An Evolutionary Model for the Emergence of Scale-Free Biological Networks

Amir Kadivar, Jan. 2017

Abstract

One of the universal properties of biological networks appears to be their scale-free connecetivity structure. Such a structure is identified by a power-law like distribution of degrees and can be found in gene regulatory, metabolic, and protein interaction networks. Recently, a computational model of optimization in biological networks has been proposed that draws analagies with the 0/1 knapsack problem. This report which is based on this computational model has two goals. First, we address some limitations of the model and propose modifications to resolve them. Second, we address the question of whether the power-law connectivity structure of biological networks, when considered as optimized subgraphs in a network of all possible genes/proteins/metabolites, can arise as a consequence of evolution under natural selection. To this end, we propose a simple evolutionary algorithm and qualitatively analyze the degree distribution of optimized networks that arise from a random (Erdős–Rényi) network. We will see that the evolutionary algorithm preserves the qualitative structure of degree distribution in the underlying network. Therefore, a scale-free connectivity structure in an optimal subgraph does not arise from evolution on random graphs and can presumably only arise from a network having the same scale-free property.

1 Background

Biological networks including gene regulatory networks, protein and metabolite interaction networks have been observed to share certain structural/topological features that appear universal [12, 10, 1, 5]. Being scale-free is one such property which is also found in a variety of other complex networks such as neural, ecological, and social networks. A scale-free network is one that follows a power law degree distribution, that is, the probability of a node having degree k varies in proportion to $k^{-\gamma}$ for some $\gamma > 0$. The universality of such structural features across such a large range, from molecules to ecosystems to man-made networks, demands a universal explanation based on generic principles. One successful attempt is the growing network model of Barabási and Albert [3, 1, 2] which explains the scale-free structure in terms of a *preferential attacahment* condition on the growing network. According to this condition, more highly connected nodes in a network are more likely to acquire new links. This condition is justifiable, and intuitively obvious, for artificial networks like the Internet backbone, friendship and citation networks, and linguistsic networks. For biological networks, however, the mechanistic explanation for such preferential attachment, although supported by empirical evidence [4], remains unclear.

A novel line of thought regarding the structural evolution of biological networks has been recently proposed in which the fitness landscape is modeled as a two dimensional cost/benefit optimization problem analagous to the 0/1 knapsack: the process of evolution by natural selection is modeled as an optimization process that seeks to maximize the total benefit obtained by advantageous interactions in the network while maintaining an upper bound on the total damage incurred by disadvantageous interactions [9].

In this report, we first review and propose modifications to the model of [9]. Then, we use an evolutionary algorithm to examine how the degree distribution of growing networks, when considered as subgraphs in a "universal network", evolve over time and how these distributions relate to the distribution of the underlying universal network. The simulation results show that the evolutionary algorithm preserves the qualitative structure of degree distribution in the universal network such that evolved subgraphs over an Erdős-Rényi universal network also have binomial-like degree distributions. However, power-law like degree distributions can arise under certain parameter regimes and at certain evolutionary times when the binomial mean approaches 1 which is explained by the well-understood association between binomial and Poisson distributions for small means. The results imply that preferential attachment does not arise statistically at the single-gene level and presumably occurs at the level of interacting protein domains [11, 7] or multiple-gene motifs [6].

2 Network Evolution Formulation

We first describe the model in [9]: a network $G = (V, E)$ consists of weighted nodes $V = \{v_i\}$ with weights $\{s_i^{(v)}\}$ and directed edges $E = \{e_i\}$ with weights $\{s_i^{(e)}\}$. Node weights correspond to whether a gene (or protein) is beneficial, neutral, or disadvantageous and are drawn randomly from $\{-1, 0, 1\}$ in such a way that

$$\Pr\{s_i^{(v)} \neq 0\} = p$$

where p represents the selection pressure. Edge weights correspond to whether gene (or protein) interactions lead to inhibition or promotion of the target and are randomly drawn from $\{-1, 1\}$. Each interaction $e_k = v_i \to v_j$ is considered beneficial if $s_k^{(e)} s_j^{(v)} > 0$ and damaging otherwise. Accordingly, e_k either contributes to the benefit b_i or damage d_i of both its incident nodes depending on the sign of $s_k^{(e)} s_j^{(v)}$. The network evolution problem is then stated as follows:

$$\max_{V^* \subset V} \sum_{v_i \in V^*} b_i$$

$$\text{subject to } \sum_{v_i \in V^*} d_i < D := t \left(\sum_{i \in G} d_i \right)$$

where D is the maximum allowable damage determined by the tolerance parameter $t \leq 1$ as a proportion of total damages in the network.

Once benefits and damages of all nodes in the graph are determined, the problem is shown to be NP-hard by reduction of the 0/1 knapsack problem. This correspondance with the knapsack problem inspires the use of what is known of the tractability of knapsack problems [8] to analyze the effective complexity of the network evolution problem. For instance, we know that highly correlated value/weight ratios make a knapsack problem more difficult to solve which may explain certain features of evolved biological networks.

In this section, however, we argue that the above formulation and the knapsack reduction of [9] have limited applicability to biological networks. First, the above formulation assumes that the total benefit of a node for a subgraph identified by $V^* \subset V$ is independent of the presence of other nodes in V. This is required for nodes to have fixed benefit and damages and thus the problem to be relatable to knapsack. However, if the model is to be taken mechanistically, the benefit and damage contributions of edges that do not connect nodes in V^* should be discounted when evaluating the fitness of the subgraph. Thus, if we incorporate the fact that knapsack weights and values of genes are not independent of the choice of items, the knapsack correspondance breaks down.

Second, NP-hardness of the network evolution problem is proved by reducing an arbitrary $0/1$ knapsack problem to a network evolution problem. However, the constructed network evolution problem contains edges with weights outside of $\{-1, 0, +1\}$. Therefore, it is possible that a $\{-1, 0, +1\}$ network evolution problem is in fact easier than the $0/1$ knapsack problem and thus not necessarily NP-hard.

Third, the calculation of benefit and damage contributions of edges can be improved to match what is biologically known from gene and protein interactions. When a beneficial gene is suppressed by another gene, it is biologically appropriate to force the benefit and damage contributions of the interaction to vanish. Instead, in the above formulation, if flipping the sign of an edge with a beneficial target not only removes the benefit contribution but it switches to a contribution to the damage of the target gene (which itself remains beneficial). This can be easily fixed by picking edge weights $s_i^{(e)}$ from $\{0, 1\}$ instead of $\{-1, +1\}$ in which case all other calculations remain valid.

Finally, we note that this formulation ignores the spatial and temporal complexity of biological networks. Although this simplification is justified for model tractability, the equilibrium assumption limits the applicability of the model to real biological networks. Essentially, by fixing p, t and the beneficial/disadvantageous assignments $s_i^{(v)}$ to nodes, we are assuming a fixed (or steady-state) environmental landscape and development state.

3 Evolutionary Algorithm

In this section we summarize the model that is used to simulate evolution by natural selection on biological networks. The *universal network* $G = (V, E)$ is constructed as in [9] but is interpreted differently: V contains all genes (or proteins) that are physically possible, i.e not limited to what is found in living organisms. The point here is to allow evolution to carve an optimal subgraph out of the universal network. The set of edges E reflect the natural interaction between genes (or proteins) and is considered to be determined by physical law and independent of evolution.

Individuals I_k are identified as subgraphs of G with nodes $V_k \subset V$ and edges

$$E_k = \{e = (v_i, v_j); \{v_i, v_j\} \subset V_k\}$$

We apply two modifications discussed in the previous section: edge weights $s_i^{(e)}$ belong to $\{0, 1\}$, and benefits $b_i^{V_k}$ and damages $d_i^{V_k}$ depend on the choice of subgraph $V_k \subset V$ and are re-calculated for every individual. Each generation is a set of individuals $\{I_1, \ldots, I_C\}$ where $C \in \mathbb{N}$ is a constant representing capacity. The fitness of an individual $f(I_k)$ is

given by:

$$f(I_k) = \sum_{v_i \in V_k} b_i^{V_k}$$

In each iteration of the algorithm the population of C individuals is replaced as follows:

1. Individuals are sorted by fitness such that $f(I_1) \leq \ldots \leq f(I_C)$.
2. The least fit $k \leq (1-s)C$ individuals are removed from the population where $s \leq 1$ represents survivorship.
3. The surviving sC individuals are copied as-is to the next generation.
4. The remaining $(1-s)C$ spots of the next generation are populated by mutated descendents of the fittest sC individuals I_k for $k > (1-s)C$ with fecundity of I_k proportional to $f(I_k)$.

The process of gene mutation contains a deletion operation and an insertion operation which wraps a duplication-and-diversification event [5, 11]:

1. Each gene in V_k is inherited by proginy unless it is deleted with probability $1 - p_d$.
2. Each inherited gene v is duplicated-and-diversified with probability p_i in which case is new gene v' is added to the proginy. The probability distribution $\Pr\{v'|v\}$ is subject to one of the experiments.

The above description can be summarized using the following pseudocode:

```
def next_generation({I_k}, C, s, D, p_d, p_i):
    survivors = []
    for I_k in {I_k}:
        benefit, damage = ... # benefits and damages keyed by node
        if sum(damage) > D:
            continue
        survivors += [I_k]
    survivors = sorted(I_k) # in increasing order of total benefit
    survivors = survivors[sC:]
    next_gen = []
    next_gen += survivors # keep the fittest as-is in next generation
    for I_k in survivors:
        fecundity = ... # proportional to fitness of I_k s.t. sum over all k = (1-s)C
        for i in range(fecundity):
            offspring_genes = []
            for v in V_k:
                if bernoulli(p_d):
                    continue
                offspring_genes += [v]
                if bernoulli(p_i):
                    new_gene = ... # drawn according to a specified distribution
                    offspring_genes += [new_gene]
            next_gen += [subgraph(offspring_genes)]
    return next_gen
```

4 Experiments and Results

The main question we ask in the following experiments is this: *"Can a scale-free optimal graph emerge, under appropriate parameter regimes, from an Erdős-Rényi universal network?"* Due to the multitude of parameters involved, we fix a following paramter regime as the "baseline" and evaluate the effect of changes in paramter values by comparing results to the baseline configuration.

