organisation's control. What we are interested in looking at in this sequence of results is how the histogram of path length sizes is

transformed as the attack progresses, along with how the different topologies and network sizes influence this transformation. Relevant model output can be seen in Figure 4.

### 3.2.1. First case (scale-free T.N.)

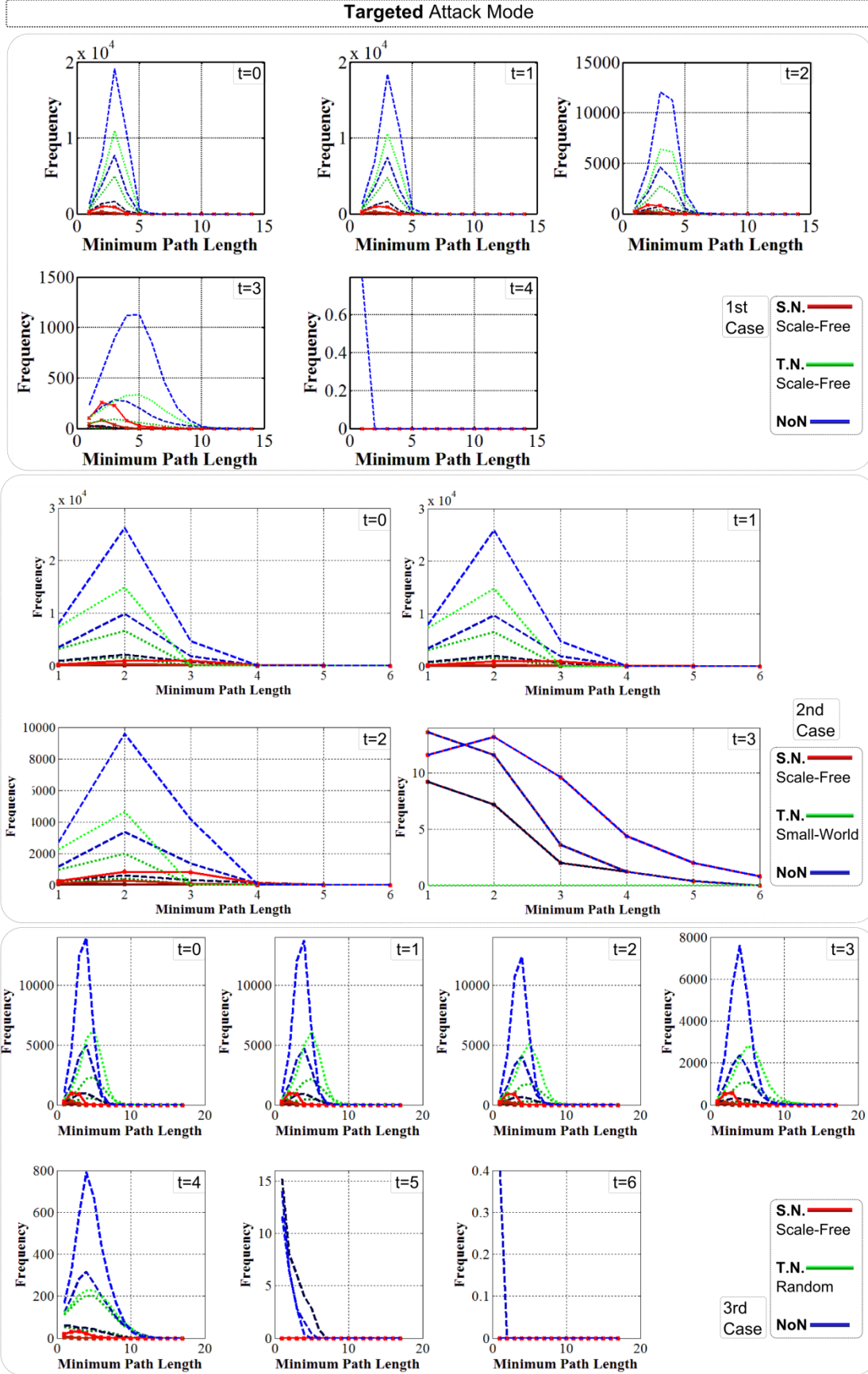
In the case of a scale-free T.N., an initial narrow distribution of the path length with extremely thin tails is firstly observed, while the size of the network appears to be exerting an influence on the peak values. Specifically, it can be seen that as the size of the network increases, the peak value of the distribution becomes more pronounced. As the attack progresses, there is a rather constant decrease of the mean average path lengths (since it is the most frequently observed path, it has a higher probability of removal) leading to a widening of the peak. Again this effect is more pronounced to the largest networks (moving from the NoN, to the T.N. and finally, to the S.N.). It is worth noting that when  $t=3$ , the tail of the distribution becomes much wider for all three networks – occurrences of paths with a length greater than 10 is now likely where this was not previously the case. With reference to Figure 3, subplot (1, 2), one can easily see that after  $t=3$ , a dramatic decrease is observed in all three networks (with T.N. being completely decomposed) which is reinforcing the emergence of the longer paths. This is mainly due to the fact that as nodes are being removed, it becomes exceedingly costly (i.e. more nodes need to be traversed) to reach any other node, assuming that they are still connected by a path. Finally, at  $t=4$  the entirety of the large majority of the networks have been decomposed, with some trivial number of paths within the NoN still remaining – at this point one can safely assume that all three networks are no longer functioning. One should note that during this iteration, the NoN is purely a SF network as both constituent networks (S.N. and T.N.) are of the same nature.

