GP and Bloat: Absorbing boundaries and spatial structures

P.A. Whigham & G. Dick

Spatial Information Research Centre Information Science Dept., Univ. of Otago, Dunedin, New Zealand ph: +64 3 4797391 pwhigham@infoscience.otago.ac.nz

Abstract. This paper examines the behaviour of bloat for GP tree structures using three different topologies: a panmictic, ring and star structure. Initially genetic drift is examined and the results showing the influence of a lower absorbing boundary are examined for each space. A simple selection model is then applied and analysed for bloat. A conjecture regarding the influence of inbreeding, due to spatial structure, is presented as one mechanism for bloat reduction. The paper shows that spatially-structured GP results in a tradeoff between convergence, diversity and the size of individuals.

1 Introduction

The concept of "bloat" in Genetic Programming(GP) is a well established phenomena [1–8] characterised by variable-length genomes gradually increasing in size during evolution. A variety of causes for this effect have been shown, including as a protective response to destructive operators [7], selection pressure [9, 5] and operator bias [10]. Three main methods for controlling bloat are commonly proposed [5]: set an upper bound to the complexity of individuals in the population [11]; introduce an explicit fitness penalty (parsimony measure) that biases against larger individuals [12, 13]; and apply genetic operators designed to target redundant code or that bias against offspring size increases [14, 10].

Other work has considered an alternative approach to reducing bloat via the use of spatially-structured populations [15–17]. Since spatial structure generally slows the effect of drift [18] and forms small local demes for breeding, these types of models have shown some promise in reducing the rate of bloat. Although not explicitly explored in these papers, spatial structure often increases inbreeding, which can lead to local convergence and therefore reduced selection pressure. This may be partly an explanation for the previously observed bloat behaviour and will be addressed in more detail in this paper.

Recently there have been a number of theoretical advances in understanding bloat [19, 6, 20–22] which have confirmed that selection pressure is one fundamental driver in bloat creation. For example, Poli [19] reduced bloat by a stochastic approach to setting the fitness of individuals to zero who were above the average length of the population.

Stringer and Wu [20, 21] showed that a shrinking effect on genome length occured for a chunking GA once the population had essentially converged and selection had become random. Skinner et al. [22] provided a theoretical argument for this observed tendency of variable-length genomes to shrink when selection is not considered (i.e. under the process of genetic drift). This work implied that the presence of an absorbing boundary (a genome size that once reached cannot change) contributed to the reduction in average size of a population under drift.

This paper will revisit the results of Skinner et al. [22] and examine the behaviour of this model when the population is structured as a network of spatial locations. Under genetic drift the results show that for both the ring and star topologies there is an increase in small tree sizes compared with a panmictic space. Subsequently, a simple selection model applied to these topologies shows that inbreeding may well contribute to a reduction in bloat, although with a subsequent tradeoff in evolutionary dynamics. The paper is organised as follows: §2 describes the basic GP model, introduces the graph-based approach to spatial structures, and describes the initial experiments; §3 presents a simple model of selection to examine the influence of spatial structures on bloat; §4 discusses the findings and implications of this work; and §5 concludes the paper with some directions for future work.

2 Genetic Drift, Bloat and Space

The original paper by Skinner et al. [22] based their analysis on a general model of crossover, where an individual was represented by a single number corresponding to the number of genes. Crossover was performed by selecting a random number of genes from each (randomly selected) parent and swapping the genes to create two offspring. Each individual was initially created with the same number of genes (20). The resulting model, and those of Stringer and Wu [20, 21], showed that without selection the average number of genes in the population reduced over time. Our initial interest is to examine these results using a GP tree model and examine in what ways spatial structure changes the response of this shrinking genome.

2.1 Spatial Structure and Drift

Spatial structure for a single population is defined as follows: a population of N individuals reside at N locations, labelled i = 1, 2, ... N. Neighbourhood structure is defined as a network, which can be represented as a square connectivity matrix, M, where an entry of 0 in M indicates that the two locations are not connected, and a value of 1 indicates that the locations are in the same deme. This matrix defines the connectivity between locations (a value M[i, j] = 1 indicates that location j is in the deme of location i), which defines the local deme structure used to determine the breeding pool for each location. For each generation, each location i is visited, and two parents are chosen at random based on M. Crossover is applied to these parents to produce two children with one child being randomly selected to produce a new individual at location i. Once all locations have been visited the current generation is replaced by the next generation. Hence this describes a two parent generational model with replacement. For example,



Fig. 1. Inbreeding (F_{st}) for panmictic, ring and star spaces.

a panmictic (fully-connected) model of space would have each entry in M[i, j] = 1. Previous work by the authors [23, 18] has shown that the time to fixation for a neutral allele depends on the spatial structure of the population: a regular structure such as a ring increases fixation time compared with a panmictic space, whereas a space dominated by one individual, such as a star, has fixation times well below that of a panmictic space (see Fig. 1 for example ring and star structures). Note that these networks are undirected, and therefore M is symmetrical.