Parameter	Symbol	Baseline value
Network size	$\|V\|$	1000
Network edge probability	p_e	0.01
Pressure on node weights $s_i^{(v)}$	p	0.9
Edge positive probability	p_+	0.7
Tolerance ratio $D/\sum d_i$	t	0.1
Node deletion probability	p_d	0.1
Node insertion probability	p_i	0.1
Population capacity	C	200
Node weight distribution for $s_i^{(v)} \neq 0$	–	uniform
Edge sign $s_i^{(e)}$ range	–	$\{0, 1\}$
Node insertion distribution	–	similarity

In addition to what was discussed in the previous section, the last three parameters capture additional variations on the evolutionary algorithm:

1. Node weights $\{s_i^{(v)}\}$ are drawn from $\{-1, 0, 1\}$ such that for nonzero weights, the proportion of which is dictated by p, the choice is uniform over $\{-1, +1\}$. Alternatively, we wish to investigate how a power law distribution of weights, i.e $|s_i^{(v)}|$ allowed to exceed 1, affects the topology of optimized subgraphs.
2. We repeat the experiment for edge signs $s_i^{(e)}$ as in the original model with values from $\{-1, +1\}$.
3. When a node v is duplicated-and-diversified with insertion probability p_i we need to randomly pick a new node v' from $V \setminus \{v\}$ to represented a mutated version of v. In the simplest case, v' is drawn uniformly. Alternatively, in order to capture the fact that the mutant node v' is probably "similar" to v in some aspects, we allow the probability of a specific choice of mutatnt v' to be proportional to its similarity to v as measured by the number of their shared neighbors.

In each experiment, C individuals are created each containing only one gene and 400 iter-

ations of the evolutionary algorithm are performed. The evolution of degree distributions, total benefits and damages and gene counts are plotted for each experiment. To visually check power-law like patterns, all distributions are plotted in log-log coordinates in which a power-law distribution appears as a straight decreasing line [1] (Fig. 1). The general outcome of all performed experiments (Fig. 3, 4 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17) can be summarized as follows:

1. Early in the evolutionary process all populations appear power-law like due to the fact that all genes have small degrees in the subgraph. Over the course of optimization the degree distributions, as well as total benefits, damages and gene counts, can be qualitatively seen to stabilize.
2. The knapsack solver based on the original formulation of [9] produces degree distributions that are consistent with the outcome of the evolutionary algorithm. This is surprising since the knapsack solver uses subgraph-independent benefit and damage assignments for nodes to reverse-reduce the problem to 0/1 knapsack.
3. The evolutionary algorithm seems to preserve the degree distribution the underlying universal network. In other words, optimized subgraphs have binomial-like distributions just like the underlying universal network. However, since the binomial distribution is not scale free, degree distributions have different shapes for small binomial means which may appear power-law like (Fig. 2). This can be verified in every case by calculating the expected binomial mean from the product of subgraph gene count and the ER edge probability. When the same algorithm is applied to a biological power-law distributed network, the optimized subgraphs also show power-law degree distributions.
4. The main effect of changing the paramters listed above is to speed up or slow down the optimization process thus snapshots of the population at fixed generation counts for different parameter regimes merely show the same degree distribution families (e.g. binomial for ER universal network).

[1] All code is available at https://gist.github.com/amirkdv/234d776abf55b1303bc03212512a5ba9.

5 Figures

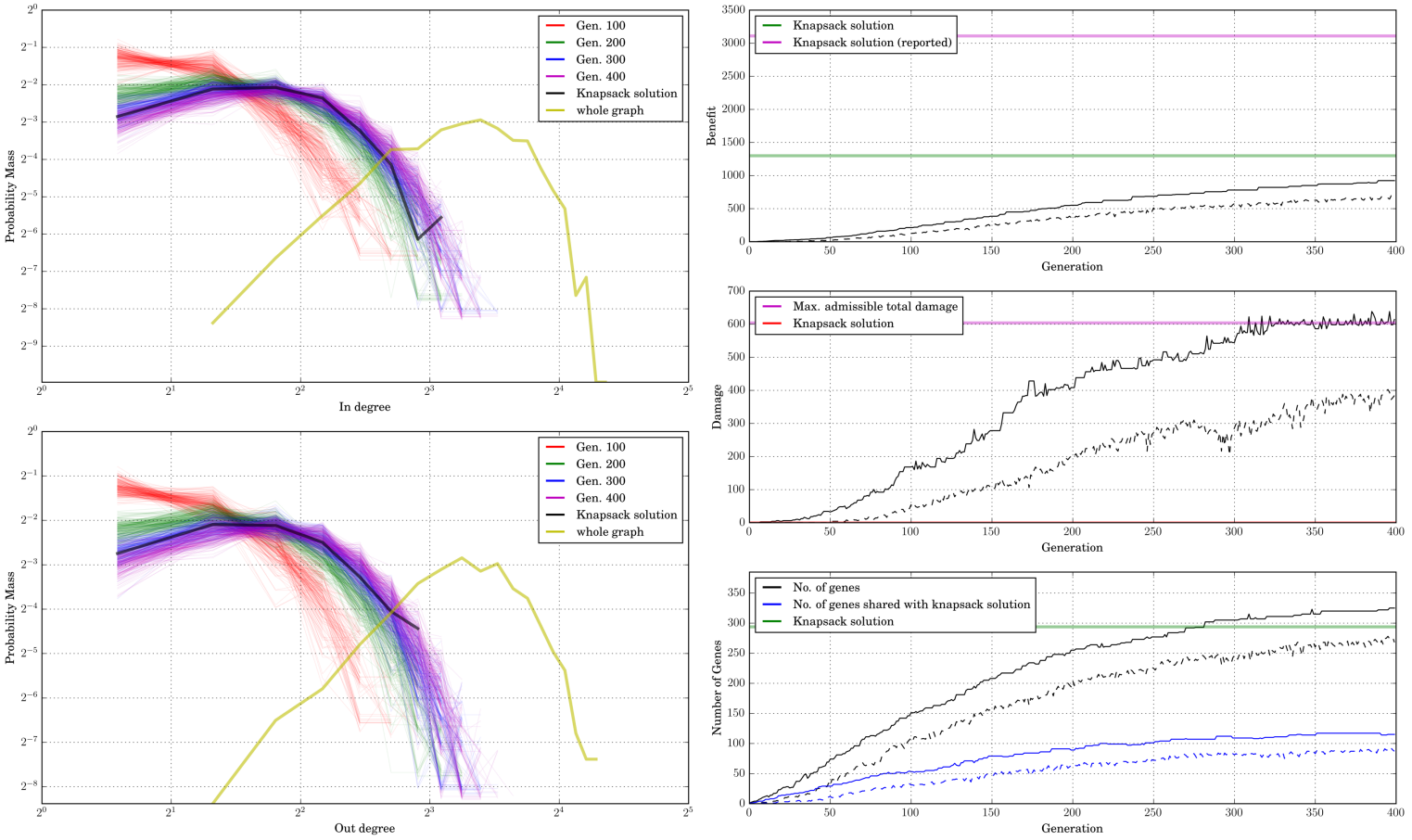

Figure 1: In/out degree distributions of all individuals at generations 100, 200, 300, and 400 (*left*) and time series of maximum (*solid*) and minimum (*dashed*) values of total benefits (*top*), total damages (*middle*), and gene count (*bottom*) over the course of 400 generations (*right*) for baseline parameters. Note the consistency of degree distributions with those obtained from the knapsack solver.

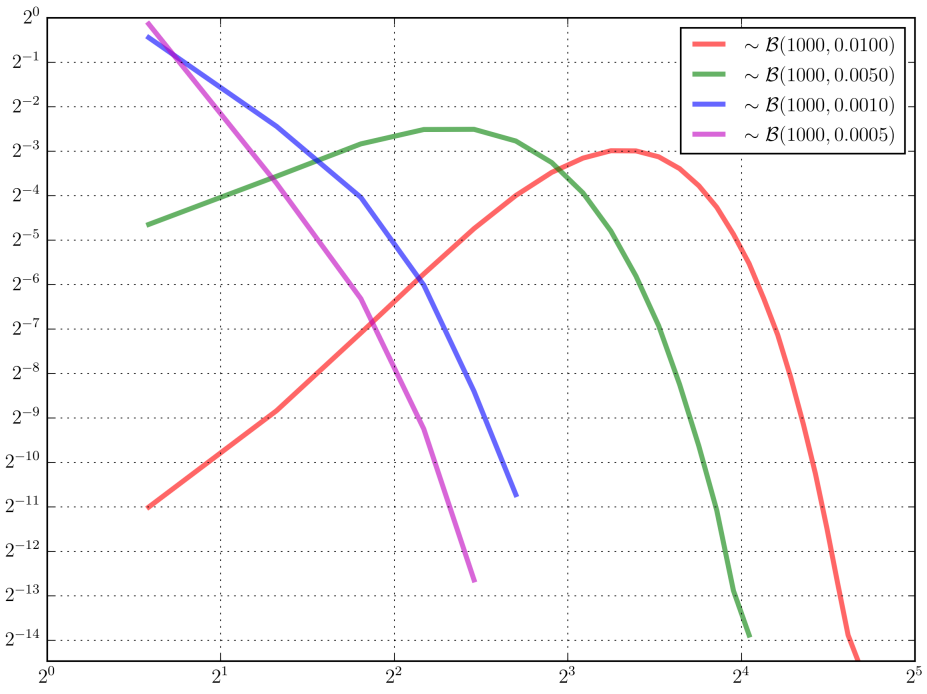

Figure 2: Binomial distributions with different means. Note that for small mean values the distribution can appear similar to a power-law distribution.

