entropy

MDPI

Article

Simulating Weak Attacks in a New Duplication–Divergence Model with Node Loss [†]

Ruihua Zhang [1,*] and Gesine Reinert [1,2]

[1] Department of Statistics, University of Oxford, 24–29 St. Giles', Oxford OX1 3LB, UK; reinert@stats.ox.ac.uk
[2] The Alan Turing Institute, London NW1 2DB, UK
[*] Correspondence: ruihua.zhang@stats.ox.ac.uk

[†] This paper is an extended version of our paper published in The 12th International Conference on Complex Networks and Their Applications, Menton Riviera, France, 28–30 November 2023.

Abstract: A better understanding of protein–protein interaction (PPI) networks representing physical interactions between proteins could be beneficial for evolutionary insights as well as for practical applications such as drug development. As a statistical model for PPI networks, duplication–divergence models have been proposed, but they suffer from resulting in either very sparse networks in which most of the proteins are isolated, or in networks which are much denser than what is usually observed, having almost no isolated proteins. Moreover, in real networks, where a gene codes a protein, gene loss may occur. The loss of nodes has not been captured in duplication–divergence models to date. Here, we introduce a new duplication–divergence model which includes node loss. This mechanism results in networks in which the proportion of isolated proteins can take on values which are strictly between 0 and 1. To understand this new model, we apply strong and weak attacks to networks from duplication–divergence models with and without node loss, and compare the results to those obtained when carrying out similar attacks on two real PPI networks of *E. coli* and of *S. cerevisiae*. We find that the new model more closely reflects the damage caused by strong and weak attacks found in the PPI networks.

Keywords: duplication–divergence model; gene loss; weak attack; protein–protein interaction networks

check for updates

Citation: Zhang, R.; Reinert, G. Simulating Weak Attacks in a New Duplication–Divergence Model with Node Loss. *Entropy* **2024**, *26*, 813. https://doi.org/10.3390/e26100813

Academic Editor: Hocine Cherifi

Received: 20 June 2024
Revised: 7 September 2024
Accepted: 17 September 2024
Published: 25 September 2024

1. Introduction

From virtual internet to practical traffic control systems, from small social networks to large biological systems, networks are ubiquitous, and so are attacks on networks. For example, an internet cyber attack can slow down information transmission or cause information leakage, and drugs can target a number of different proteins. Reference [1] shows that partial inactivation of multiple nodes simultaneously in a network can be more effective than the complete elimination of a node, by measuring the sum of the inverse of the shortest path between any two nodes of biological networks (the *network efficiency*).

This result motivates the study of weak attacks in pharmaceutical designs. For example, broader-specificity, lower-affinity compounds or multidrug therapies may cause larger damage in network efficiency than high-affinity, high-specificity compounds. The success of multitarget drugs, like non-steroidal anti-inflammatory drugs (NSAIDs) [2], metformin [3], and Gleevec [4], to treat diseases including AIDS, cancer, atherosclerosis, and Alzheimer's disease, all suggest that attacking multiple targets may be a useful therapeutic strategy.

To anticipate the effect of an attack, a well-fitting parametric network model could help gain insights. For protein–protein interaction (PPI) networks, duplication–divergence (DD) models have been suggested, see for example [5–7]. This paper hence starts with practically simulating weak attacks in a duplication–divergence model. Simulations from [8] suggest that DD models can generate networks which resemble PPI networks more than a basic Bernoulli random graph model. However, ref. [9] found that while Monte Carlo tests based on network comparison statistics do not reject the DD model for some small-virus PPI networks, they do reject it (at the 5% level) for *E. coli*, *worm*, *fly*, *S. cerevisiae*, and *human* PPI networks. Indeed, DD models are known not to be very realistic; for example, ref. [10] proved that as the number of nodes tends to infinity, the proportion of isolated nodes in a standard DD model converges to either 0 or 1, neither of which is realistic.

To understand theoretically how weak attacks damage PPI networks, it is instructive to consider a simple Bernoulli $G(n, M)$ random graph with n nodes and M edges. We derive a Poisson approximation for the number of isolated nodes in a $G(n, M)$ via Stein's method, which gives explicit bounds in total variation distance, and we prove similar bounds for the number of isolated nodes after different attack strategies. These results lead to a clear statistical rejection of the hypothesis that the real PPI networks in this paper follow a $G(n, M)$ model.

To identify a more realistic model for PPI networks, we notice that the current DD models ignore gene losses, a biological function [11] which can potentially balance the proportion of isolated nodes. As genes code for proteins, it is plausible that a model with node loss may perform better than standard duplication–divergence models for PPI networks. This paper introduces a new DD model with node loss, where a node can be lost with probability q if it is isolated. We compare the simulation results of weak attacks in a standard DD model and the DD model with node loss, and conclude that the new model indeed generates a more realistic performance.

This paper is structured as follows. Section 2 describes the datasets and attack strategies that are employed, as well as the damage strategy and measures of damage. Section 4 introduces the new DD model with node loss. Simulations of various attack strategies on PPI networks on real and synthetic networks are provided in Section 5. The results are discussed in Section 6. Appendix A contains details of the Poisson approximation results and Appendix B contains additional figures. The code is available at https://github.com/rh-zhang/Entropy_CNC2023 (accessed on 24 August 2024).

2. Data and Methods

2.1. Datasets

We use PPI networks for *E. coli* and *S. cerevisiae* downloaded from STRING (version 12.0, accessed on 11 March 2024), restricted to physical interactions between proteins only. The resulting networks are unweighted, undirected physical subnetworks representing direct interactions between proteins only, excluding indirect functional associations. We remove interactions with a STRING score [12] less than 0.500 for the *E. coli* PPI network and less than 0.400 for the *S. cerevisiae* PPI network, taking all evidence channels into account. The 0.400 threshold is the default threshold in STRING; the 0.500 threshold for *E. coli* is chosen such that the number of isolated nodes is of a similar magnitude (around 1100) in both networks, see Table 1. As shown in Figure 1, the number of isolated nodes increases as the threshold of STRING scores increases. However, the overall trend regarding the impact of weak attacks on the networks remains consistent in our results, as shown in Figure A9.

We note that there is no claim that all possible protein–protein interactions have been detected, and hence the STRING database is unlikely to contain all true interactions; it may also contain some false positive interactions. Our study is conceptual and hence not severely affected by such false positives and false negatives, under the assumption that there is no strong systematic connection between errors in the data and isolated proteins.

We assign a uniform weight of 1 to all the remaining edges in the datasets, with the summary statistics shown in Table 1. The reason for ignoring weights is conceptual simplicity.

Table 1. Summary statistics for the analysed networks; No. stands for *Number of*.

Networks	***E. coli***	***S. cerevisiae***
No. nodes	3043	5925
No. edges	52,914	140,402
No. isolated nodes	1141	1100
Average degree	28.05	59.51
Average local clustering coefficient	0.31	0.40
Global clustering coefficient	0.25	0.80

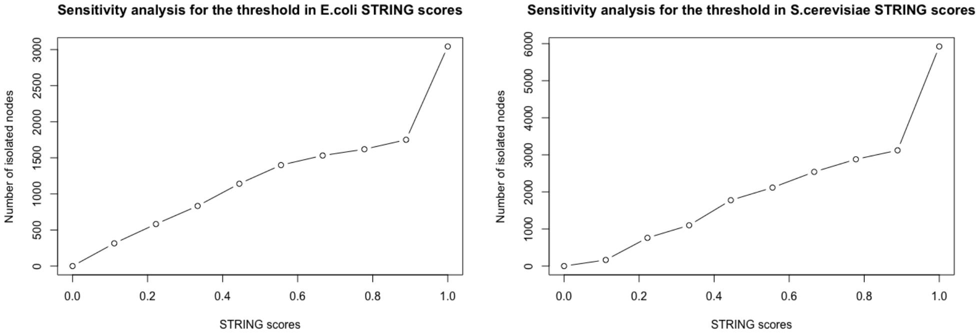


Figure 1. Sensitivity analysis for the number of isolated nodes in the *E. coli* and *S. cerevisiae* PPI networks across varying STRING score thresholds.

2.2. Attack Strategies

The attack strategies used in this paper follow those from [1]. While in [1], networks with weighted edges are allowed, in our investigative study we set all edge weights equal to 1 initially; some attacks lead to a reduction in some of the edge weights. The attack strategies are split into three categories.

Type A: Complete knockout: the attack of a single target by eliminating all interactions of a given node, as shown in Figure 2A.

Type B: Partial inactivation of a target, as shown in Figure 2: B, which is modelled in two different ways:

B1: Partial knockout: half of the interactions of a given node are removed (the number of interactions removed is rounded down when the degree of the target is odd). If a node is attacked partially once, it will not be attacked again to ensure no node is completely knocked out. This is shown in Figure 2B1.

B2: Attenuation: all interactions of a given node are attenuated by halving their weight.

Type C: A distributed, system-wide attack, which can affect any interactions (i.e., edges) within a network. Again, such an attack is modelled in two different ways:

C1: Distributed knockout: edges are deleted independently at random, with the same deletion probability, as shown in Figure 2C1.

C2: Distributed attenuation: edges are chosen independently at random, with the same probability, and their weights are halved.

These attacks can be interpreted in pharmaceutical terms; a high-affinity drug completely eliminates an interaction while a low-affinity drug attenuates it, and a highly specific drug targets one single interaction only, while less specific drugs affect some or all interactions of a given node.

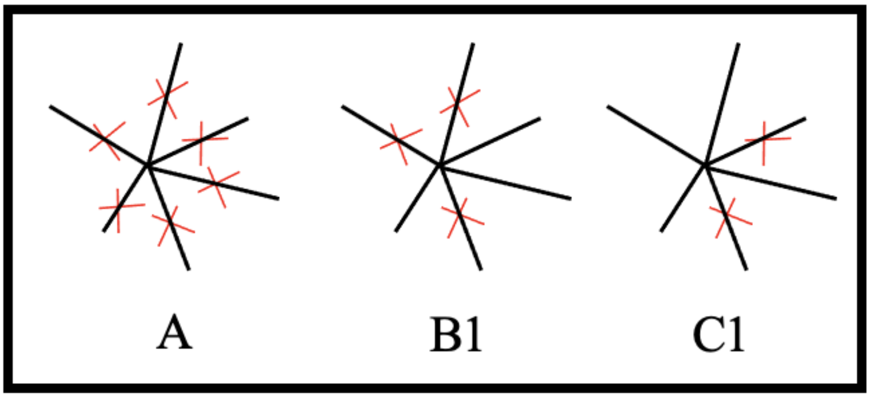


Figure 2. Attack strategies. (**A**) Complete knockout attack: all edges connected to the attacked node are eliminated. (**B1**) Partial knockout attack: half of the edges connected to the attacked node are eliminated. (**C1**) Distributed knockout attack: randomly selected edges are eliminated. Adapted from FIG.1 in [1].

2.3. Successive Maximal Damage Strategy

As in [1], the nodes being attacked in the simulation of this paper are selected based on a successive maximal damage strategy. The search for maximal damage caused by multiple attacks is computationally very intensive. For instance, to determine which 5 of the 1000 edges of a given network need to be deleted in order to produce a maximal effect on the network efficiency, one would need to test $1000!/(5!995!) \approx 8.25 \times 10^{12}$ cases in a single-simulation experiment.

Instead, we use a greedy algorithm: for each type of attack, in each step we choose the action that produces the largest damage. The greedy algorithm is carried out by first determining the damage caused by the removal of each individual node or edge, depending on the strategy. The node or edge causing the maximum damage is selected for removal in the subsequent attack. We note that the damage calculated in this manner is only an estimate of the maximal damage, since there may be more efficient combinations.

2.4. Measures of Damage

The damage induced by the attacks on the networks is measured by three metrics: the network efficiency for the transcriptions regulator networks, as used in [1], the average number of edges in the 1-step ego network for the PPI networks, as proposed in [13] for assessing the robustness of network metrics, and the number of isolated nodes.

The **network efficiency** (NE) of an undirected, unweighted graph of n nodes is $\sum_{i \neq j} \frac{1}{d_{ij}}$, where d_{ij} is the length of a shortest path between nodes i and j. If the network is weighted, d_{ij} is the weight of a path between nodes i and j with a minimum weight. If any two nodes $i \neq j$ are disconnected, then $d_{ij} = \infty$, and their contribution to the calculation of network efficiency is 0. NE measures how efficiently a network exchanges information. The underlying idea is that the more distant two nodes are in a network, the less efficient their exchange of information will be.

The second measure is the **average number of edges in the 1-step ego network,** where a 1-step ego network consists of a focal node (the *ego*), the nodes to which the ego is directly connected (the *alters*), and the edges, if any, among the alters.

The third measure is the **number of isolated nodes.** We add this measure because an ideal attack would isolate a deleterious node. Moreover, in a Bernoulli random graph model this measure can be analysed analytically and thus is useful for providing theoretical underpinning.

2.5. Bernoulli Random Graphs

Given the number n nodes and the number M of edges in a simple network, in the absence of further information one may model the network as a $G(n, M)$ graph. This is a random graph that is chosen uniformly at random from the collection of all simple graphs which have n nodes and M edges, where $0 \leq M \leq \binom{n}{2}$.

The distribution of the degree of a node v, $D(v)$, in a $G(n, M)$ graph is hypergeometric; there are $n-1$ edges that are adjacent to v, out of the $\binom{n}{2}$ potential edges, of which we choose M. Abbreviating the number of node pairs by $N = \binom{n}{2}$ we thus have

$$\mathbb{P}(D(v) = k) = \frac{\binom{n-1}{k}\binom{N-(n-1)}{M-k}}{\binom{N}{M}}, \quad k = 0, 1, \ldots, M.$$

We can calculate the expected number of isolated nodes from this distribution, but not its variance, due to the dependence between edges. To clarify the dependence, for example, if we know that the first $n-1$ nodes have degree 0, then node n necessarily must have degree 0. As this dependence is usually weak, we derive a Poisson approximation for the number of isolated nodes in the total variation distance. The *total variation distance* d_{TV} measures the largest absolute difference between the probabilities of the actual probability distribution and the Poisson approximation. For distributions P and Q taking values in $\mathbb{Z}_+ = \{0, 1, \ldots\}$, the total variation distance is defined as

$$d_{TV}(P, Q) = \sup_{A \subset \mathbb{Z}_+} |P(A) - Q(A)|. \tag{1}$$

For $M \leq N - (n-1)$, the probability that node i is isolated is

$$\mathbb{P}(I_i = 1) = \frac{\binom{N-(n-1)}{M}}{\binom{N}{M}} := \pi.$$

With W denoting the number of isolated nodes, its expectation is $\mathbb{E}(W) = n\pi =: \lambda$, and this is the parameter which we choose for the approximating Poisson distribution.