### 3.2.2. Second case (small-world T.N.)

In this case, the T.N. is now defined by a small-world network topology while the S.N. remains by *de facto* SF. Compared with before; the NoN is now a mixture of both SW and SF topology. We notice that at the initial stage, the path length distribution is very similar to the first case, although the peaks are less pronounced for the T.N. and NoN; S.N. is qualitatively similar to the previous case. As the attack progresses, there is no widening of the distribution nor any emerging longer paths. Interestingly, by  $t=3$  both S.N. and T.N. are now severely decomposed with only some reminiscent connections found within the NoN – again by this point, one can assume that the functionality of all three networks has been efficiently reduced to zero. It is also worth pointing out that under this scenario, the entirety of the networks are completely decomposed after  $t=3$  – this is by far the fastest deterioration as both first and third case are more resilient (lasting up to  $t=4$  and  $t=6$  respectively).

### 3.2.3. Third case (random T.N.)

Following our original test hypothesis, T.N. is now defined as a random network while S.N. remains a SF network. Similarly, the NoN is now a compounded version of both network topologies. As in both previous cases, the distributions are relatively well defined with a high mean peak although they appear to have a larger variance – for example the most frequent path within the NoN is of length 4 – for case 1 and 2 this was equal to 3 and 2 respectively. Up to and including  $t=3$ , the observed behaviour appears to be qualitatively similar to case 2 where there is a uniform reduction of all path lengths – remember that this was not the outcome in case 1, where the majority of paths removed were of the most frequent one making the percentage composition of the second most frequent, path to increase. Conversely, the path percentage composition is much more balanced in case 3 (case 2 remains somewhere in between). At  $t=4$ , a rather interesting phenomenon is observed – the same populating process of longer paths (again, this is due to removal of a critical number of nodes which leads to the emergence of much longer paths) is observed but this effect is counterbalanced by the increase in the frequency of much smaller paths. The latter is an effect of the emergence of smaller clusters which are inevitably composed by small paths. By referencing the Figure 3, subplot (3, 2), one can see the sudden drop of average path length on all three networks at exactly  $t=4$ , which further reinforces our casual explanation. This is similar to the effect observed at  $t=3$  in the 1st case. However, the rate of decomposition is much more consistent over all three networks than in the other two cases.



**Figure 4.** Temporal evolution of the frequency (y-axis) of various path lengths (x-axis) with respect to the S.N., T.N. and NoN during a targeted attack. All three cases are shown.

### 3.3. Future model improvements

By further refining the features that can be parameterised in the network generation process, the degree of generalizability of the model would improve. For example, the algorithms that generate both the S.N. and the T.N. have no specified purpose other than matching a number of user-defined, statistical criteria (e.g. degree distribution). Although such features appear to be universal [Barabási, Albert 1999], they are context *independent* and thus, tailoring of the methodology needs to take place in a case by case fashion to ensure specific features (e.g. the tendency of social networks to be assortative while technological networks being disassortative [Newman 2002]) are taken into account.

## 4. Design Implications

We have assumed that the S.N. cannot be designed in any way and since it has been reported numerous times in the literature that such networks tend to be SF [Newman 2009] with an exponent of 2.3 [Nicosia, De Domenico, Latora 2013], this has been the prevailing topology used. However, one can explicitly design the topology of the underlying T.N. and thus, in this section, we will present some of the insight induced by the model in order to strive for increased robustness in terms of the T.N. (and subsequently, the NoN).

By combining observations presented in section 3.1., 3.2., along with Figure 3 and 4 respectively, three design considerations can be inferred. 1) The fact that the SW and random topologies take longer to fail suggest that these are more robust to both random and targeted attacks. 2) The SW and random networks have a gradual increase in average path-length before their sudden decomposition. These designs could thus provide a more gradual and measurable system failure. Coupled with longer reaction times could allow a system intervention before a catastrophic failure occurs. 3) The larger the size of the network, the more time a system tends to have before a catastrophic disaster. However, this is most likely to the system's discredit as the more extensive the response would have to be. The only imaginable scenario where this could be beneficial is if the T.N. were in fact multiple small networks that are easily compartmentalised and intervened.