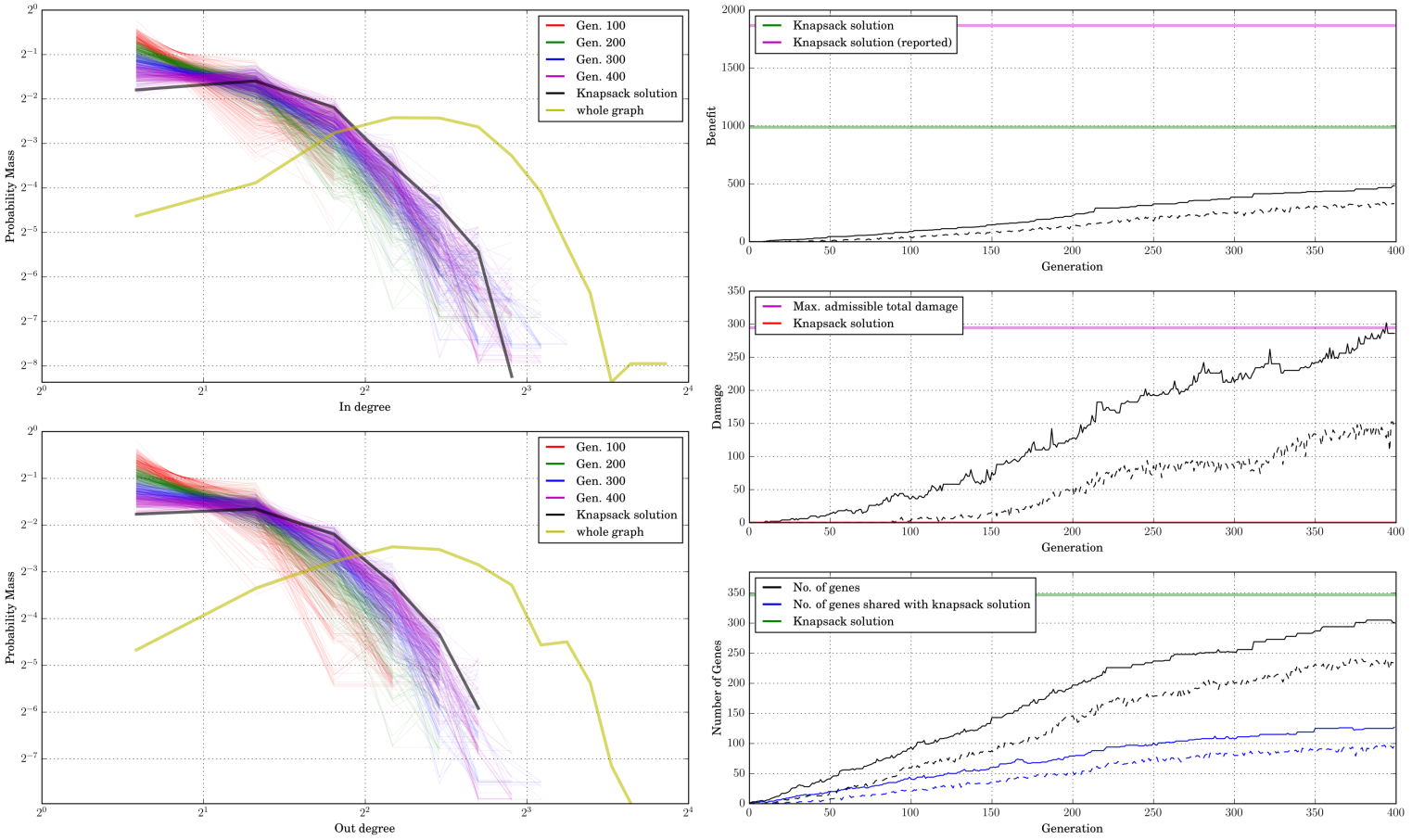

Figure 3: Similar experiment as in Fig. 1 but with sparser connectivity $p_e = 0.005$. Note the appearance of power-law like distributions, however, see Fig. 4.

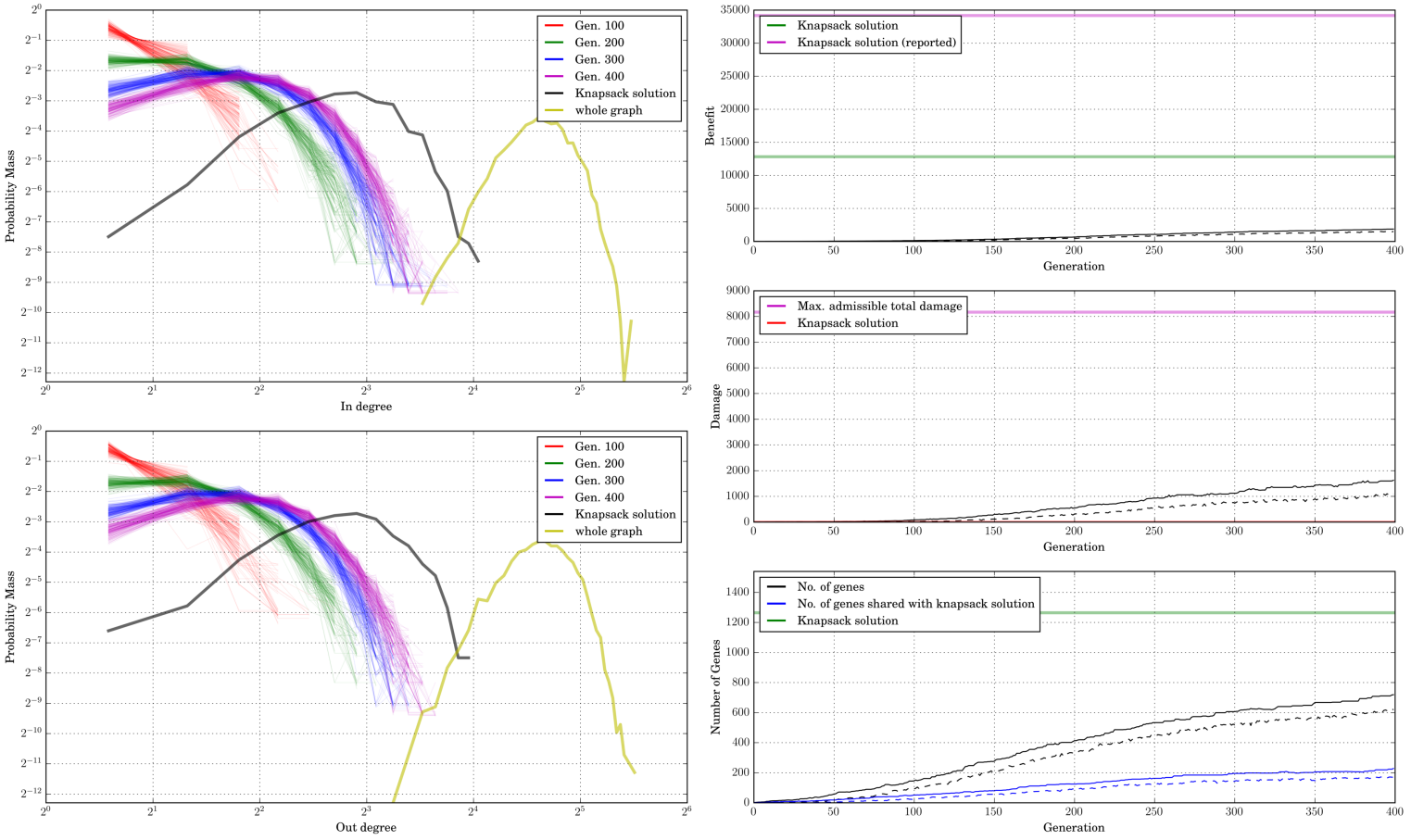

Figure 4: Similar experiment as in Fig. 3 but with larger universal network size $|V| = 5000$. This experiment shows that the power-law like distributions seen in 3 are merely due to small binomial means and not a consequence of universal network sparsity.

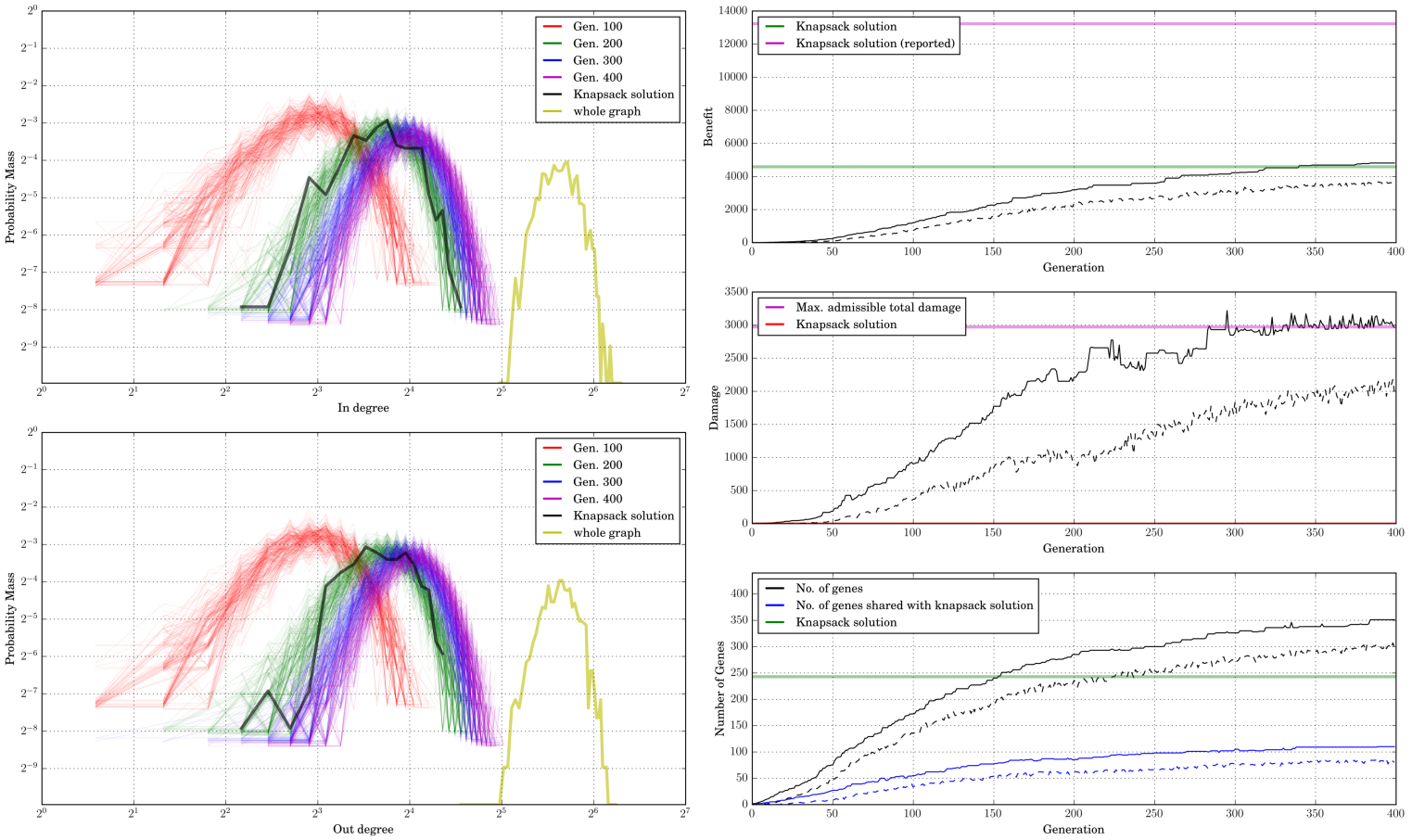

Figure 5: Similar experiment as in Fig. 1 but with higher connectivity $p_e = 0.05$. The preservation of underlying network degree distribution is most clear in this example. Note the shifting binomial mean as the gene count (*right bottom*) increases.