Theorem 1. *It holds that*

$$d_{TV}(\mathcal{L}(W); Po(\lambda)) \leq \min(1, \lambda^{-1}) e^{-np(1+\frac{n-2}{N+2-n})+p(1+\frac{n-2}{N+2-n})+\frac{2-3n+n^2}{N+2-n}} \left(1 + np\left(1 + \frac{n + Np - 2}{N - Np - n + 2}\right)\right).$$

This bound tends to zero as $p := M/N \to 1$. The proof and more details can be found in Appendix A.1. Appendix A.1 also gives Poisson approximations for the number of isolated nodes after an attack for the different attack strategies. These results may be of independent interest.

3. Duplication–Divergence Models

Simulations suggest that duplication–divergence (DD) models generate networks which provide a better fit to protein interaction networks than the standard models [8]. There are different variations of duplication–divergence models in the literature, see for example [6,7,14,15]. Here, we use a version, from [15], which incorporates the parameters of the probability of edge divergence, p, but we exclude the possibility of a parent–child edge.

A standard duplication–divergence model $DD(t_0; p)$ starts from a complete graph G_{t_0} on t_0 nodes (labelled from 1 to t_0), and then repeats the following steps until a graph of the desired size is obtained:

- **Duplication:** at time t, a node u is selected uniformly at random. A node labelled as $t + 1$ is added, as well as the edges between node $t + 1$ and the neighbours of node u in the graph.
- **Divergence:** edges involving node $t + 1$ are randomly retained with probability p.

An illustration of a DD model is shown in Figure 3.

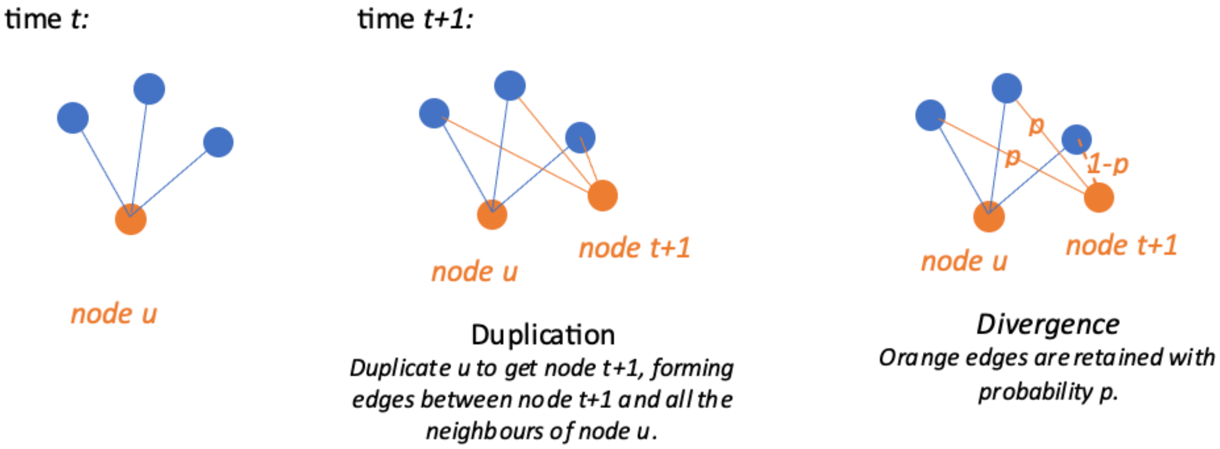

time *t:*
node u
time *t+1:*
node u
node t+1
Duplication
Duplicate u to get node t+1, forming edges between node t+1 and all the neighbours of node u.
p
p
1-p
node u
node t+1
Divergence
Orange edges are retained with probability p.

Figure 3. Graph illustration of a duplication–divergence model.

Reference [15] found that the degree distributions of the DD model described above are in reasonable agreement with the distributions observed in real protein networks, and tuning the parameter p reveals a rich behaviour of the model. When p is large, the network growth lacks self-averaging and results in a great diversity of networks grown out of the same initial condition. For $p < 0.5$, the average degree increases very slowly or tends to a constant, and the degree distribution has a power-law tail. Several real protein–protein networks are estimated to have a p value of around 0.4 [15]. As shown in Figure A1, the choice of p does not affect the qualitative behaviour of the models against attacks.

4. A New Duplication–Divergence Model Which Allows for Node Loss

Although simulations have shown that the DD model described above is more realistic than a $G(n, M)$ model, ref. [10] proved that the proportion of isolated nodes in a DD model either converges to 0 or 1. This behaviour does not match biological intuition, and other network models do not exhibit it; for example, we prove in Appendix A that the proportion of isolated nodes in a $G(n, M)$ model does not have to converge to either 0 or 1.

The quality of a network model has to be judged by the research question to be addressed. In a series of Monte Carlo tests for *E. coli*, *worm*, *fly*, *S. cerevisiae*, and *human* PPI networks and some small-virus PPI networks [9], a DD model (allowing for a non-zero probability of a parent–child edge) is rejected as a model for the large PPI networks based on network comparison statistics including graphlet correlation distance, graphlet degree distribution agreement, Netal, and Netdis. In contrast, in the small-virus PPI networks investigated in [9], the DD model is not rejected by most of these network comparison statistics. These statistics do not include the number of isolated nodes, but Netdis is based on subgraph counts in ego networks, and is thus related to our outcome measure of the average number of edges in 1-step ego networks. Hence, these Monte Carlo results indicate that the DD model may not be a good fit for larger PPI networks when the interest is in modelling the effect of attacks.

From a biological viewpoint, genes and the proteins they code for can not only duplicate, but can also be lost. For example, gene loss can occur during natural mutations and frameshifts [16]. Furthermore, many examples support the idea that gene loss can be an adaptive evolutionary force that is especially common when organisms are faced with abrupt environmental challenges [11]. Adaptive gene loss, or gene loss in general, can be of potential interest in the study of both biomedicine and evolution.

Therefore, we modify the DD model to allow for both node addition and for the loss of nodes. In addition to the process that generates a DD model, a node loss step is added after every duplication-and-divergence step. In particular, we focus on the node loss mechanism that a node can be lost with probability q if it is isolated.

- **Duplication:** at time t, a node u is selected uniformly at random. A node labelled as $t+1$ is added, as well as the edges between node $t+1$ and the neighbours of node u in the graph.
- **Divergence:** edges involving node $t+1$ are randomly retained with probability p.
- **Node loss:** a node is randomly lost with probability q if it is isolated.

A graph illustration of our new model is present in Figure 4.

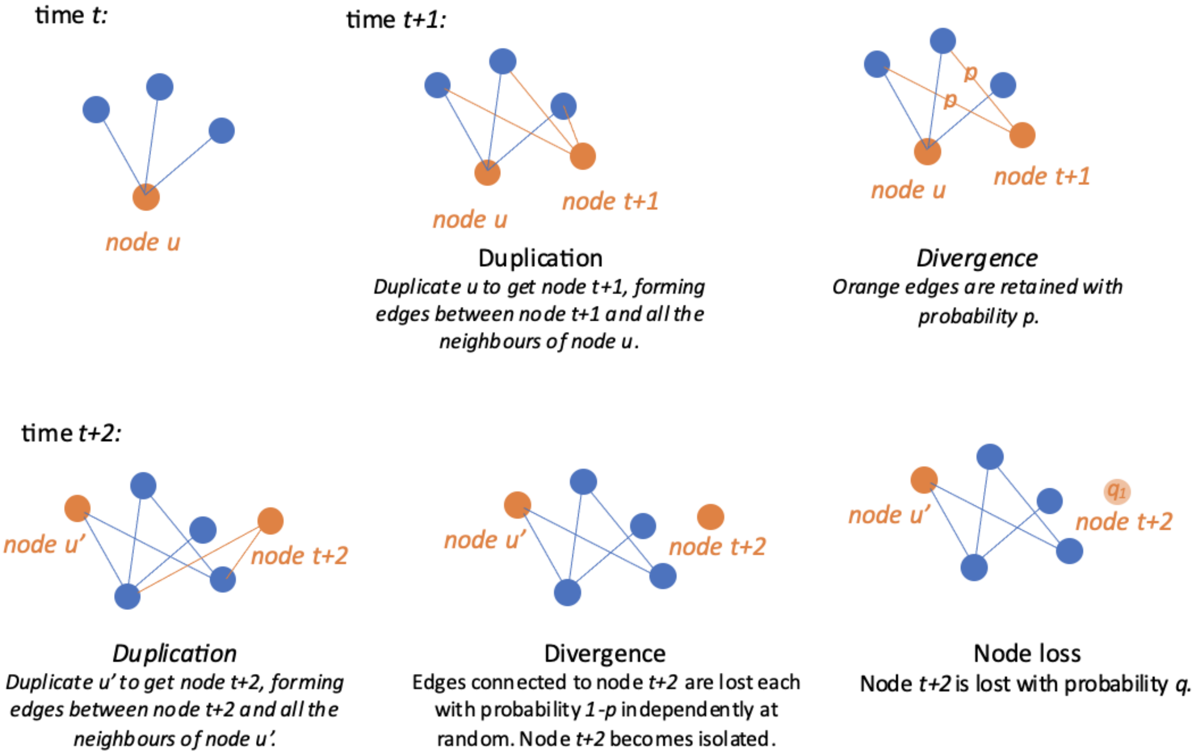

time t:
node u
time t+1:
node u
node t+1
Duplication
Duplicate u to get node t+1, forming edges between node t+1 and all the neighbours of node u.
p
p
node u
node t+1
Divergence
Orange edges are retained with probability p.
time t+2:
node u'
node t+2
Duplication
Duplicate u' to get node t+2, forming edges between node t+2 and all the neighbours of node u'.
node u'
node t+2
Divergence
Edges connected to node t+2 are lost each with probability 1-p independently at random. Node t+2 becomes isolated.
node u'
q1
node t+2
Node loss
Node t+2 is lost with probability q.

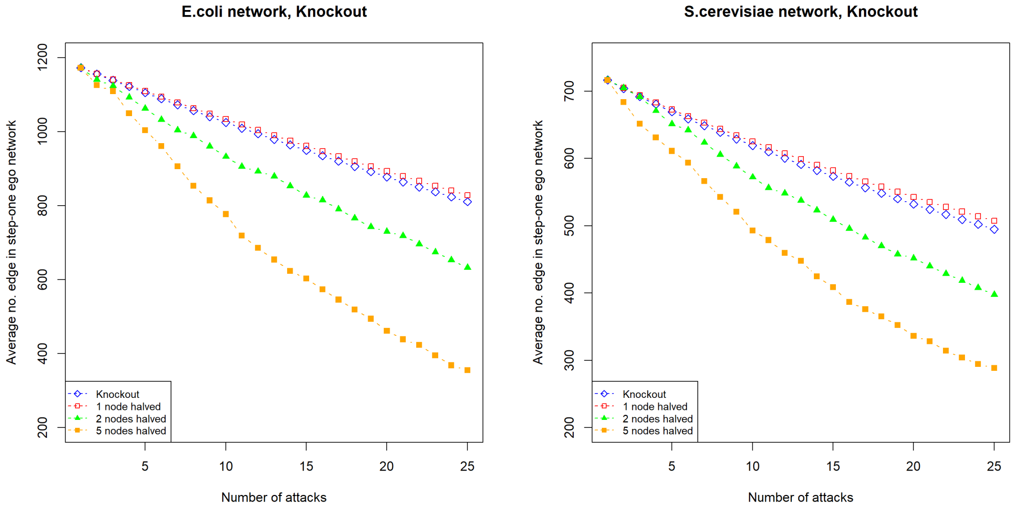
Figure 4. Graph illustration of a new duplication divergence model with node loss.

5. Results

5.1. Simulation of Weak Attacks in Real PPI Networks

Here, we apply the various attack strategies to our PPI networks datasets with 10 repeats. Figure 5 shows that as for the PPI networks of *E. coli* and *S. cerevisiae* the number of targets that are subject to weak attacks increases, and the damage caused by weak attacks becomes larger and is significantly greater than the damage caused by complete knockout.

To understand the expected effects of attacks, a parametric model may be useful. Next, we investigate two such models.



(**a**) *E. coli* (**b**) *S. cerevisiae*

Figure 5. The average number of edges in the 1-step ego network of an *E. coli* and *S. cerevisiea* PPI network after 25 attacks. (**a**) shows the average number of edges in the 1-step ego network in a *E. coli* PPI network under 25 knockout attacks. Blue line: complete knockout; red line: partial knockout with half of the edges connected to one node being removed at each attack; green line: partial knockout with half of the edges connected to two nodes being removed at each attack; orange line: partial knockout with half of the edges connected to five nodes being removed at each attack. (**b**) shows the average number of edges in the 1-step ego network in a *S. cerevisiae* PPI network under 25 attenuation attacks. Since a one-node halved knockout only deletes half of the edges connected to the selected node, when a node has a degree of at least 2 it causes less damage than a complete knockout which removes all the edges connected to the selected node.

5.2. *The Number of Isolated Nodes in a Bernoulli Random Graph*

As a baseline model for a PPI network, we use a $G(n, M)$ model. In Appendix A we derive an upper bound for the total variation distance for the number of isolated nodes in real PPI networks using a $G(n, M)$ graph under Poisson approximation, see Appendix A.1. The Poisson approximation comes with an explicit bound, which we abbreviate here as Δ, on the total variation distance (1). If W denotes the number of isolated nodes, λ its expectation under the $G(n, M)$ model, and Z a Poisson-distributed random variable with mean λ, then it follows that for all k,

$$\mathbb{P}(Z \geq k) - \Delta \leq \mathbb{P}(W \geq k) = \mathbb{P}(Z \geq k) + (\mathbb{P}(W \geq k) - \mathbb{P}(Z \geq k)) \leq \mathbb{P}(Z \geq k) + \Delta.$$

Thus, the Poisson approximation can be used to assess statistical significance.