We have identified that a SF topology may provide some early warning signs in terms of reduced performance (due to the emergence of longer paths – Figure 4) as a proxy of a potential reduction in the functionality of the T.N. (and subsequently, NoN) – such early warnings are completely absent when the topology of the T.N. is either SW or random (i.e. any form of topology that has an underlying homogeneity in terms of node degree). Even worse, as discussed briefly in the section regarding the third case, when the T.N. possesses a random topology, an increase is seen in both short path lengths (i.e. improvement in performance) along with the emergence of longer paths (fitted to a reduction, and thus, counterbalance of performance). If monitored correctly, this may be a better proxy to use in terms of identifying whether your system is being maliciously attacked. Either way, having small local clusters (i.e. SW topology) is something commonly experienced in organisations as T.N. systems are built around specific departments which are then connected together with other departments – such approach provides no early warning signs of an eminent attack and thus such design appears to be the least efficient in terms of SA. It should, however, be noted that all three architectures experience sudden phase changes which do not scale with the size of the networks i.e. there is a sudden switch from fully operational to highly reduced operational capacity, thus, the capacity to introduce any sort of early warning signs is extremely important.

Given the progression of the average path-length of the SW and SF topologies, the mediated response of the random topology is in comparison, the obvious choice given that it presents by far, the best survivability. However, one must consider the trade-off between reliability and efficiency. As such, a greater number of short path-lengths would suggest greater efficiency, but the trade-off between peaking at ~2 versus ~4 cannot be easily transferable without more specific, context-dependent model definitions. Such type of study needs a specific definition of what the sociotechnical system is and our given level of aggregation merely provides a methodological framework rather than easily transferable results.

## 5. Summary

This paper has reasoned that any organisation can be viewed in terms of its functional connections; specifically by viewing it as a Network-of-Networks compromised by a network of social agents who operate/monitor it and a technological network composed by a number of aggregated, information-processing artefacts. It has been proposed that the functionality of the NoN is directly related to the topology and size of its composing networks and by understanding their reaction under two attack modes, important design decisions can be taken to improve its resilience.

The path length (in its raw and averaged form) has been used as a proxy to infer the effect of the attack in terms of the networks' functionality and have noted the following:

- Size of networks only plays a role in the initial conditions but has not a significant effect in terms of the network's path deterioration, and subsequently, ability to perform under an attack.
- All three network topologies illustrate sudden phase changes in which they exhibit a dramatic loss of functionality in a very short period of time (e.g. Figure 3, subplot (1, 2) moving from  $t=3$  to  $t=4$ ). This has been a universal result, regardless of attack mode, topology and size.
- Finally, since phase shifts are eminent universal behaviours, early warning signs in the form of reduced performance before a complete breakdown (i.e. emergence of longer paths) is desirable (see Figure 4). In such terms, only the SF topology exhibited such behaviour (section 3.2.1) while the SW topology (a typical arrangement for a number of organisations who tend to design technological systems within specific department, with only few connection between departments). Random topologies have the largest capacity to provide early warning signs (as there is also an increase in the number of short paths, along with the emergence of largest paths) but there is an increased cost in utilising such a topology as it implies that every entity within the T.N. needs to be approximately connected with an equal number of connections as any other node (i.e. implying increased number of redundancies).

## Acknowledgement

The authors would like to thank Prof. Chris McMahon for steering this work and the staff at the IDC in Systems. Funding from EPSRC, Systemic Consult Ltd, EADS and Aero Engines Control is gratefully acknowledged.

## References

- Albert, R., Jeong, H., Barabási, A.-L., "Error and attack tolerance of complex networks", *Nature*, 406, 2000, pp. 378-382.
- Arthur, W. B. "The nature of technology: What it is and how it evolves", Penguin Group England, 2009.
- Barabási, A.-L., Albert, R., "Emergence of scaling in random networks", *Science*, 286, 1999, pp. 509-512.
- Barford, P. et al. "Cyber Situational Awareness", Springer US, 2010.
- Demetrius, L., Manke, T., "Robustness and network evolution—an entropic principle", *Physica A: Statistical Mechanics and its Applications*, 346, 2005, pp. 682-696.
- DSTL, "Cyber Situational Awareness – Centre for Defence Enterprise", MoD UK, 2012.
- Endsley, M. R. "Toward a Theory of Situational Awareness in Dynamic Systems", *The Journal of the Human Factors and Ergonomics Society*, 1995, pp. 32-64.
- Erdős, P., Rényi, A., "On random graphs", *Publicationes Mathematicae Debrecen*, 6, 1959, pp. 290-297.
- Newman, M. E.J., "Assortative mixing in networks", *Physical Review Letters*, 89, 2002, pp. 208701 – 208704.
- Newman, M.E.J., "Networks: An Introduction", Oxford Press London, 2009.
- Nicosia, V., De Domenico, M., Latora, V., "Characteristic exponents of complex networks", *arXiv preprint arXiv:1306.3808*, 2013.
- Watts, D. J., Strogatz, S. H., "Collective dynamics of 'small-world' networks", *Nature*, 393, 1998, pp. 440-442.

Mr Christos Ellinas  
Research Engineer  
University of Bristol, IDC in Systems  
+44 (0) 7823559283  
ce12183@bristol.ac.uk