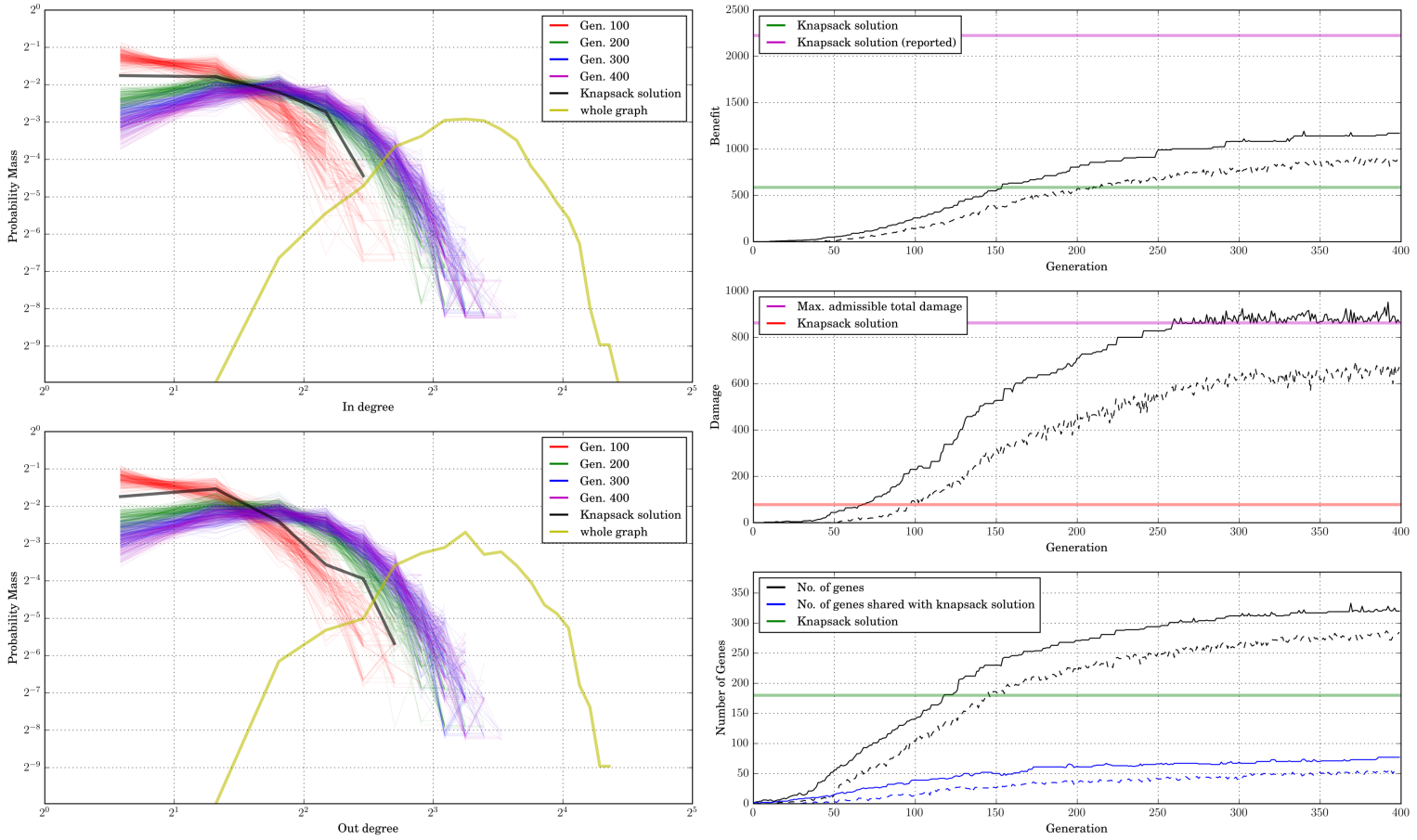

Figure 6: Similar experiment as in Fig. 1 but with edge signs $s_i^{(e)}$ taken from $\{-1, +1\}$, as in [9], instead of $\{0, 1\}$. This means that suppressing an advantageous gene contributes to the total damages.

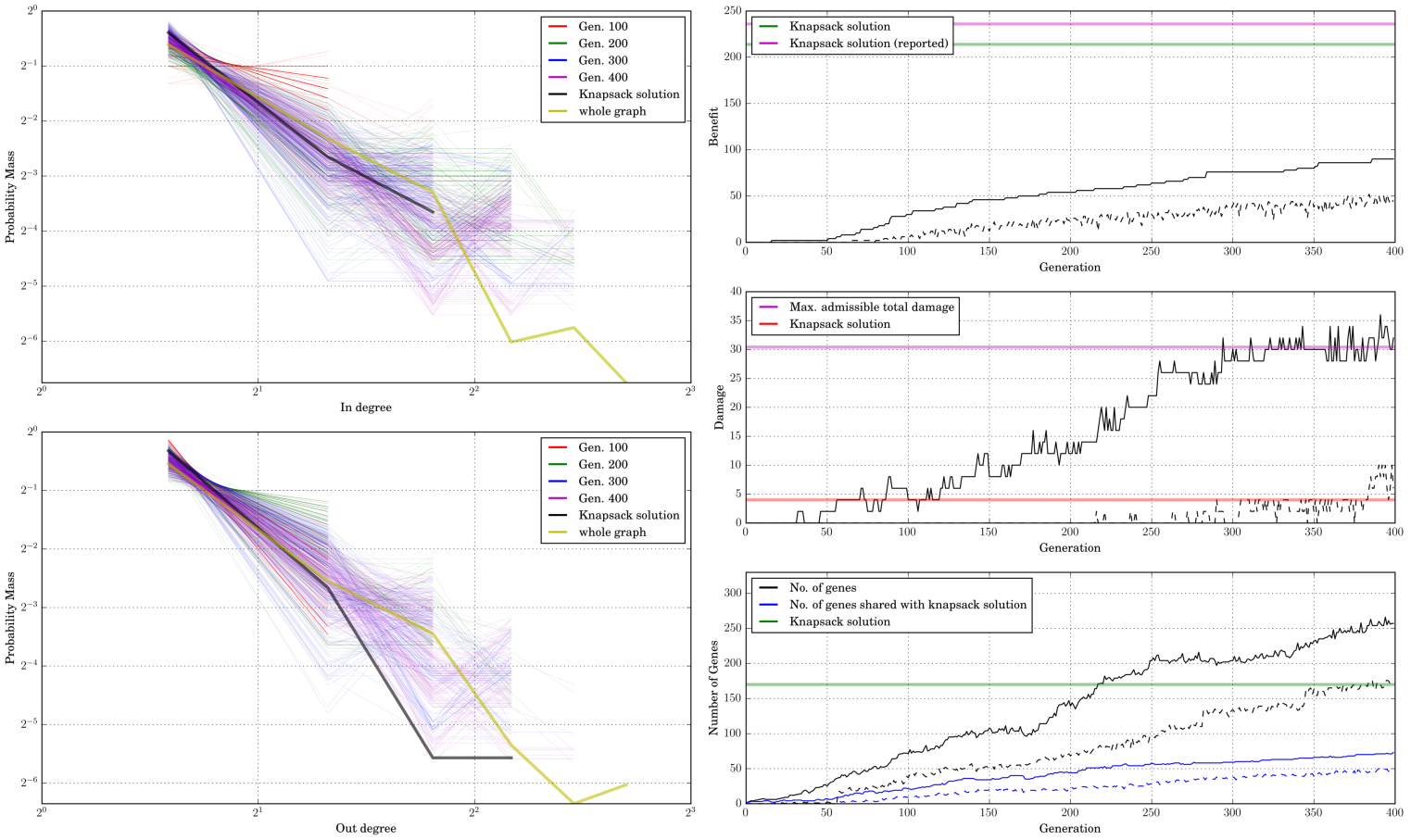

Figure 7: Similar experiment as in Fig. 1 but with a biological scale-free universal network instead of the simulated ER network and with uniform choice of inserted nodes (neighbor similarity ignored). Note that the emerging degree distributions are all power-law supporting the conclusion that the evolutionary algorithm preserves the degree distribution of the underlying network.