For our *E. coli* and *S. cerevisiae* data, the estimated upper bound for the total variation distance is 3.73×10^{-15} and 9.28×10^{-19}, respectively. While these bounds are small, the p-values associated with these bounds are 0 up to 6 significant digits under a two-sided test in which the null hypothesis of the $G(n, M)$ model is rejected for very small or very large numbers of isolated nodes, lending evidence to the explanation that the $G(n, M)$ model does not explain the observed number of isolated nodes well. The observed number of isolated nodes in *E. coli* and *S. cerevisiae* is 833 and 1100, respectively, whereas the expected number of isolated nodes under the $G(n, M)$ model is 2.99×10^{-14} and 1.27×10^{-17}. This suggests that a $G(n, M)$ graph may not be suitable for modelling these real PPI networks when the interest is in the number of isolated nodes as a summary statistics.

We further derived upper bounds for the total variation distance under Poisson approximation for the number of isolated nodes after different types of attack, see Appendices A.2–A.4. Again, the results are highly significant, with p-values equal to 0 up to 6 significant digits, indicating that after an attack, the $G(n, M)$ model still does not fit the data well. Hence, a different model for the data is needed. Next, we investigate the standard duplication–divergence model from Section 3.

5.3. Simulation of Weak Attacks in Duplication–Divergence Model

In this section, we present the simulation results of applying weak attacks to realisations of the standard duplication–divergence model $DD(t_0; p)$ from Section 3. The model is undirected, and all edges are set to have unit weight; we take $t_0 = 3$, and start the simulation of the graph with a triangle. This choice ensures that the generated networks can include triangles, resulting in non-zero local and global clustering coefficients; thus they are able to match this key characteristic of PPI networks. In contrast, if the graph is initiated with just a connected pair of nodes, the generated graphs cannot have any triangles; the corresponding simulation results, shown in Appendix B, are, however, similar regarding the effect of attacks. Reference [10] proves that p^* solving the equation $pe^p = 1$ is a critical value, in the sense that for $p > p^*$ there is no limiting degree distribution. In this paper, we take p to be 0.4, a value smaller than $p^* \approx 0.567$. The simulations are run for 1000 steps, with five repeats.

The top two plots of Figure 6 show how partial attacks damage a DD network compared to complete knockout attacks. As illustrated in the top left plot of Figure 6, while increasing the number of nodes being attacked weakly eventually enhances the damage efficiency for a large number of attacks, complete knockout attacks serve as a robust method to destroy the network.

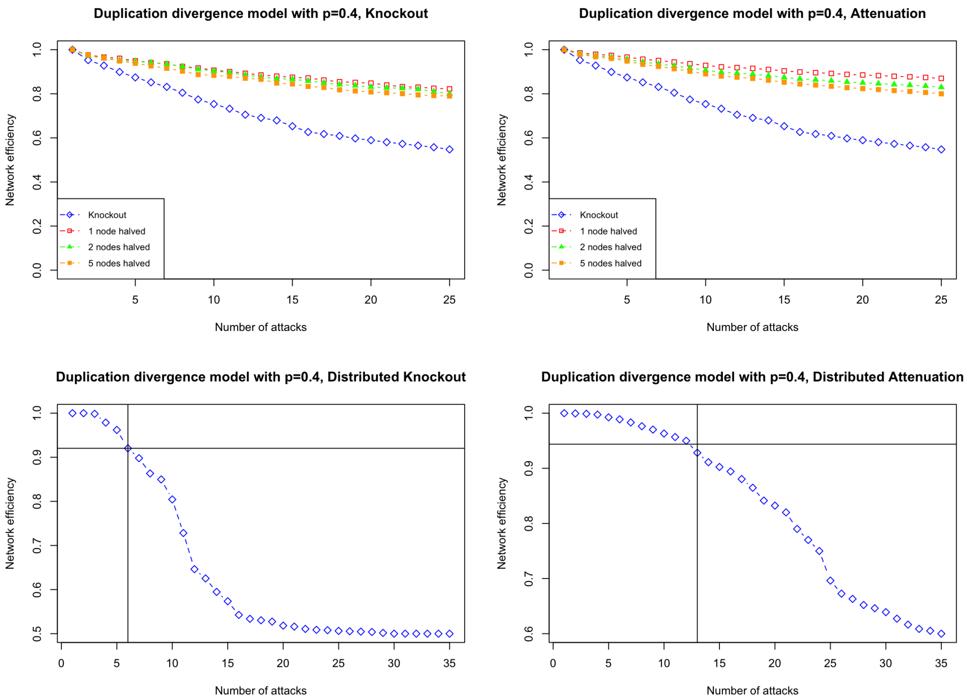


Figure 6. Network efficiency after up to 25 weak attacks on simulations from the duplication–divergence model starting with a triangle with a divergence rate $p = 0.4$. **Top left:** knockout attacks. Blue line: complete knockout; red line: partial knockout with half of the edges connected to one node being removed at each attack; green line: partial knockout with half of the edges connected to two nodes being removed at each attack; orange line: partial knockout with half of the edges connected

to five nodes being removed at each attack. **Top right:** attenuation attacks. Blue line: complete knockout; red line: partial attenuation with all the edges connected to one node being halved at each attack; green line: partial attenuation with all the edges connected to two nodes being halved at each attack; orange line: partial attenuation with all the edges connected to five nodes being halved at each attack. **Bottom left**: distributed attacks, with edges drawn from a random distribution; the horizontal line represents equivalent damage to the network achieved by one complete knockout. **Bottom right**: distributed attenuation attacks, with the weight of edges drawn from a random distribution to be halved; the horizontal line represents equivalent damage to the network achieved by one complete knockout.

The bottom two plots of Figure 6 show how distributed attacks damage a DD network compared to complete knockout attacks. The horizontal line representing the damage caused by one complete knockout suggests that the effect of 6 distributed knockout or 13 distributed attenuation attacks is equivalent to the effect of one complete knockout. This indicates that distributed attacks are less effective than both complete knockout attacks and partial attacks.

5.4. Simulation of Weak Attacks in the New Node Loss Model

Now, we present the simulation results of applying weak attacks onto the new node loss model introduced in Section 4. Again, the model is undirected with all edges assigned unit weight. The simulations are run with 10 repeats and the average network efficiency values are reported to account for randomness. We note here that we did not carry out a grid search for the optimal parameter choices for the DD models without and with gene loss for the different organisms, as the focus of this paper is the qualitative behaviour of the new DD model with gene loss, and not detailed modelling of observed PPI networks.

Figure 7 shows the results for $p = 0.4$ and $q = 0.2$ under different weak attacks. We observe that in 25 attacks, a complete knockout attack is more effective than a partial attenuation when half of the edges connected to one node are eliminated, but less effective than a partial attenuation when halving two nodes or five nodes. Our results indicate that as the number of halved nodes increases, the weak attacks damage networks more efficiently. Furthermore, distributed attacks are less effective than complete knockout and partial attacks, mirroring the qualitative impact observed in real PPI networks.

We observe that the pattern of Figure 7 for the new node loss model is more similar to the pattern of Figure 5 for the real datasets than the pattern of Figure 6 for a standard DD model. This suggests that the new node loss model can mimic the effect of weak attacks on protein–protein interaction networks more realistically than the standard $DD(t_0, p)$ model.

Regarding the effect of the probability of node loss on weak attacks in the new node loss model, we notice that the number of distributed attacks required to achieve the equivalent effect as one complete knockout attack increases as q increases. This raises a natural question regarding how the value of q affects the efficiency of weak attacks in the new node loss model. In our simulations, shown in Figure 8, the resilience of the new node loss model to weak attacks results in a slower rate of network degradation. This can be attributed to the fact that higher q values correspond to an increased likelihood of losing isolated nodes, which in turn leads to a more connected graph structure.

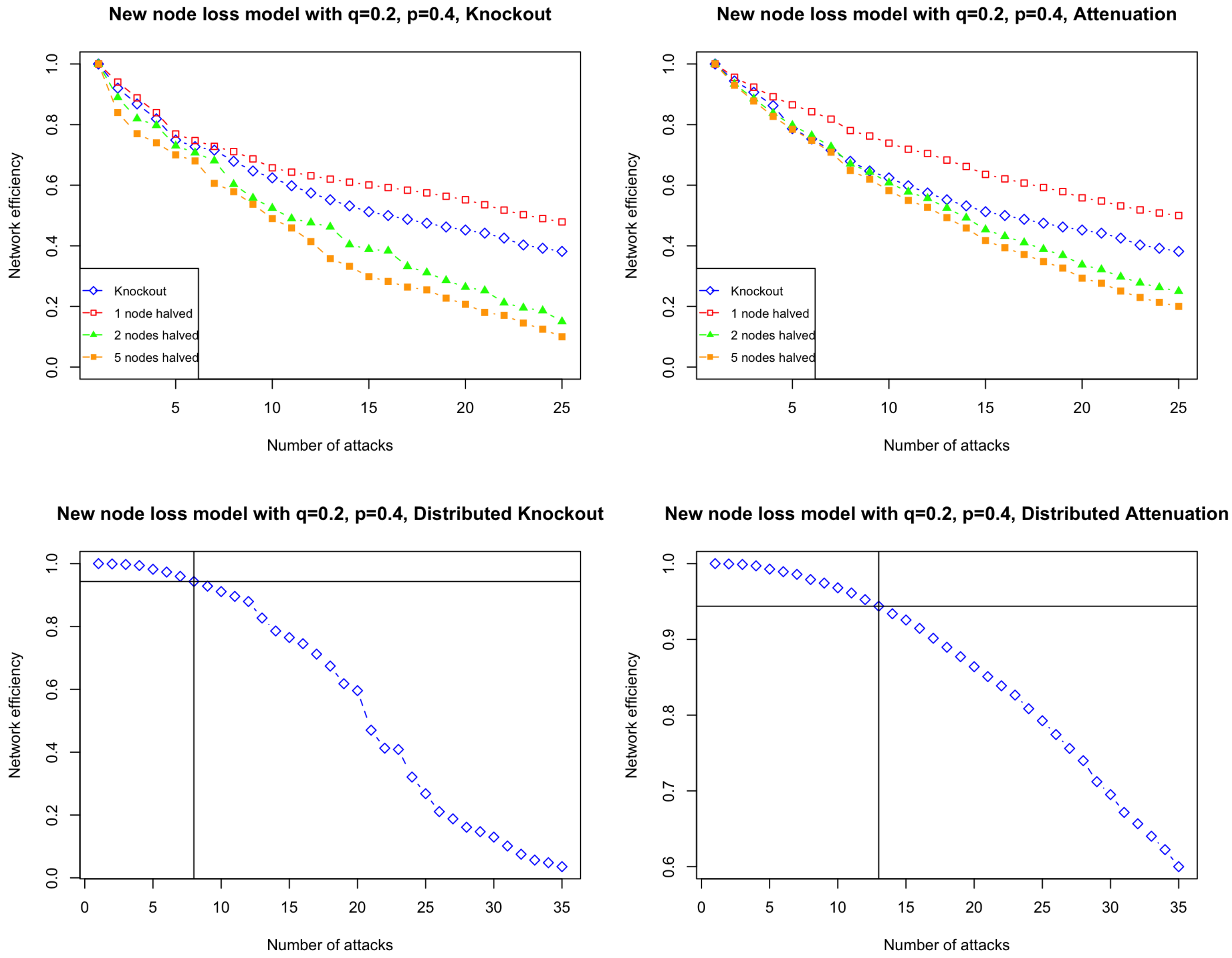
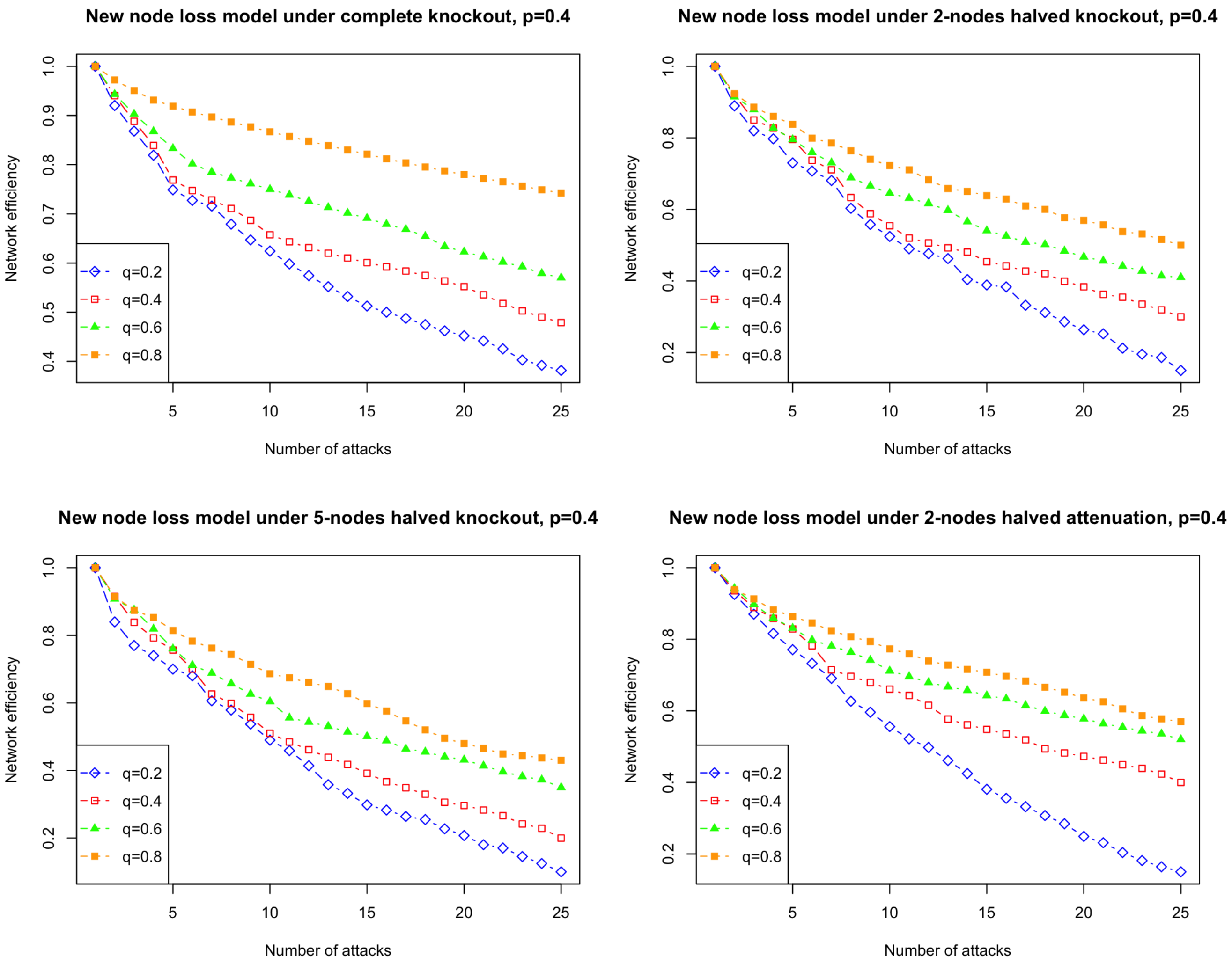


Figure 7. Network efficiency after up to 25 weak attacks on simulations from the new node loss model starting with a triangle; a node can be lost with probability $q = 0.2$, using a divergence rate $p = 0.4$. The graph is undirected and has unit edge weight. **Top left:** knockout attacks. Blue line: complete knockout; red line: partial knockout with half of the edges connected to one node being removed at each attack; green line: partial knockout with half of the edges connected to two nodes being removed at each attack; orange line: partial knockout with half of the edges connected to five nodes being removed at each attack. **Top right:** attenuation attacks. Blue line: complete knockout; red line: partial attenuation with all the edges connected to one node being halved at each attack; green line: partial attenuation with all the edges connected to two nodes being halved at each attack; orange line: partial attenuation with all the edges connected to five nodes being halved at each attack. **Bottom left**: distributed attacks, with edges drawn from a random distribution; the horizontal line represents equivalent damage to the network achieved by one complete knockout. **Bottom right**: distributed attenuation attacks, with the weight of edges drawn from a random distribution to be halved; the horizontal line represents equivalent damage to the network achieved by one complete knockout.