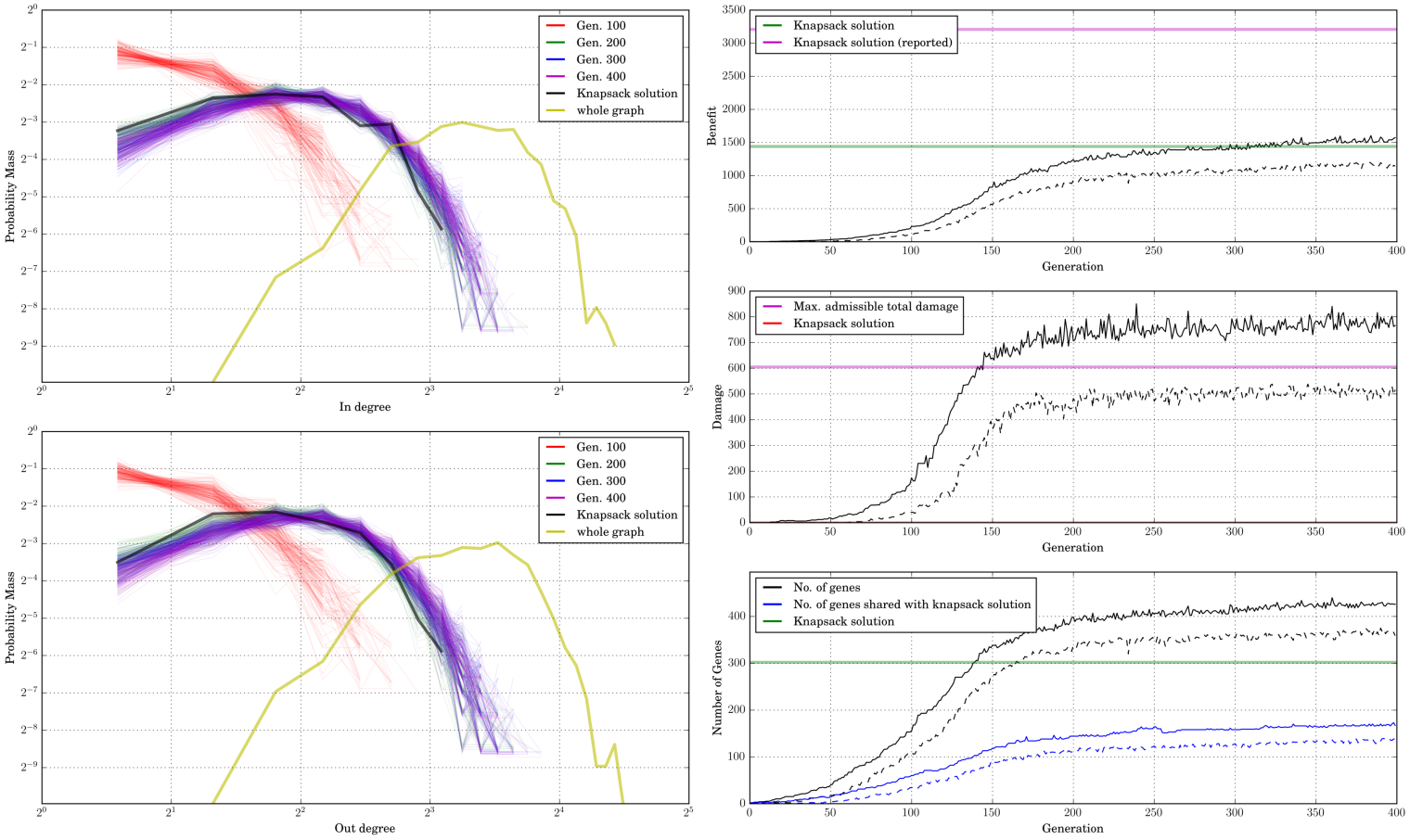

Figure 8: Similar experiment as in Fig. 1 but with uniform choice of inserted nodes instead of using neighbor-based similarity scores.

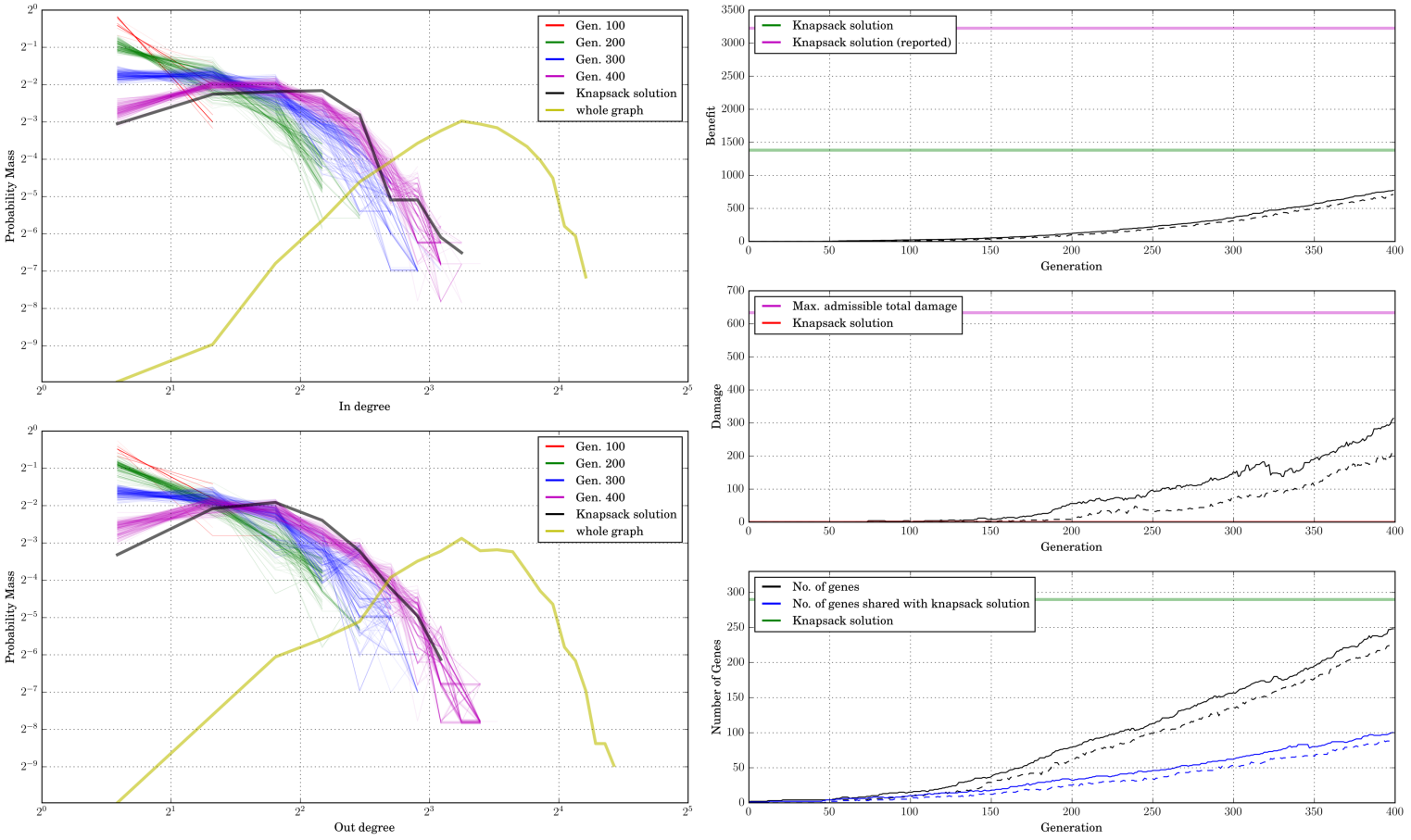

Figure 9: Similar experiment as in Fig. 1 but with decreased mutation rates $p_i = p_d = 0.01$. Note that the only effect of this change appears to be a reduction in convergance rate. This slowing down leads to small binomial means and the appearance of power-law degree distributions, however, cf. Fig. 10.

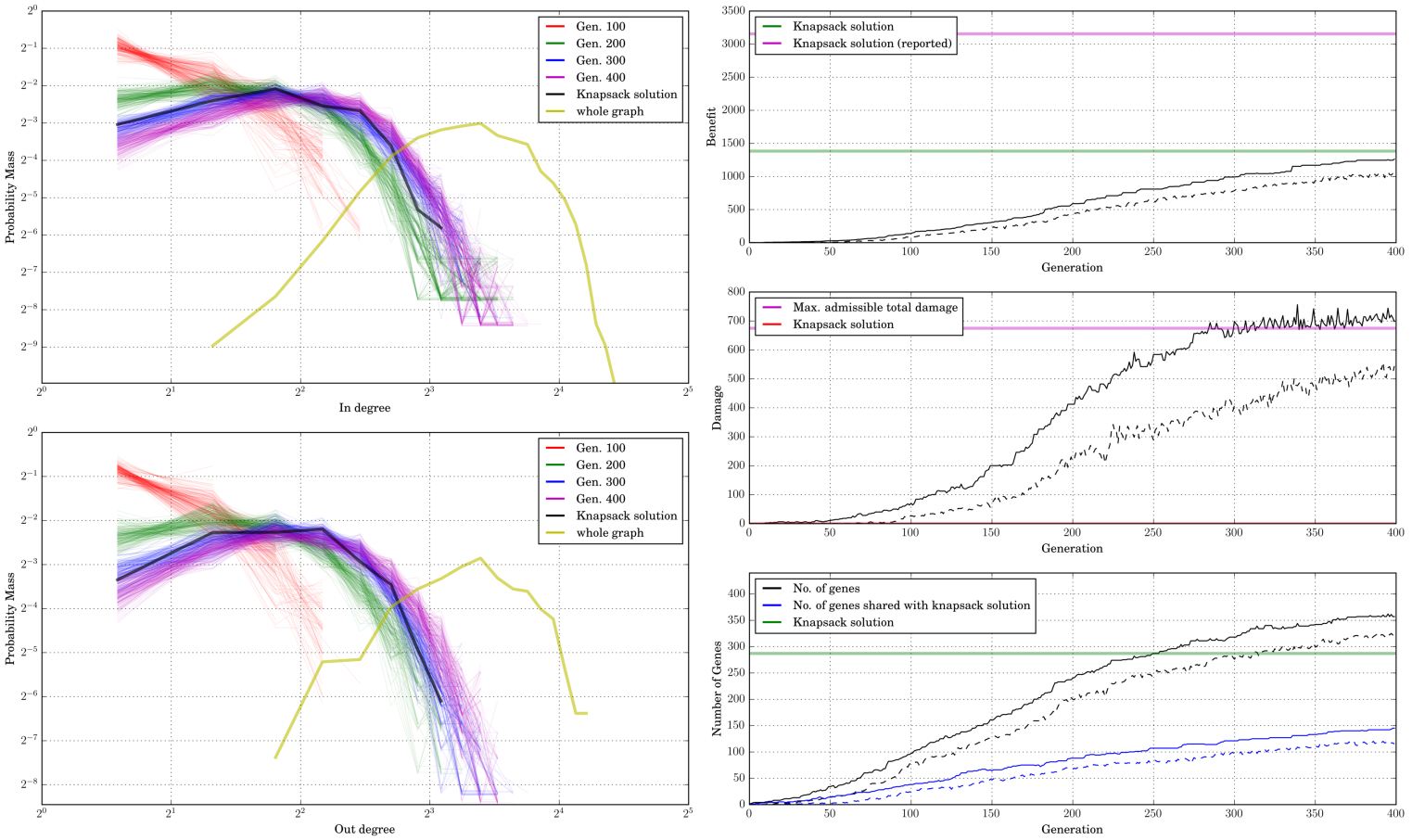

Figure 10: Similar experiment as in Fig. 1 but with decreased mutation rates $p_i = p_d = 0.05$. This experiments confirms that the power-law like distributions seen in Fig. 9 are merely due to small binomial means.