New node loss model under complete knockout, p=0.4

New node loss model under 2-nodes halved knockout, p=0.4

New node loss model under 5-nodes halved knockout, p=0.4

New node loss model under 2-nodes halved attenuation, p=0.4

Figure 8. Effect of q on the efficiency of weak attacks on simulated networks from the node loss model starting from a triangle with $p = 0.4$, and q ranges from 0.2, 0.4, 0.6, to 0.8.

6. Discussion

In this paper, we have assessed standard models for PPI networks and we have introduced a new node loss model which is motivated by observed gene loss in organisms. We show that our new node loss model captures the effect of weak attacks in a protein–protein interaction network more realistically than a standard DD model (i.e., $q = 0$).

To further enhance the robustness of our results, as future work we aim to derive analytical results for the average number of edges in a 1-step ego network and for the network efficiency before and after attacks in the new node loss model.

It is perhaps not surprising that the new node loss model performs better due to its incorporation of a natural and common biological adaptation, namely, gene loss, occurring throughout evolution. As a next step, variants of the new node loss model could be examined; for example, one could include the case where the probability of a parent–child node edge is not zero. In order to understand how node loss affects duplication–divergence behaviour, we also aim to investigate other parameters that can affect a node loss in a network; for example, a pair of nodes may be more likely to be lost if they are connected by an isolated edge.

Regarding the network representation of PPIs, we chose the PPI networks from the STRING database, which represents each protein-coding gene locus by only a single, representative protein. The datasets contain non-binary data which could be incorporated in the analysis. Moreover, future work will assess the effect of restricting the protein interactions from the STRING database to physical interactions, by repeating the analysis for the full STRING PPI networks. Hypergraph representations as in [17] may also be fruitful.

Author Contributions: Writing—original draft, R.Z. and G.R.; supervision, G.R. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to acknowledge a discussion with Alan Whitmore and Jonny Wray at e-Therapeutics which sparked this project. G.R. acknowledges supports from EPSRC grants EP/T018445/1, EP/W037211/1, EP/V056883/1, and EP/R018472/1.

Data Availability Statement: The data presented in this study are openly available in GitHub at https://github.com/rh-zhang/Entropy_CNC2023.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Poisson Approximation for the Number of Isolated Nodes in a $G(n, M)$ Graph Before and After Attacks

Appendix A.1. Poisson Approximation for the Number of Isolated Nodes in a $G(n, M)$ Graph

For a $G(n, M)$ graph G, $n \geq 2$, we define the edge indicators E_{ij} so that $E_{ij} = 1$ if there is an edge between i and j belonging to the edge set $E(G)$ of G, and 0 otherwise. These edge indicators are not independent, as can be seen by the requirement that $\sum_{i<j} E_{ij} = M$. Since M edges are chosen uniformly at random from $N = \binom{n}{2}$ possible edges, we have

$$\mathbb{P}(E_{ij} = 1) = \frac{M}{N} := p$$

so that $E_{ij} \sim Be(p)$. We let

$$I_i : I_i(n) = \prod_{j \neq i}(1 - E_{ij})$$

be the indicator of the event that node i is isolated in $G(n, M)$. Then, for $M \leq N - (n-1)$,

$$\mathbb{P}(I_i = 1) = \frac{\binom{N-(n-1)}{M}}{\binom{N}{M}} := \pi \tag{A1}$$

is the same for each i, and for $M > N - (n-1)$ it is 0. Our quantity of interest is $W = \sum_{i=1}^{n} I_i$, the number of isolated nodes in $G(n, M)$. From Equation (A1), $\lambda = \mathbb{E}(W) = n\pi$. We point out that I_i's are not independent, but when π is small the dependence is weak.

While Equation (A1) can be difficult to evaluate numerically for large N and M, we note that

$$\pi = \frac{\binom{N-(n-1)}{M}}{\binom{N}{M}} = \left(1 - \frac{M}{N}\right)\left(1 - \frac{M}{N-1}\right)\cdots\left(1 - \frac{M}{N-n+2}\right).$$

Thus, setting $p = \frac{M}{N}$, we can bound $n(1 - \frac{M}{N-n})^{n-1} \leq \lambda \leq n(1 - \frac{M}{N})^{n-1} = n(1-p)^{n-1}$. To understand the distribution of W, Theorem 1 in the main text gives a Poisson approximation for which we provide a proof here. For convenience, we re-state the result.

Theorem A1. *If* $W := \sum_{i=1}^{n} I_i$*, we have for a* $G(n, M)$ *graph,*

$$d_{TV}(\mathcal{L}(W); Po(\lambda)) \leq \min(1, \lambda^{-1}) e^{-np(1+\frac{n-2}{N+2-n})+p(1+\frac{n-2}{N+2-n})+\frac{2-3n+n^2}{N+2-n}} \left(1 + np\left(1 + \frac{n + Np - 2}{N - Np - n + 2}\right)\right).$$

Before we prove this result, we note that the bound, which we could call Δ as in Section 5.2, on the total variation distance is explicit; no limiting behaviour is assumed. However, it can be seen that Δ tends to 0 as $p = \dfrac{M}{N} \to 1$.

Proof. When assessing the goodness of fit of Poisson approximations, Stein's method has become a strong tool under various dependence structures [18]. In our case, notice that given any realisation of $G(n, M)$, an associated realisation of $G(n, M)$ conditional on $I_i = 1$ is obtained simply by setting all the edge indicators $(E_{ij}, E_{ji}, i \leq j \leq n, j \neq i)$ equal to zero. This may create additional isolated nodes, but cannot destroy any. To exploit this fact, we use so-called *size bias coupling*, constructing a random variable W_i^* in the same probability space as W, which has the conditional distribution $\mathcal{L}(W - I_i \mid I_i = 1)$. Theorem 2.A in [19] gives that

$$d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq J_1 \sum_{i=1}^{n} p_i \mathbb{E}|W - W_i^*| \tag{A2}$$

with $J_1 \leq \min\{1, \lambda^{-1}\}$.

To construct such size bias coupling, as in [20], we introduce $Z_j = \prod_{l \neq i,j}(1 - E_{jl})$, so that $Z_j = 1$ if j is not connected to any nodes excluding i and itself, and $Z_j = 0$ otherwise. Then, for each i we can take as a size-biased variable

$$\begin{aligned} W_i^* &= \sum_{j:j \neq i} Z_j \\ &= \sum_{j:j \neq i} (1 - E_{ij} + E_{ij}) \prod_{l \neq i,j} (1 - E_{jl}) \\ &= \sum_{j:j \neq i} \prod_{l \neq j} (1 - E_{jl}) + \sum_{j:j \neq i} U_j \\ &= W - I_i + \sum_{j:j \neq i} U_j, \end{aligned}$$

where

$$U_j = \prod_{l \neq i,j} (1 - E_{jl}) - \prod_{l \neq i,j} (1 - E_{jl})(1 - E_{ji}) = E_{ij} \prod_{l \neq i,j} (1 - E_{jl}).$$

For a $G(n, M)$ graph, we have

$$\mathbb{E}\left[\sum_{j=1, j \neq i}^{n} E_{ij} \prod_{l \neq i,j} (1 - E_{jl})\right] = \frac{\binom{n-1}{1}\binom{N-(n-1)}{M-1}}{\binom{N}{M}}$$

since to make sure node j is only connected to node i, we need an edge between i and j chosen from $n - 1$ nodes, and all the other $M - 1$ edges are chosen from the edge set excluding j. Hence,

$$\mathbb{E}|W_i^* - W| \leq \frac{(n-1)\binom{N-(n-1)}{M-1}}{\binom{N}{M}} + \frac{\binom{N-(n-1)}{M}}{\binom{N}{M}}. \tag{A3}$$

Therefore, with $p = \frac{M}{N}$ and $N = \binom{n}{2}$, by Equation (A2) we have

$$\begin{aligned} d_{TV}(\mathcal{L}(W); Po(\lambda)) &= \min(1, \lambda^{-1})\lambda \left\{ (n-1)\frac{\binom{N-(n-1)}{M-1}}{\binom{N}{M}} + \frac{\binom{N-(n-1)}{M}}{\binom{N}{M}} \right\} \\ &\leq \min(1, \lambda^{-1}) e^{-np(1+\frac{n-2}{N+2-n})+p(1+\frac{n-2}{N+2-n})+\frac{2-3n+n^2}{N+2-n}} \\ &\quad \left(1 + np\left(1 + \frac{n+Np-2}{N-Np-n+2}\right)\right), \end{aligned}$$

where the last step follows from standard inequalities. □

This bound tends to 0 as $p = \dfrac{M}{N} \to 1$ as long as $M \leq N - n + 1$.

Appendix A.2. Poisson Approximation for the Number of Isolated Nodes in a $G(n, M)$ Graph after One Complete Knockout Attack

A complete knockout attack removes all the edges of a randomly picked node U. Assume that $U = i$. Let $N' = \binom{n-1}{2}$, set $I_j = \mathbb{1}(j \text{ is isolated in the graph before the attack})$, and

$$I'_j = \mathbb{1}(j \text{ is isolated in the graph after the attack}).$$

Before the attack, denoting by $deg(i)$ the degree of i we have

$$\mathbb{P}_{G(n,M)}(deg(i) = k) = \frac{\binom{n-1}{k}\binom{N-(n-1)}{M-k}}{\binom{N}{M}} := p_{deg_k,n}, \tag{A4}$$

for $k \leq \min(n-1, M)$ and 0 otherwise. Suppose that node i is attacked. Then, for $j \neq i$ and $k \leq \min(n-1, M)$, as the edges in $G(n, M)$ are distributed uniformly, if the attacked node has degree k then the graph after the attack is a $G(n-1, M-k)$ graph. Hence, for $k \leq \min(n-1, M)$,

$$\mathbb{P}(I'_j = 1 | deg(i) = k) = \frac{\binom{N'-(n-2)}{M-k}}{\binom{N'}{M-k}} =: \pi_k(n-1),$$

which is the same for each $j \neq i$ in the graph after the attack. Now, let W' be the number of isolated nodes after one attack. We have

$$\begin{aligned} \mathbb{E}(W') &= \frac{1}{n}\sum_{v=1}^{n} \mathbb{E}(W' | v \text{ is the vertex for duplication}) \\ &= \frac{1}{n}\sum_{v=1}^{n}\sum_{k=0}^{\min(n-1,M)} \mathbb{P}(deg(v) = k)\mathbb{E}(W' | v \text{ is the vertex for duplication}, deg(v) = k) \\ &= \sum_{k=0}^{\min(n-1,M)} \lambda_k p_{deg_k,n} \end{aligned}$$

where

$$\lambda_k := \mathbb{E}(W' | v \text{ is the vertex for duplication}, deg(v) = k).$$

Let Λ be a random variable taking values in $\lambda_k, k = 1, \ldots, \min(n-1, M)$, with

$$\mathbb{P}(\Lambda = \lambda_k) = p_{deg_k,n}, \tag{A5}$$

using the notation (A4). Then , $E\Lambda = \sum_{k=0}^{\min(n-1,M)} p_{deg_k}(v)\lambda_k = E(W')$. We now approximate the distribution of W' by a mixed Poisson distribution. Let $Z \sim Po(\Lambda)$. Then, for any function h,

$$\mathbb{E}h(W') - \mathbb{E}h(Z) = \sum_k \{\mathbb{E}(h(w)|deg(U) = k) - \mathbb{E}h(Z_k)\} p_{deg_k,n}, \tag{A6}$$

where $Z_k \sim Po(\lambda_k)$. For each choice of k, bounding $|\mathbb{E}(h(w)|deg(U) = k) - \mathbb{E}h(Z_k)|$ can then be carried out in a similar vein as for Theorem 1, as follows.

Theorem A2. *For a $G(n, M)$ graph after one complete knockout attack, we have*

$$d_{TV}(\mathcal{L}(W'); Po(\Lambda)) \leq \sum_{k=0}^{\min(n-1,M)} \frac{\binom{n-2}{k}\binom{N'-(n-2)}{N'p_k-k}}{\binom{N'}{N'p_k}} \min(1, \lambda_k^{-1}) \\ e^{-np_k(1+\frac{n-3}{N'+3-n})+p_k(2+\frac{n-3}{N'+3-n})+\frac{6-5n+2n^2}{N'+3-n}} \\ \left(1 + np_k\left(1 + \frac{n + N'p_k - 3}{N' - n - N'p_k + 3}\right)\right).$$

where $W' := \sum_{j=1}^{n-1} I'_j(n-1)$, $p_k = \frac{M-k}{N'}$, and Λ given in (A5).

Proof. After one attack on node i of degree k the graph is a realisation of the $G(n-1, M-k)$ model, together with an isolated node i. Again, we use size bias coupling. Given any realisation of $G(n-1, M-k)$, an associated realisation of $G(n-1, M-k)$ conditional on $I_j = 1$ is obtained simply by setting all the edge indicators $(E_{lj}, E_{jl}, l \leq j \leq n-1, j \neq l)$ equal to zero. This may create additional isolated nodes, but cannot destroy any. By (A3), we have for $k \leq \min(n-1, M)$

$$\mathbb{E}\Big[|W_j^* - W'|\Big|deg(i) = k\Big] \leq \frac{(n-2)\binom{N'-(n-2)}{M-k-1}}{\binom{N'}{M-k}} + \frac{\binom{N'-(n-2)}{M-k}}{\binom{N'}{M-k}}. \tag{A7}$$

Setting $p_k = \frac{M-k}{N'}$, using (A6) we have

$$d_{TV}(\mathcal{L}(W'); Po(\Lambda)) \\ \leq \sum_{k=0}^{\min(n-1,M)} p_{deg_k,n-1} \min(1, \lambda_k^{-1}) \left\{ (n-2)\frac{\binom{N'-(n-2)}{M-1-k}}{\binom{N'}{M-k}} + \frac{\binom{N'-(n-2)}{M-k}}{\binom{N'}{M-k}} \right\} \\ \leq \sum_{k=0}^{\min(n-1,M)} \frac{\binom{n-2}{k}\binom{N'-(n-2)}{N'p_k-k}}{\binom{N'}{N'p_k}} \min(1, \lambda_k^{-1}) e^{-np_k(1+\frac{n-3}{N'+3-n})+p_k(2+\frac{n-3}{N'+3-n})+\frac{6-5n+2n^2}{N'+3-n}} \\ \left(1 + np_k\left(1 + \frac{n + N'p_k - 3}{N' - n - N'p_k + 3}\right)\right).$$

□

To further bound this bound, we could bound $p_{deg_k,n-1}$ by $\max_k p_{,n-1}$. More crudely, we can bound

$$d_{TV}(\mathcal{L}(W'); Po(\Lambda)) \leq \sum_{k=0}^{\min(n-1,M)} e^{-np_k(1+\frac{n-3}{N'+3-n})+p_k(2+\frac{n-3}{N'+3-n})+\frac{6-5n+2n^2}{N'+3-n}} \\ \left(1 + np_k\left(1 + \frac{n + N'p_k - 3}{N' - n - N'p_k + 3}\right)\right).$$

Appendix A.3. Poisson Approximation for the Number of Isolated Nodes in a $G(n, M)$ Graph after One Partial Knockout Attack

A partial knockout attack on node i randomly removes half of its edges with other nodes. Here, we round down the number of edges removed, which means node i would never be isolated after one attack if its degree is 0 or 1 in the original graph $G(n, M)$. So, if the node i has degree k before the attack, then $\lfloor k/2 \rfloor$ edges are removed. Again, we let

$$I'_j = \mathbb{1}(j \text{ is isolated in the graph after the attack}).$$

Theorem A3. *For a $G(n, M)$ graph after one partial knockout attack, we have*

$$\begin{aligned} & d_{TV}(\mathcal{L}(W'); Po(\Lambda)) \\ & \leq \sum_{k=0}^{n-2} \frac{\binom{n-2}{k}\binom{N'-(n-2)}{N'p_k-k}}{\binom{N'}{N'p_k}} \min(1, \lambda_k^{-1}) \left(1 + np_k\left(1 + \frac{n + N'p_k - 3}{N' - n - N'p_k + 3}\right)\right) \\ & \quad e^{-np_k(1+\frac{n-3}{N'+3-n})+p_k(2+\frac{n-3}{N'+3-n})+\frac{6-5n+2n^2}{N'+3-n}}, \end{aligned}$$

where $W' := \sum_{j=1}^{n-1} I'_j(n-1)$, $p_k = \frac{M-\lfloor \frac{k}{2} \rfloor}{N'}$, and Λ is given in (A5).

Proof. Firstly, letting $deg(i)$ denote the degree of node i in $G(n, m)$,

$$\begin{aligned} \mathbb{P}\Big(I'_i = 1 \Big| deg(i) = k\Big) &= \mathbb{P}\Big(I'_i = 1 \Big| deg(i) > 0\Big)\mathbb{P}\Big(deg(i) > 0\Big) + \\ & \quad \mathbb{P}\Big(I'_i = 1 \Big| deg(i) = 0\Big)\mathbb{P}\Big(deg(i) = 0\Big) \\ &= 0 + \pi = \pi \end{aligned}$$

and

$$\mathbb{P}(I'_j = 1) = \textstyle\sum_{k=1}^{n-1} p_{deg_k,n} \mathbb{P}\Big(I'_j = 1 \Big| deg(i) = k\Big) \text{ for } j \neq i.$$

In particular, we have

$$\begin{aligned} & \mathbb{P}\Big(I'_j = 1 \Big| deg(i) = k\Big) \\ & = \mathbb{P}\Big(I_j = 1 \Big| deg(i) = k\Big) \\ & + \mathbb{P}\Big(i \sim j, deg(j) = 1 \Big| deg(i) = k\Big)\mathbb{P}\Big(i \sim j \text{ is deleted } \Big| i \sim j, deg(j) = 1, deg(i) = k\Big). \end{aligned} \tag{A8}$$

For $k \leq min(n-1, M)$,

$$\begin{aligned} \mathbb{P}\Big(I_j = 1 \Big| deg(i) = k\Big) &= \frac{\binom{N-k-(n-1)}{M-k}}{\binom{N}{M}}; \\ \mathbb{P}\Big(i \sim j, deg(j) = 1 \Big| deg(i) = k\Big) &= \frac{\binom{n-2}{k-1}\binom{N-(2(n-2)+1)}{M-k}}{p_{deg_k,n}\binom{N}{M}}. \end{aligned}$$

Also,

$$\mathbb{P}\Big(i \sim j \text{ is deleted} \Big| i \sim j, deg(j) = 1, deg(i) = k\Big) = \begin{cases} \frac{1}{2} & \text{if } k \text{ is even} \\ \frac{k-1}{2k} & \text{if } k \text{ is odd.} \end{cases}$$

Therefore,

$$\mathbb{P}\Big(I'_j = 1\Big|deg(i) = k\Big) = \begin{cases} \frac{\binom{N-k-(n-1)}{M-k}}{\binom{N}{M}} + \frac{\binom{n-2}{k-1}\binom{N-(2(n-2)+1)}{M-k}}{p_{deg_k,n}\binom{N}{M}}\frac{1}{2} & \text{if } k \text{ is even} \\ \frac{\binom{N-k-(n-1)}{M-k}}{\binom{N}{M}} + \frac{\binom{n-2}{k-1}\binom{N-(2(n-2)+1)}{M-k}}{p_{deg_k,n}\binom{N}{M}}\frac{k-1}{2k} & \text{if } k \text{ is odd} \end{cases}$$

where for each case the first term represents the probability before attack and the second term represents the probability after attack.

Then, $\lambda_k := \mathbb{E}[W'|v \text{ is the vertex picked for duplication}, deg(v) = k]$ is

$$\lambda_k = \begin{cases} \pi + (n-1)\left(\frac{\binom{N-k-(n-1)}{M-k}}{\binom{N}{M}} + \frac{\binom{n-2}{k-1}\binom{N-(2(n-2)+1)}{M-k}}{p_{deg_k,n}\binom{N}{M}}\frac{1}{2}\right) & \text{if } k \text{ is even} \\ \pi + (n-1)\left(\frac{\binom{N-k-(n-1)}{M-k}}{\binom{N}{M}} + \frac{\binom{n-2}{k-1}\binom{N-(2(n-2)+1)}{M-k}}{p_{deg_k,n}\binom{N}{M}}\frac{k-1}{2k}\right) & \text{if } k \text{ is odd.} \end{cases}$$

After one partial knockout attack, the graph is a realisation of the $G(n-1, M - \lfloor \frac{k}{2} \rfloor)$ model, combined with node i and its remaining edges if i does not become isolated, or the graph is a realisation of the $G(n-1, M - \lfloor \frac{k}{2} \rfloor)$ model combined with an isolated i if i becomes isolated. Conditioning on the different cases,

$$\begin{aligned} &\mathbb{E}h(W') - \mathbb{E}h(Z) \\ &= \sum_{k=0}^{n-1} \{\mathbb{E}[h(W')|deg(i) = k, i \text{ becomes isolated})\mathbb{P}(i \text{ becomes isolated}) \\ &+ \mathbb{E}[h(W')|deg(i) = k, i \text{ does not become isolated})\mathbb{P}(i \text{ does not become isolated}) \\ &- \mathbb{E}h(Z_k)\}p_{deg_k,n-1} \end{aligned}$$

where

$$Z_k = \begin{cases} Z_k^{iso} \sim Po(\lambda_k) + 1 & \text{if } v \text{ becomes isolated} \\ Z_k^{non-iso} \sim Po(\lambda_k) & \text{if } v \text{ does not become isolated.} \end{cases}$$

Hence,

$$\begin{aligned} \mathbb{E}h(W') - \mathbb{E}h(Z) = &\sum_{k=0}^{n-1} p_{deg_k,n}\Big\{\Big(\mathbb{E}[h(W')|deg(i) = k, i \text{ becomes isolated}] - \mathbb{E}h\Big(Z_k^{iso}\Big)\Big) \\ &\mathbb{P}(i \text{ becomes isolated}) \\ &+ \Big(\mathbb{E}[h(W')|deg(i) = k, i \text{ does not become isolated}] - \mathbb{E}h\Big(Z_k^{non-iso}\Big)\Big) \\ &\mathbb{P}(i \text{ does not become isolated})\Big\} \\ = &\sum_{k=0}^{n-2} p_{deg_k,n-1}\Big\{\Big(\mathbb{E}[g(W')|deg(i) = k, i \text{ becomes isolated}] - \mathbb{E}g\Big(Z_k^{iso}\Big)\Big) \\ &\mathbb{P}(i \text{ becomes isolated}) \\ &+ \Big(\mathbb{E}[g(W')|deg(i) = k, i \text{ does not become isolated}] - \mathbb{E}g\Big(Z_k^{non-iso}\Big)\Big) \\ &\mathbb{P}(i \text{ does not become isolated})\Big\} \end{aligned}$$

where $g(x) = h(x+1)$. Now, for each of the two cases we apply size bias coupling as in Theorem A2. We approximate $\mathbb{E}[g(W')|deg(i) = k, i \text{ becomes isolated}]$ by $\mathbb{E}[g(Z_k^{iso})]$ and we approximate $\mathbb{E}[g(W')|deg(i) = k, i \text{ does not become isolated}]$ by $\mathbb{E}[g(Z_k^{non-iso})]$.

Combining the bounds (A7) for the two cases,

$$
\begin{aligned}
&\mathbb{E}[|W' - W| deg(i) = k] \qquad \text{(A9)}\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{n-2} p_{deg_k,n-1}\left\{\left(\frac{(n-2)\binom{N'-(n-2)}{M-\lfloor k/2\rfloor-1}}{\binom{N'}{M-\lfloor k/2\rfloor}} + \frac{\binom{N'-(n-2)}{M-\lfloor k/2\rfloor}}{\binom{N'}{M-\lfloor k/2\rfloor}}\right)\right.\\
&\quad \mathbb{P}(i \text{ becomes isolated})\\
&\quad + \left(\frac{(n-2)\binom{N'-(n-2)}{M-\lfloor k/2\rfloor-1}}{\binom{N'}{M-\lfloor k/2\rfloor}} + \frac{\binom{N'-(n-2)}{M-\lfloor k/2\rfloor}}{\binom{N'}{M-\lfloor k/2\rfloor}}\right)\mathbb{P}(i \text{ does not become isolated})\\
&\leq \sum_{k=0}^{n-2} p_{deg_k,n-1}\left(\frac{(n-2)\binom{N'-(n-2)}{M-\lfloor k/2\rfloor-1}}{\binom{N'}{M-\lfloor k/2\rfloor}} + \frac{\binom{N'-(n-2)}{M-\lfloor k/2\rfloor}}{\binom{N'}{M-\lfloor k/2\rfloor}}\right) \qquad \text{(A10)}
\end{aligned}
$$

Letting $p_k = \frac{M-\lfloor \frac{k}{2}\rfloor}{N'}$, we have

$$
\begin{aligned}
&d_{TV}(\mathcal{L}(W'); Po(\Lambda)) \qquad \text{(A11)}\\
&\leq \sum_{k=0}^{n-2} p_{deg_k,n-1}\min(1,\lambda_k^{-1})\left\{(n-2)\frac{\binom{N'-(n-2)}{M-\lfloor\frac{k}{2}\rfloor-1}}{\binom{N'}{M-\lfloor\frac{k}{2}\rfloor}} + \frac{\binom{N'-(n-2)}{M-\lfloor\frac{k}{2}\rfloor}}{\binom{N'}{M-\lfloor\frac{k}{2}\rfloor}}\right\}\\
&= \sum_{k=0}^{n-2}\frac{\binom{n-2}{k}\binom{N'-(n-2)}{N'p_k-k}}{\binom{N'}{N'p_k}}\min(1,\lambda_k^{-1})\left\{e^{-np_k(1+\frac{n-3}{N'+3-n})+p_k(2+\frac{n-3}{N'+3-n})+\frac{6-5n+2n^2}{N'+3-n}}\right.\\
&\quad \left.\left(1+np_k\left(1+\frac{n+N'p_k-3}{N'-n-N'p_k+3}\right)\right)\right\}.
\end{aligned}
$$

□

Again, we can use the crude upper bound

$$
\begin{aligned}
d_{TV}(\mathcal{L}(W'); Po(\Lambda)) \leq \sum_{k=0}^{n-1} & e^{-np_k(1+\frac{n-3}{N'+3-n})+p_k(2+\frac{n-3}{N'+3-n})+\frac{6-5n+2n^2}{N'+3-n}}\\
&\left(1+np_k\left(1+\frac{n+N'p_k-3}{N'-n-N'p_k+3}\right)\right), \qquad \text{(A12)}
\end{aligned}
$$

which tends to 0 as $p = \dfrac{M}{N} \to 1$.

Appendix A.4. Poisson Approximation for the Number of Isolated Nodes in a $G(n, M)$ Graph after One Distributed Knockout Attack

A distributed knockout attack on node i of degree k randomly removes its edges with other nodes according to a random distribution.