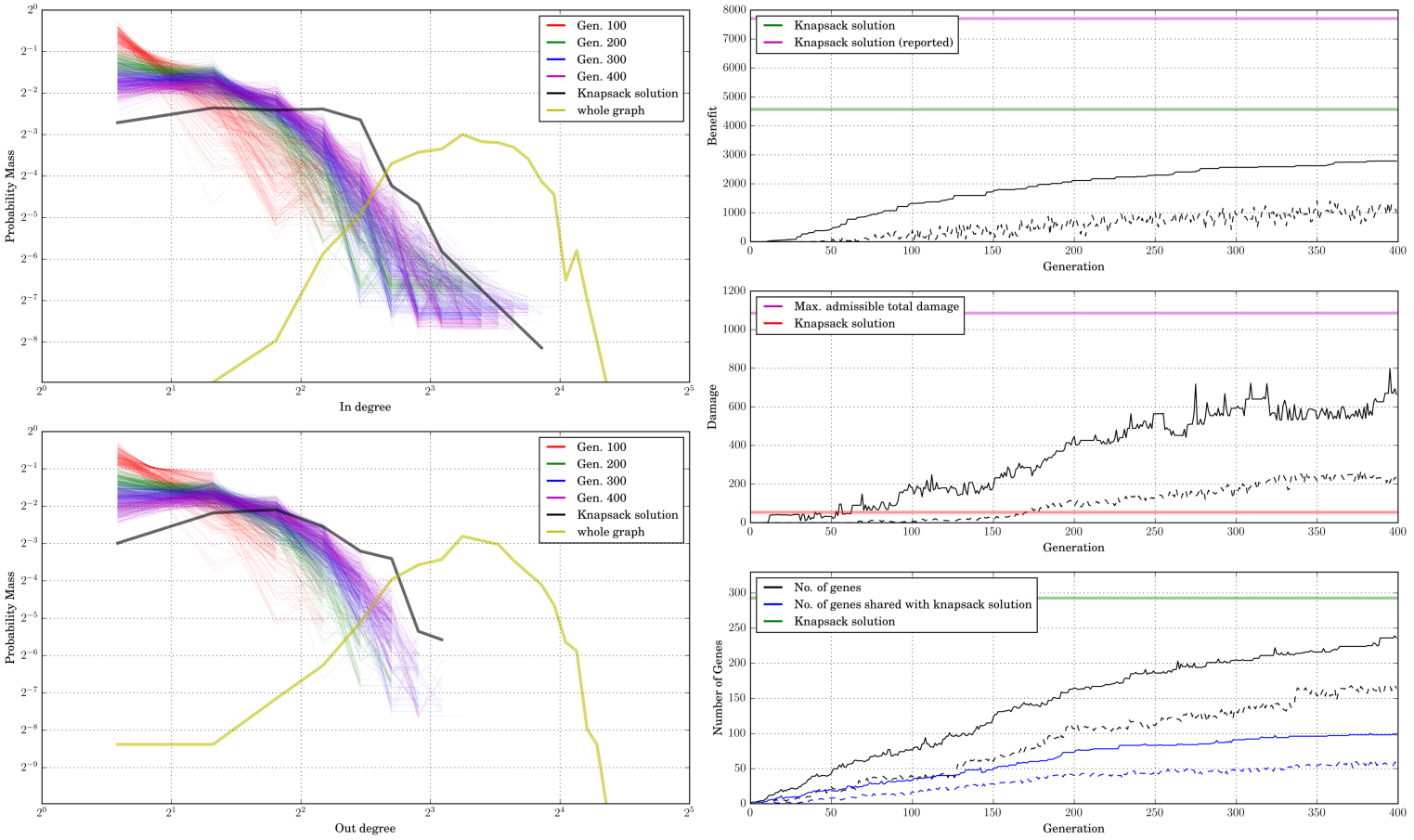

Figure 11: Similar experiment as in Fig. 1 but with node weights $s_i^{(v)}$ drawn from a power law distribution with exponent $\gamma = 1$. Note that this does not change the qualitative structure of the optimized subgraph degree distributions.

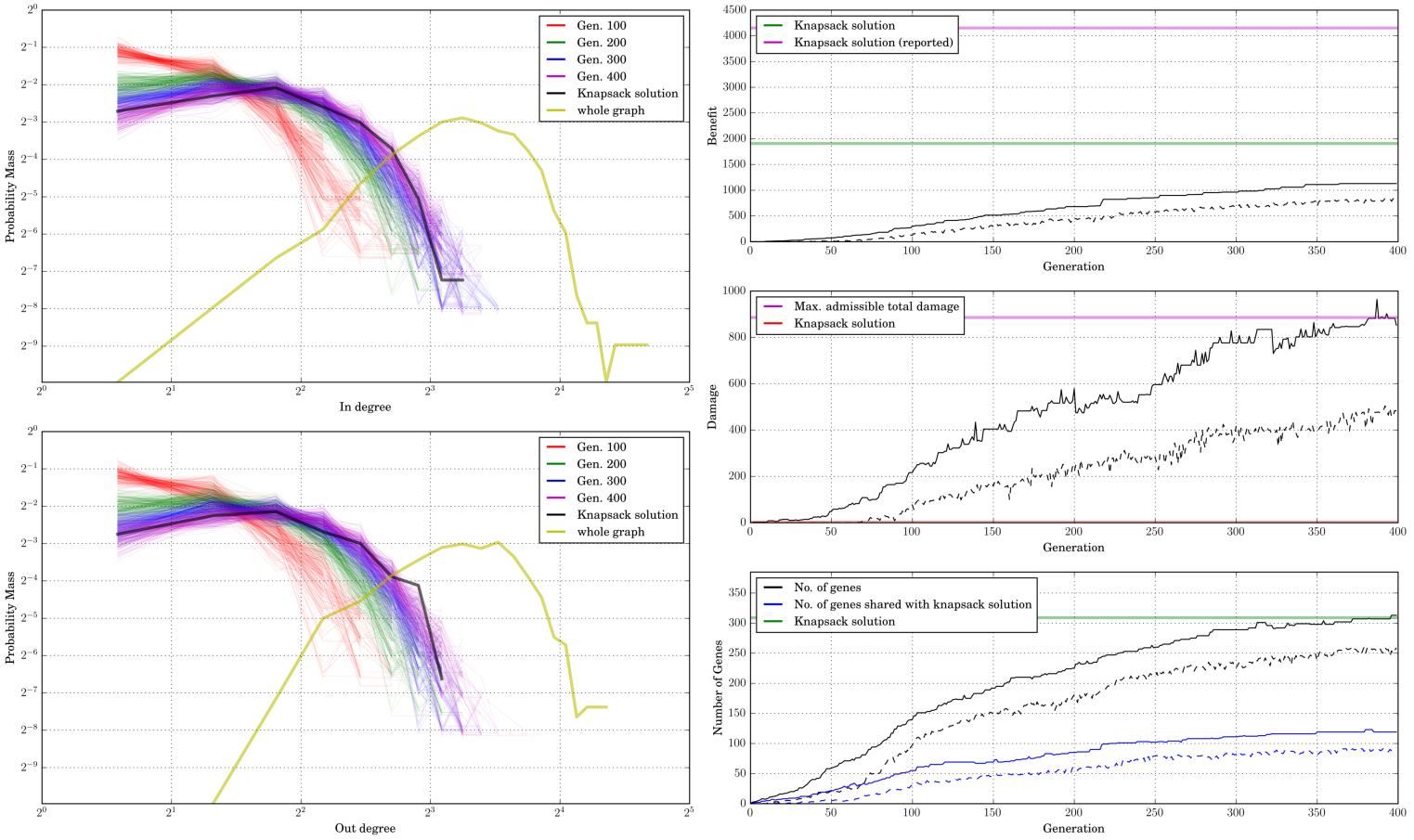

Figure 12: Similar experiment as in Fig. 1 but with node weights $s_i^{(v)}$ drawn from a power law distribution with exponent $\gamma = 2$. Note that this does not change the qualitative structure of the optimized subgraph degree distributions.