Theorem A4. *In a $G(n, M)$ graph, we have for W' the number of isolated nodes after one distributed knockout attack,*

$$
\begin{aligned}
d_{TV}(\mathcal{L}(W'); Po(\Lambda)) \leq \sum_{k=0}^{n-2}\sum_{x=0}^{k} & \frac{\binom{n-2}{k}\binom{N'-(n-2)}{N'p_x-k}}{\binom{N'}{N'p_x}}\binom{k}{x}q^x(1-q)^{k-x}\min(1,\lambda_k^{-1})\\
& e^{-np_x\left(1+\frac{n-3}{N'+3-n}\right)}e^{p_x\left(2+\frac{n-3}{N'+3-n}+\frac{6-5n+2n^2}{N'+3-n}\right)}\\
&\left(1+np_x\left(1+\frac{n+N'p_x-3}{N'-n-N'p_x+3}\right)\right).
\end{aligned}
$$

where $W' := \sum_{j=1}^{n-1} I'_j(n-1)$, $\lambda_k = q^k + (n-1)\left(\frac{\binom{N-k-(n-1)}{M-k}}{\binom{N}{M}} + \frac{\binom{n-2}{k-1}\binom{N-(2(n-2)+1)}{M-k}}{p_{deg_k,n}\binom{N}{M}}q\right)$, $p_x = \frac{M-x}{N'}$, *and* Λ *is given in* (A5).

Proof. Let $X_k \sim Bin(k,q)$ denote the number of removed edges if the attacked node i has degree k. Let I'_j denote the indicator that node j is isolated after the attack. Then, after one attack, $\mathbb{P}(I'_i = 1|deg(i) = k) = q^k$, and for $j \neq i$,

$$\mathbb{P}(I'_j = 1) = \sum_{k=1}^{n-1} p_{deg_k,n-1}\mathbb{P}\Big(I'_j = 1\Big|deg(i) = k\Big)$$

In particular,

$$\mathbb{P}\Big(I_j = 1\Big|deg(i) = k\Big) = \frac{\binom{N-k-(n-1)}{M-k}}{\binom{N}{M}},$$
$$\text{and } \mathbb{P}\Big(i \sim j, deg(j) = 1\Big|deg(i) = k\Big) = \frac{\binom{n-2}{k-1}\binom{N-(2(n-2)+1)}{M-k}}{p_{deg_k,n}\binom{N}{M}}.$$

Also, we notice

$$\mathbb{P}\Big(i \sim j \text{ is deleted}\Big|i \sim j, deg(j) = 1, deg(i) = k\Big)$$
$$= \sum_{x=0}^{k} \mathbb{P}(X_k = x)\mathbb{P}\Big(i \sim j \text{ is deleted}\Big|X_k = x\Big) = \sum_{x=0}^{k}\binom{k}{x}q^k(1-q)^{k-x}\frac{x}{k} = q.$$

Therefore, substituting into Equation (A8), we obtain for $k \leq min(n-1, M)$

$$\mathbb{P}\Big(I'_j = 1\Big|deg(i) = k\Big) = \frac{\binom{N-k-(n-1)}{M-k}}{\binom{N}{M}} + \frac{\binom{n-2}{k-1}\binom{N-(2(n-2)+1)}{M-k}}{p_{deg_k,n}\binom{N}{M}}q.$$

Hence,

$$\lambda_k := \mathbb{E}[W'|i \text{ is the vertex picked for duplication}, deg(i) = k]$$
$$= q^k + (n-1)\left(\frac{\binom{N-k-(n-1)}{M-k}}{\binom{N}{M}} + \frac{\binom{n-2}{k-1}\binom{N-(2(n-2)+1)}{M-k}}{p_{deg_k,n}\binom{N}{M}}q\right)$$

After one distributed knockout attack on node i, let $X_k = x_k$. The graph is a realisation of the $G(n-1, M-k)$ model if node i becomes isolated, and it is a realisation of the model $G(n-1, M-x_k)$ combined with node i and its remaining edges if node i does not become isolated. Any additional isolated nodes can only appear in the $G(n-1, M-k)$ or $G(n-1, M-x_k)$ part of the model. With $p_x = \frac{M-x}{N'}$, a similar argument as for Equation (A10) gives as the upper bound for the total variation distance in a distributed knockout attack

$$\begin{aligned}
&d_{TV}(\mathcal{L}(W'); Po(\Lambda)) \\
&\leq \sum_{k=0}^{n-2}\sum_{x=0}^{k} p_{deg_k,n-1}\mathbb{P}(X_k = x)\min(1, \lambda_k^{-1})e^{-np_x\left(1+\frac{n-3}{N'+3-n}\right)} \\
&\quad e^{p_x\left(2+\frac{n-3}{N'+3-n}+\frac{6-5n+2n^2}{N'+3-n}\right)}\left(1+np_x\left(1+\frac{n+N'p_x-3}{N'-n-N'p_x+3}\right)\right) \qquad (A13) \\
&= \sum_{k=0}^{n-2}\sum_{x=0}^{k}\frac{\binom{n-2}{k}\binom{N'-(n-2)}{N'p_x-k}}{\binom{N'}{N'p_x}}\binom{k}{x}q^x(1-q)^{k-x}\min(1, \lambda_k^{-1})e^{-np_x\left(1+\frac{n-3}{N'+3-n}\right)} \\
&\quad e^{p_x\left(2+\frac{n-3}{N'+3-n}+\frac{6-5n+2n^2}{N'+3-n}\right)}\left(1+np_x\left(1+\frac{n+N'p_x-3}{N'-n-N'p_x+3}\right)\right).
\end{aligned}$$

□

Again, this bound can be bounded as

$$d_{TV}(\mathcal{L}(W'); Po(\Lambda)) \leq \sum_{x=0}^{n-1} e^{-np_x(1+\frac{n-3}{N'+3-n})+p_x(2+\frac{n-3}{N'+3-n})+\frac{6-5n+2n^2}{N'+3-n}} \left(1 + np_x\left(1 + \frac{n + N'p_x - 3}{N' - n - N'p_x + 3}\right)\right),$$

which tends to 0 as $p = \dfrac{M}{N} \to 1$.

Appendix B. Additional Figures

Below, we present additional figures for this paper.

Appendix B.1. Additional Results for Simulated Networks Starting with a Triangle

Figure A1 shows the effect of p on duplication–divergence models starting with a triangle against different attacks. As p increases, the network efficiency decreases faster; however, the relative behaviour between strong and weak attacks remains unchanged. We show in Figure A2 that for a new node loss model starting with a triangle, the network efficiency also follows the same trend.

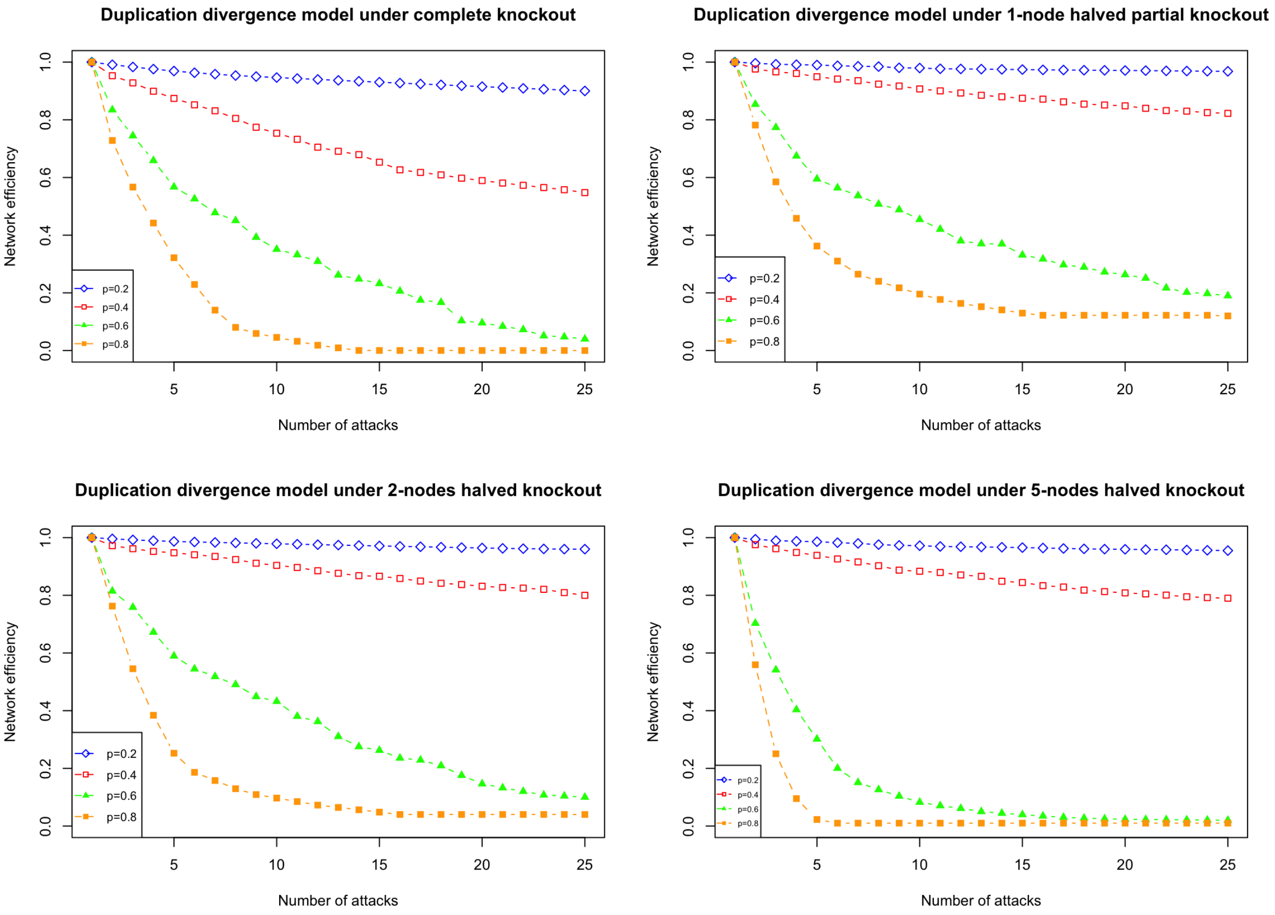


Figure A1. Effect of p when applying complete or weak knockout attacks on simulated networks from duplication–divergence models (i.e., $q = 0$) starting with a triangle.

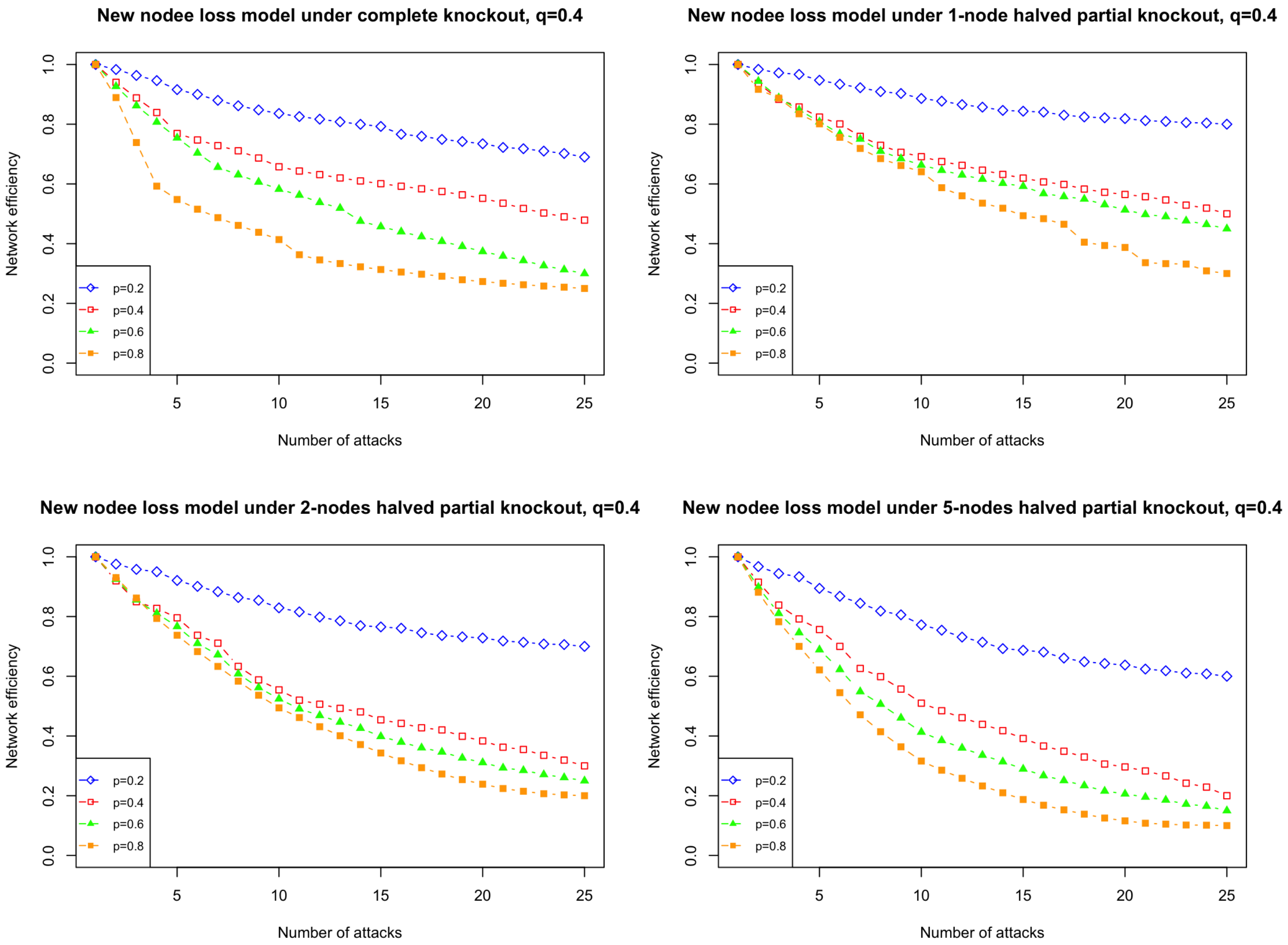
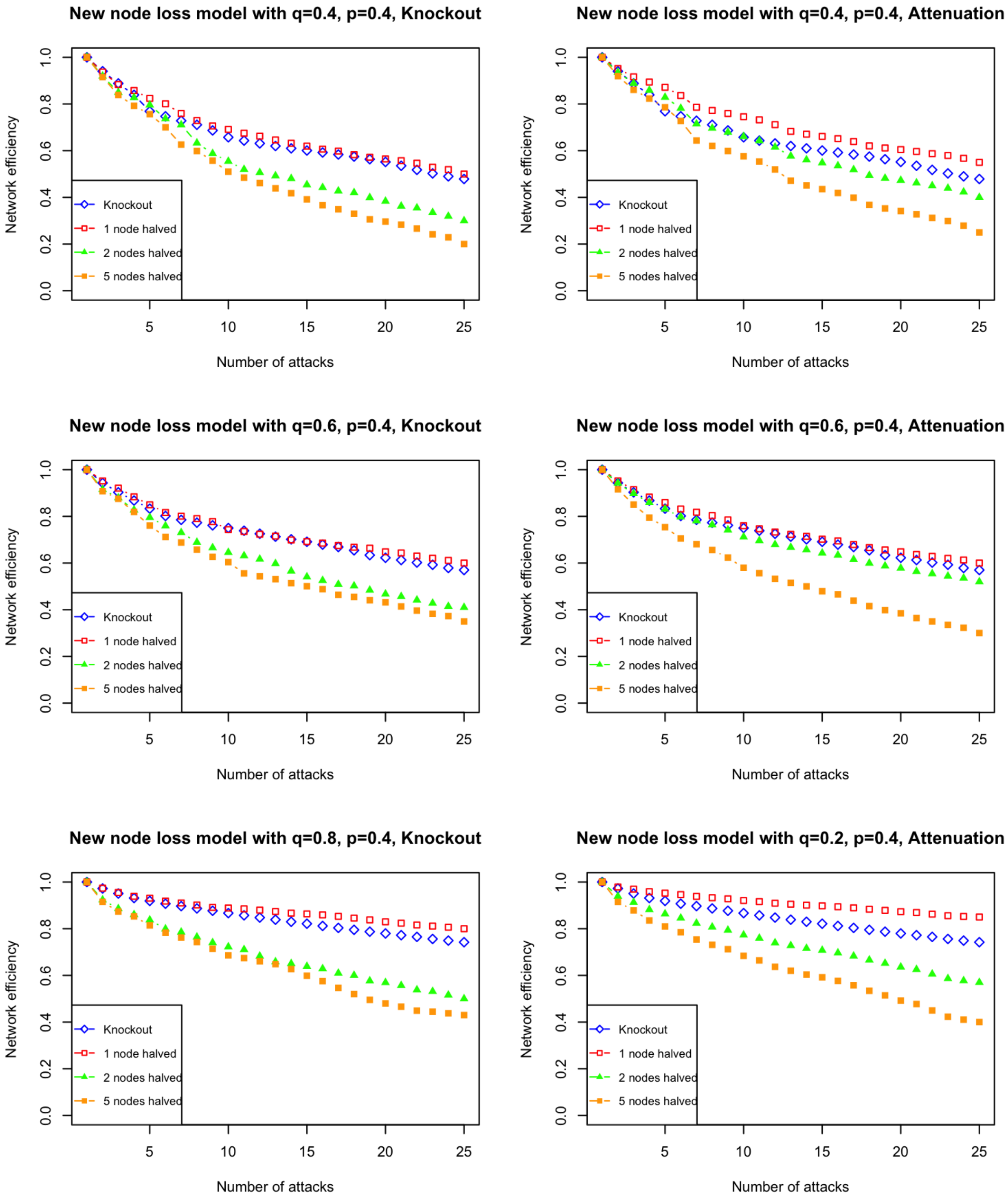


Figure A2. Effect of p when applying complete or weak knockout attacks on simulated networks from the new node loss model starting with a triangle, with $q = 0.4$.

Figure 7 in the main text shows the effect of different attacks on simulated networks from the new node loss model starting with a triangle, with $q = 0.2$, $p = 0.4$. In Figure A3, we present simulation results on the new node models with other sets of parameters; we find that these display qualitatively similar behaviour against attacks. Compared to the simulations of new node loss model that begin with a single edge, starting with a triangle provides a more realistic representation. This is not only because the triangle-based simulations have a non-zero local and global clustering coefficient unlike the edge-based simulations, but we also notice that, when $p = 0.4$, the network efficiency in the triangle-based simulations does not decline as rapidly as in Figure A7, more closely mirroring the behaviour of real PPI networks.



New node loss model with q=0.4, p=0.4, Knockout

New node loss model with q=0.4, p=0.4, Attenuation

New node loss model with q=0.6, p=0.4, Knockout

New node loss model with q=0.6, p=0.4, Attenuation

New node loss model with q=0.8, p=0.4, Knockout

New node loss model with q=0.2, p=0.4, Attenuation

Figure A3. Network efficiency after up to 25 weak attacks on simulated networks from the new node loss model starting with a triangle with a divergence rate $p = 0.4$, where a node can be lost with probability $q = 0.4$, 0.6, and 0.8. **Left plots:** knockout attacks. Blue line: complete knockout; red line: partial knockout with all the edges connected to one node being halved at each attack; green line: partial knockout with all the edges connected to two nodes being halved at each attack; orange line: partial knockout with all the edges connected to five nodes being halved at each attack. **Right plots:** attenuation attacks. Blue line: complete knockout; red line: partial attenuation with all the edges connected to one node being halved at each attack; green line: partial attenuation with all the edges connected to two nodes being halved at each attack; orange line: partial attenuation with all the edges connected to five nodes being halved at each attack.

Appendix B.2. Simulated Networks Starting with a Single Edge

In the main text, we present simulations of DD models and new node loss models initialised with a triangle. The figures below demonstrate that starting these models with an edge shows similar behaviour regarding the effect of attacks. However, when beginning with a single edge, no triangles are formed during the graph generation process, making the resulting networks less realistic for modelling PPI networks.

Figures A4 and A5 show the effect of p in simulations from the DD model without and with node loss, respectively, starting with a single edge, for different attacks. We notice that for both models, the relative behaviour between strong and weak attacks remains unchanged for different values of p and the starting configuration of the simulations.

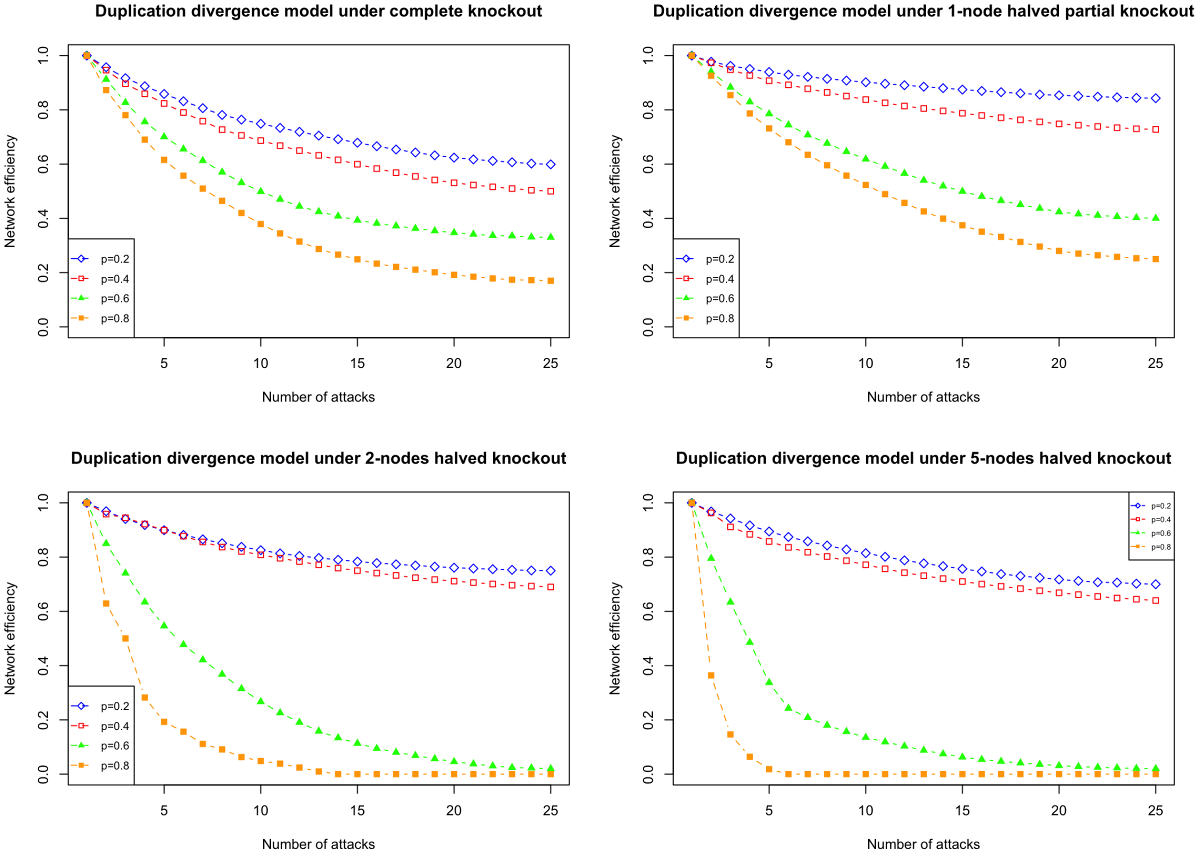


Figure A4. Effect of p when applying complete or weak knockout attacks on simulated networks from the duplication–divergence model starting with an edge.

Figure 7 in the main text gives the results for simulations from the new node loss model starting with a triangle, with $q = 0.2$ and $p = 0.4$. Figure A6 shows similar behaviour when the model starts with a single edge and weakly attacks a greater number of nodes, namely that partial attacks can generate greater damage to networks than complete knockout attacks as the number of targeted nodes increases, while distributed attacks are less efficient than complete and partial attacks. Moreover, weak attacks show the same qualitative behaviour as for the real PPI networks, see Figure 5.

A similar conclusion on the effect of weak attacks in the new node loss model starting with an edge is obtained when q, the probability of node loss, equals 0.4, 0.6, and 0.8, as shown in Figure A7. Figure A8 also shows that the qualitative behaviour is similar when $p = 0.2$.

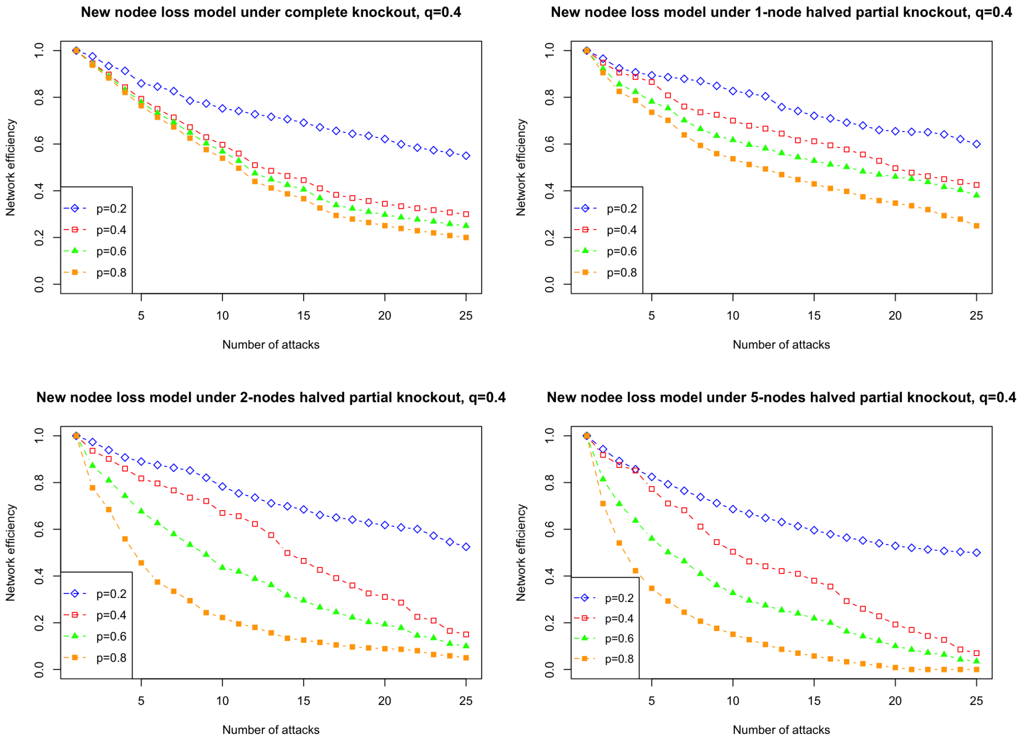


Figure A5. Effect of p when applying complete or weak knockout attacks on simulated networks from the node loss model starting with an edge, with $q = 0.4$.

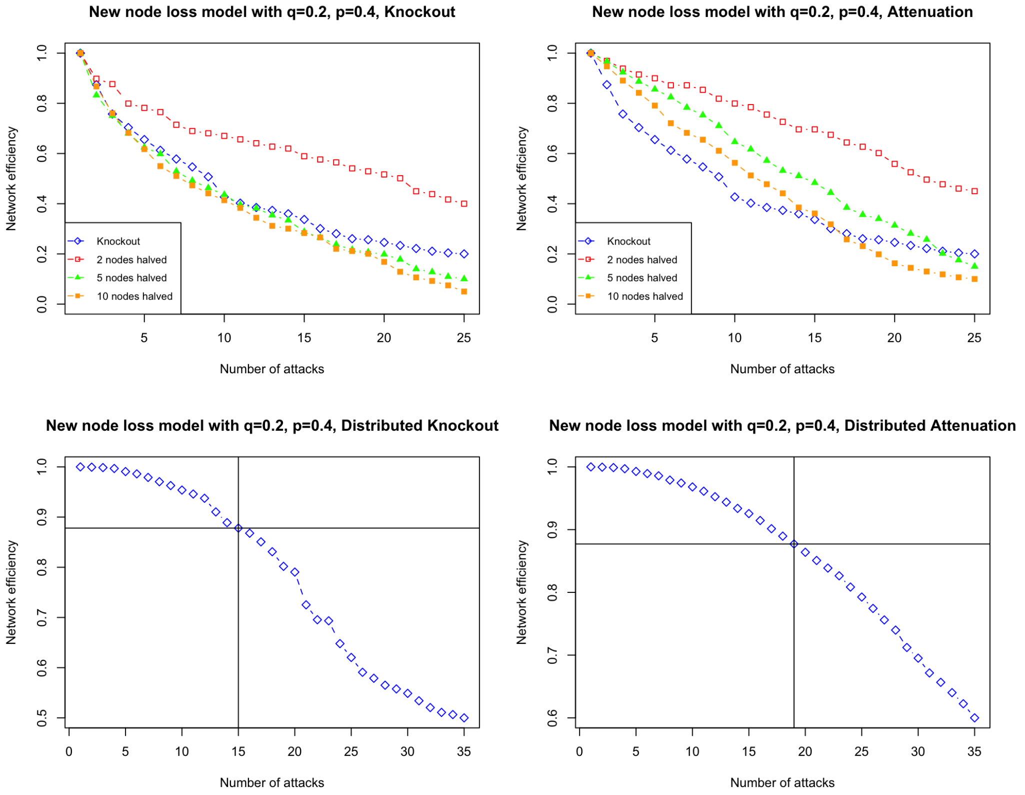


Figure A6. Weak attacks on simulated networks from the new node loss model starting with an edge where a node can be lost with probability $q = 0.2$, using a divergence rate $p = 0.4$. The graph is undirected

and has unit edge weight. Edges selected for distributed attacks are drawn from a random distribution. **Top left:** knockout attacks. Blue line: complete knockout; red line: partial knockout with all the edges connected to two nodes being halved at each attack; green line: partial knockout with all the edges connected to five nodes being halved at each attack; orange line: partial knockout with all the edges connected to ten nodes being halved at each attack. **Top right:** attenuation attacks. Blue line: complete knockout; red line: partial attenuation with all the edges connected to two nodes being halved at each attack; green line: partial attenuation with all the edges connected to five nodes being halved at each attack; orange line: partial attenuation with all the edges connected to ten nodes being halved at each attack. **Bottom left**: distributed attacks, with edges drawn from a random distribution; the horizontal line represents equivalent damage to the network achieved by one complete knockout. **Bottom right**: distributed attenuation attacks, with the weight of edges drawn from a random distribution to be halved; the horizontal line represents equivalent damage to the network achieved by one complete knockout.