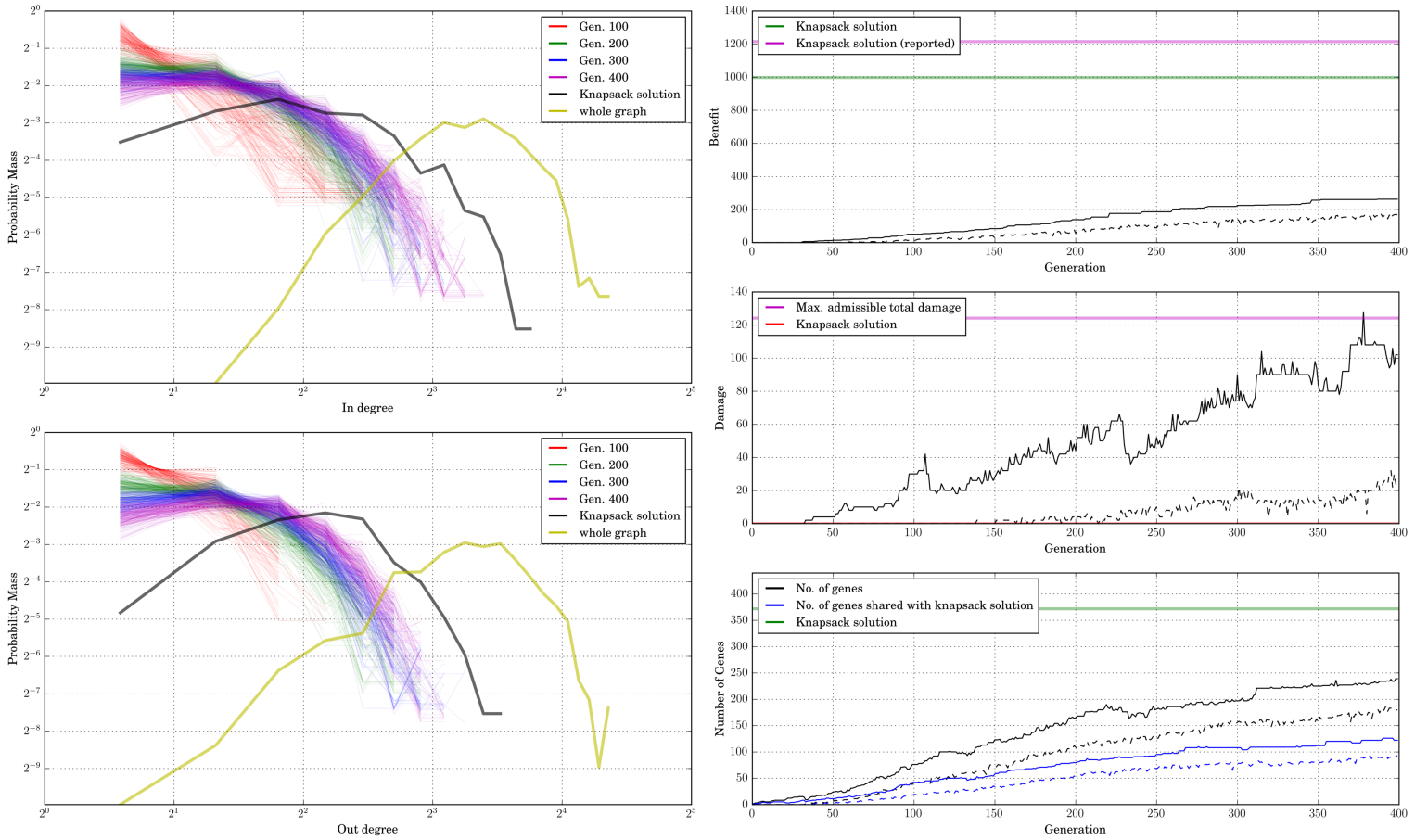

Figure 13: Similar experiment as in Fig. 1 but with reduced pressure $p = 0.2$. Note that the large number of neutral nodes slows down the convergance rate with effects consistent with general findings.

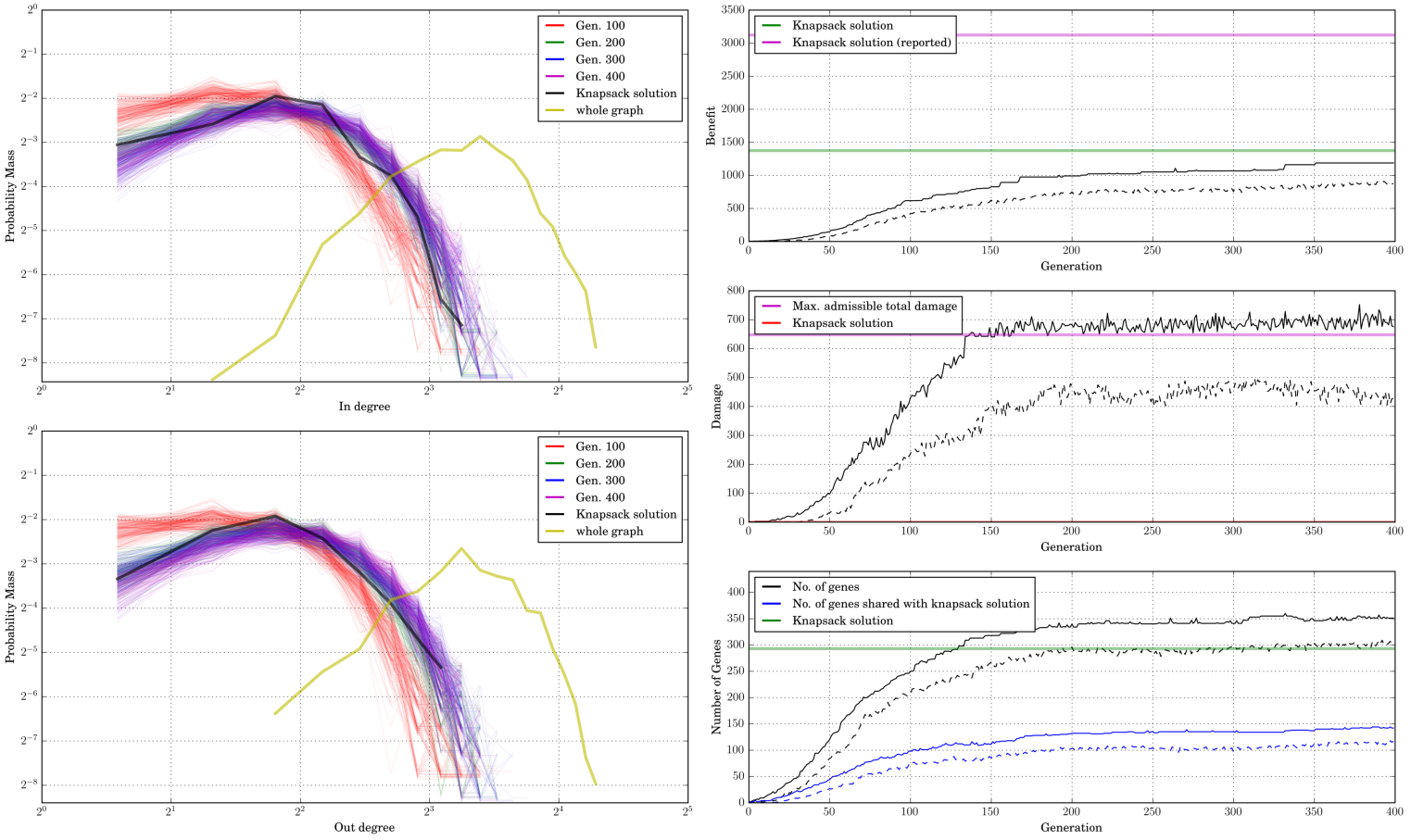

Figure 14: Similar experiment as in Fig. 1 but with reduced survivorship $s = 0.2$ (increased competition). Note that this significantly increases the convergence rate, as expected biologically, without changing the qualitative structure of the outcome (see Fig. 15 for the opposite scenario).

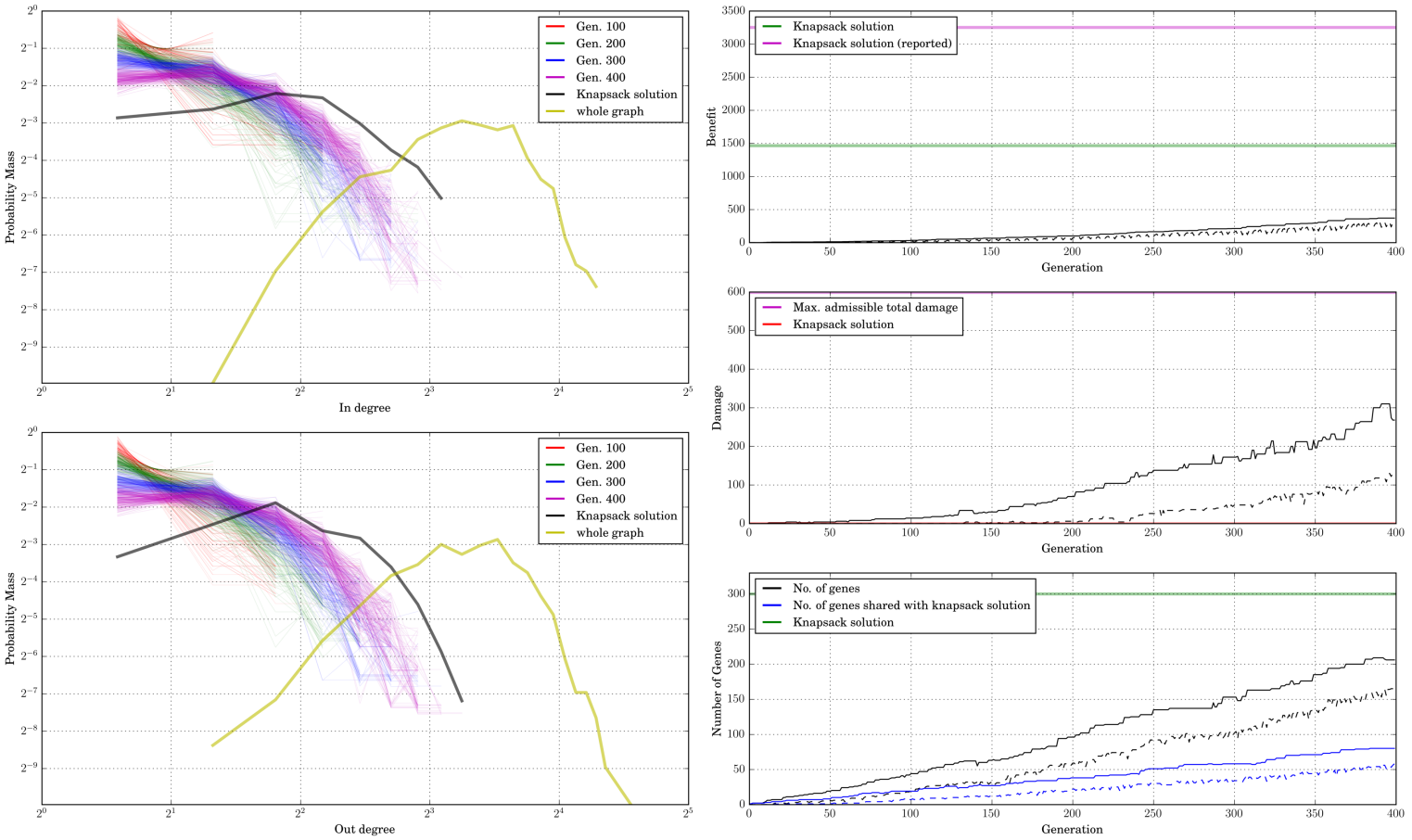

Figure 15: Similar experiment as in Fig. 1 but with increased survivorship (reduced competition) $s = 0.8$. Under a fixed capacity model, this necessarily reduces the number of mutations per generation and, thus, reduces the convergance rate.

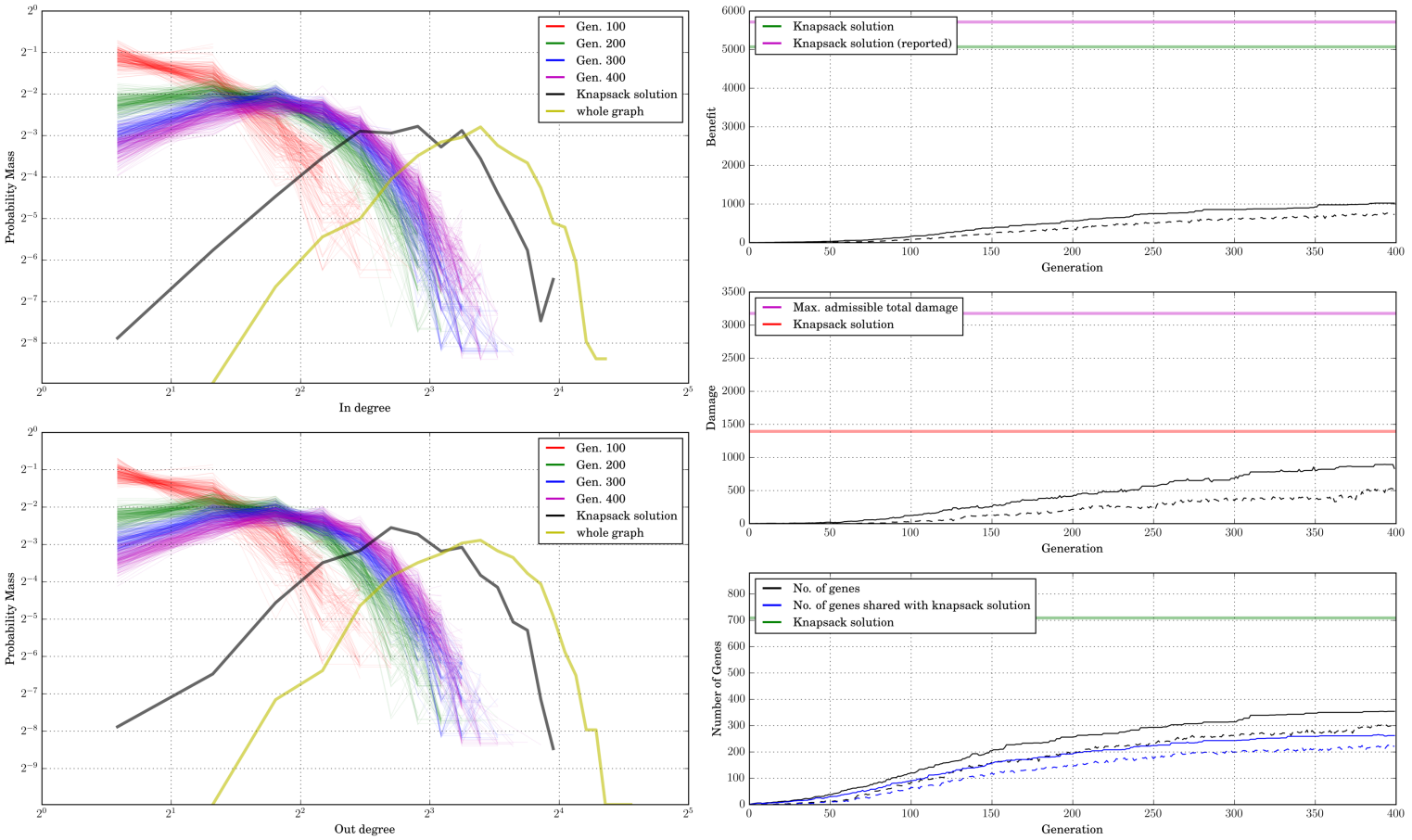

Figure 16: Similar experiment as in Fig. 1 but with increased tolerance $t = 0.5$. Naturally, this implies a larger size of optimized subgraphs. Note the approach of degree distributions to the whole graph distribution confirming the general findings.

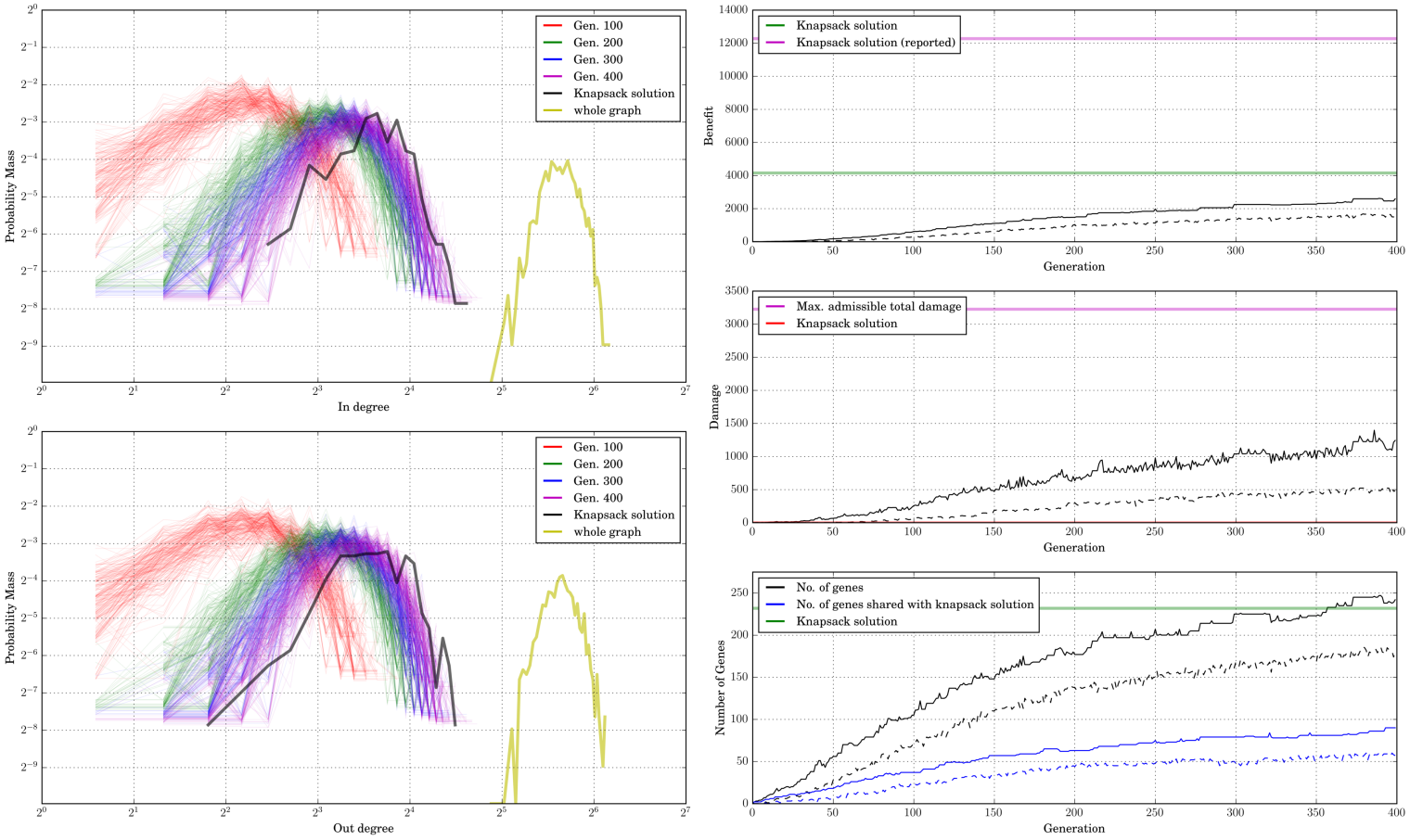

Figure 17: Similar experiment as in Fig. 5 but with fitness defined as $f(I_k) = \sum b_i - \sum d_i$ instead of $f(I_k) = \sum b_i$. No qualtiative difference can be observed.

References

[1] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Reviews of Modern Physics* 74.1 (2002), p. 47.

[2] Reka Albert. "Scale-free networks in cell biology". In: *Journal of Cell Science* 118.21 (2005), pp. 4947–4957.

[3] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *Science* 286.5439 (1999), pp. 509–512.

[4] Eli Eisenberg and Erez Y Levanon. "Preferential attachment in the protein network evolution". In: *Physical Review Letters* 91.13 (2003), p. 138701.

[5] Eugene V Koonin. "Are there laws of genome evolution?" In: *PLoS Computational Biology* 7.8 (2011).

[6] Ron Milo et al. "Network motifs: simple building blocks of complex networks". In: *Science* 298.5594 (2002), pp. 824–827.

[7] Miguel Nicolau and Marc Schoenauer. "On the evolution of scale-free topologies with a gene regulatory network model". In: *Biosystems* 98.3 (2009), pp. 137–148.

[8] David Pisinger. "Where are the hard knapsack problems?" In: *Computers and Operations Research* 32.9 (2005), pp. 2271–2284.

[9] Mohammed Shamrani, François Major, and Jérôme Waldispühl. "Evolution by Computational Selection". In: *CoRR* abs/1505.02348 (2015). URL: http://arxiv.org/abs/1505.02348.

[10] Steven H Strogatz. "Exploring complex networks". In: *Nature* 410.6825 (2001), pp. 268–276.

[11] Zhi Wang and Jianzhi Zhang. "In search of the biological significance of modular structures in protein networks". In: *PLoS Computional Biology* 3.6 (2007).

[12] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world' networks". In: *Nature* 393.6684 (1998), pp. 440–442.