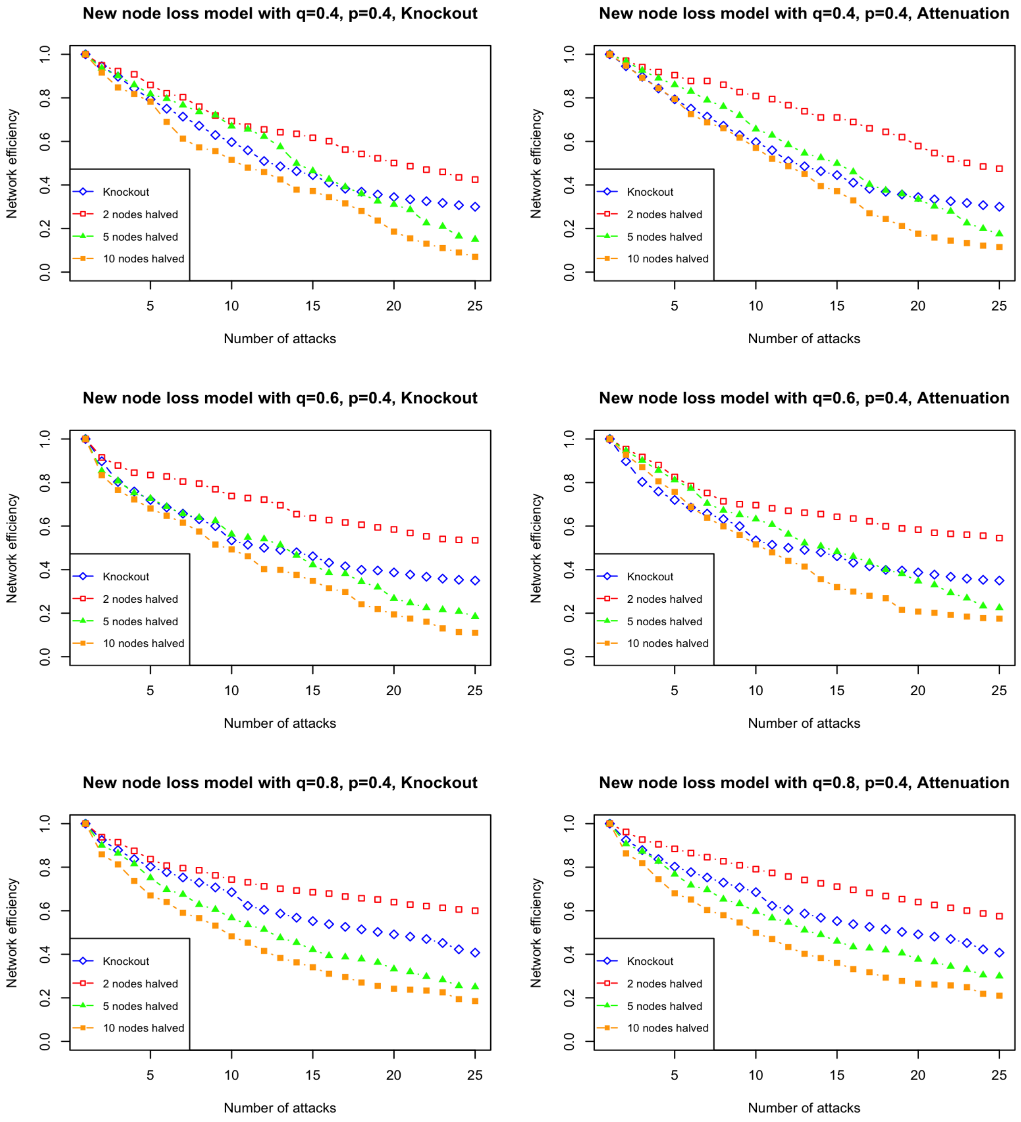


Figure A7. Weak attacks on simulated networks from the new node loss model starting with an edge; $p = 0.4$, $q = 0.6$, or $q = 0.8$, respectively. All the graphs are undirected with unit edge weight. Edges selected for distributed attacks are drawn from a random distribution.

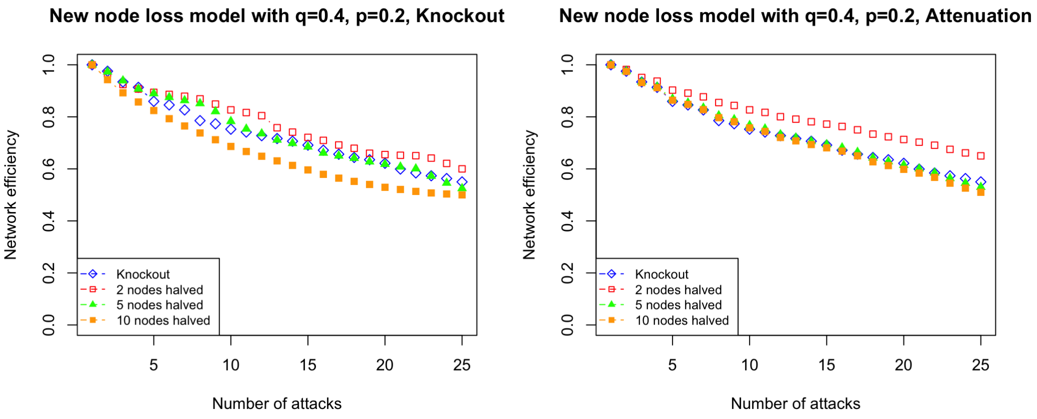


Figure A8. Weak attacks on simulated networks from the new node loss model starting with an edge; $p = 0.2$ and $q = 0.4$. All the graphs are undirected with unit edge weight. Edges selected for distributed attacks are drawn from a random distribution.

Appendix B.3. More Results for PPI Networks

Figure A9 shows that when the thresholds of STRING scores which are used to filter *E. coli* and *S. cerevisiae* PPI networks are changed from 0.400 to 0.200 or 0.600, the qualitative impact of complete and weak attacks on the datasets are the same.

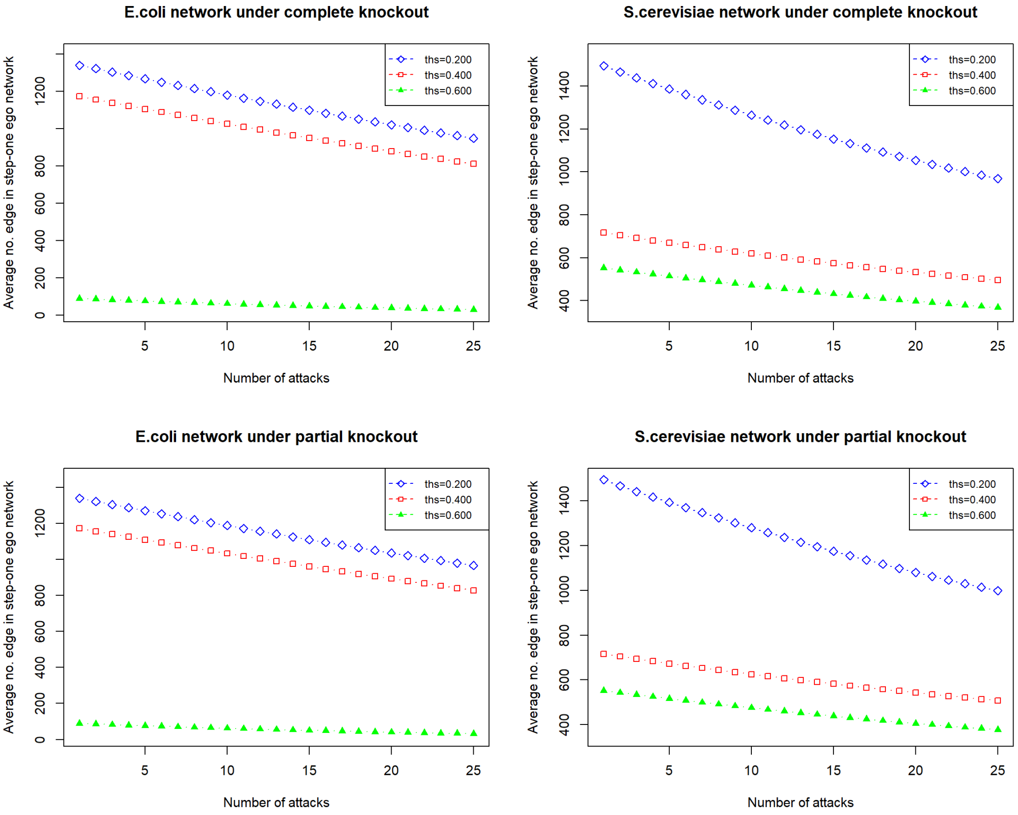


Figure A9. Effect of thresholds of STRING scores when applying complete or weak knockout attacks on real PPI networks, using the thresholds 0.200, 0.400, and 0.600.

References

1. Ágoston, V.; Csermely, P.; Sándor, P. Multiple weak hits confuse complex systems: A transcriptional regulatory network as an example. *Phys. Rev.* **2005**, *71*, 051909. [CrossRef] [PubMed]
2. Huang, S. Rational drug discovery: What can we learn from regulatory networks? *Drug Discov. Today* **2002**, *7*, s163–s169. [CrossRef] [PubMed]
3. Prato, S.D.; Volpe, L. Rosiglitazone plus metformin: Combination therapy for Type 2 diabetes. *Expert Opin. Pharmacother.* **2004**, *5*, 2051.
4. Kaelin, W.G. Gleevec: Prototype or Outlier? *Sci. STKE* **2004**, *2004*, pe12. [CrossRef] [PubMed]
5. Solé, R.; Pastor-Satorras, R.; Smith, E.; Kepler, T.B. A model of large-scale proteome evolution. *Adv. Complex Syst.* **2002**, *5*, 43–54. [CrossRef]
6. Chung, F.; Lu, L.; Dewey, G.; Galas, D. Duplication Models for Biological Networks. *J. Comput. Biol.* **2003**, *10*, 677–687. [CrossRef] [PubMed]
7. Gibson, T.A.; Goldberg, D.S. Improving evolutionary models of protein interaction networks. *Bioinformatics* **2011**, *27*, 376–382. [CrossRef] [PubMed]
8. Pastor-Satorras, R.; Smith, E.; Solé, R.V. Evolving protein interaction networks through gene duplication. *J. Theor. Biol.* **2003**, *222*, 199–210. [CrossRef] [PubMed]
9. Ospina-Forero, L.; Deane, C.; Reinert, G. Assessment of model fit via network comparison methods based on subgraph counts. *J. Complex Netw.* **2019**, *17*, 226–253. [CrossRef]
10. Hermann, F.; Pfaffelhuber, P. Large-scale behavior of the partial duplication random graph. *Lat. Am. J. Probab. Math. Stat.* **2016**, *1408*, 687–710. [CrossRef]
11. Albalat, R.; Cañestro, C. Evolution by gene loss. *Nat. Rev. Genet.* **2016**, *17*, 379–391. [CrossRef] [PubMed]
12. von Mering, C.; J Jensen, L.; Snel, B.; D Hooper, S.; Krupp, M.; Foglierini, M.; Jouffre, N.; Huynen, M.A.; Bork, P. STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucl. Acids Res.* **2005**, *33*, D433–D437. [CrossRef] [PubMed]
13. Bozhilova, L.V.; Whitmore, A.V.; Wray, J.; Reinert, G.; Deane, C.M. Measuring rank robustness in scored protein interaction networks. *BMC Bioinform.* **2019**, *20*, 446. [CrossRef] [PubMed]
14. Barbour, A.; Lo, T.Y. The expected degree distribution in transient duplication divergence models. *Lat. Am. J. Probab. Math. Stat.* **2021**, *19*, 69–107. [CrossRef]
15. Ispolatov, I.; Krapivsky, P.L.; Yuryev, A. Duplication-divergence model of protein interaction network. *Phys. Rev. E* **2005**, *71*, 061911. [CrossRef] [PubMed]
16. Stern, D. *The Chlamydomonas Sourcebook: Organellar and Metabolic Processes*; Academic Press: Cambridge, MA, USA, 2008.
17. Battiston, F.; Petri, G. *Higher-Order Systems*; Springer: Cham, Switzerland, 2022.
18. Goldstein, L.; Rinott, Y. Multivariate Normal Approximations by Stein's Method and Size Bias Couplings. *J. Appl. Probab.* **1996**, *33*, 1–17. [CrossRef]
19. Barbour, A.; Holst, L.; Janson, S. *Poisson Approximation*; Oxford University Press: Oxford, UK, 1992.
20. Barbour, A.; Reinert, G. Networks: Probability and Statistics. 2024, book manuscript in preparation.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.