2.2 The F-statistic and inbreeding

The F-statistic can be used to measure the degree of inbreeding for a population by describing the reduction in heterozygosity when compared with the expected loss predicted by a panmictic population under Hardy-Weinberg equilibrium [24]. Fig. 1 shows the F-statistic for panmictic, ring and star structures for a two-allele, one locus diploid model due to genetic drift. For the ring structure each location has a deme size of 3, being connected only to its immediate neighbours. The star structure has one location (location 0) with the entire population in its deme, while all other locations are only connected to location 0 and themselves. Hence the ring represents a highly distributed spatial structure, while a star has characteristics similar to a scale-free network. Fig. 1 shows that the ring and star structures both have greater inbreeding than a panmictic population. The star inbreeding F_{st} reduces after a short number of generations, since the population rapidly converges to a single homozygote. In comparison, the ring maintains a high inbreeding measure since the convergence of a ring is slow but small local demes rapidly form homogeneous blocks [18]. Hence a population with a star structure will amplify both drift and selection more rapidly then a ring or panmictic structure, while a ring is likely to maintain diversity for a larger number of generations compared with either a panmictic or star, with high local deme inbreeding.

2.3 Experimental Setup

Each individual of the population is represented as a binary tree. Following the model of Skinner et al. [22] the individuals are initialised with 20 nodes, although the shape of the trees are randomly generated to produce a mixture of tree depths, as defined in Algorithm 1. Here the function BuildTree takes the same form as Algorithm 1 to construct the nodes below the top node, resulting in individuals with a mixture of shapes. The evolution of this population of individuals, given a spatial structure M, is

input : Number of nodes n
output: Binary Tree individual t
1 t ← newNode();
2 n ← n - 1;
3 lsize ← RandomInt(n);
4 rsize ← n - lsize;
5 tleft ← BuildTree(t,lsize);
6 tright ← BuildTree(t,rsize);
7 return t;

Algorithm 1: Creating an individual for the initial GP tree population

described in Algorithm 2. The main points to note are: parents are randomly selected with replacement from the previous generation, defined by the connectivity matrix for each location (lines 6 and 7); and the crossover operation randomly selects one offspring to be placed at location i of the next generation (line 8). In addition, crossover is limited so any tree created from crossover must have at least 2 nodes. For each type of space

```
input : Population P_0 at time zero, Connectivity Matrix W, Max Generations G
   output: P_t after t time steps
 1 gen \leftarrow 0:
2 while gen < G do
        gen \leftarrow gen + 1;
3
4
        P_{gen} = null;
        foreach Location I_i \in W do
5
             p_1 \leftarrow \texttt{SelectParent}(P_{gen-1}, W_i);
6
             p_2 \leftarrow \texttt{SelectParent}(P_{gen-1}, W_i);
7
             P_{gen}(I_i) \leftarrow Crossover(p_1, p_2);
8
 9
        end
10 end
11 return P_{gen};
```

Algorithm 2: Evolution using a spatial structure under neutral selection

M a population size of 100 was initially created and the population evolved for 1000 generations. The results were averaged over 100 runs.

2.4 Drift Results

The left panel of Fig. 2 shows the average evolution of tree sizes for a panmictic space, and confirms the previous research [22] that indicated neutral selection produces a ten-



Fig. 2. The evolution of tree size distribution due to drift for each space, averaged over 100 runs.



Fig. 3. The average tree size for each population, averaged over 100 runs, and the frequency of trees of size 3.

dency for the size of individuals to decrease towards the absorbing boundary. This process becomes more interesting when a spatial structure is imposed on the population. The results for each space, as shown in Fig. 2, confirm the assumption that changing the interaction dynamics of the population would change the drift bloat behaviour. Surprisingly, Fig. 2 shows that the ring and star spaces both have a more rapid decrease in tree size compared with the panmictic population. Although the ring and star are very different forms of spatial structure they both appear to have similar characteristics. For example, Fig. 3 shows the average size of the population and the frequency of trees of size 3. The panmictic space has on average the largest tree sizes, although there is a large variance (> 20) in this average measure. Note that the panmictic space has a close to linear increase in trees of size 3, whereas the ring and star both exhibit an early exponential increase in these tree sizes. Here we have chosen the tree size 3 since this was the most common lowest tree size due to the bias imposed by the crossover operator. The star produced on average the smallest tree sizes once the generations passed 400 with the ring being intermediate and the panmictic space being the largest.

One conjecture regarding this pattern is that inbreeding is reducing bloat. Fig. 4 adds some support for the inbreeding conjecture, however it needs to be noted that the ring produced the largest tree structures, even though over the 100 runs all ring tree sizes were approaching the lower absorbing boundary by the end of 1000 generations. This figure also shows that the star is definitely reducing bloat, although in the upper panel there are several large trees created prior to generation 400. This can be understood in terms of the diffusion characteristics of the ring and star. For the ring local demes may



Fig. 4. Distribution of largest trees each generation for 100 runs of each space. The upper figures show the distribution up to tree size 20000. The lower figures show the distribution for sizes up to 50.

continue to inbreed with large trees, however eventually those trees in distant parts of the ring that have converged towards the absorbing boundary gradually diffuse through the population. The star structure behaves in a different manner due to the dominant central location. Since this location participates in all demes, but all other locations only have themselves and the central location in their deme, when a small tree is placed at the central location it will rapidly propagate to other locations, and therefore amplify the effect of the absorbing boundary. Hence the inbreeding and rapid diffusion characteristics of the star result in smaller tree sizes rapidly taking over the population. Note also that the panmictic space was producing some large trees after 1000 generations, which shows that without an explicit upper absorbing boundary there is still some (albeit small) probability of producing larger trees. Surprisingly, this was not explicitly noted in previous work [22].

It is also informative to examine how space affects bloat for a typical run. Fig 5 shows a typical run for each spatial structure, and shows the well-mixed behaviour of the panmictic space versus the localised deme structures of the ring. The ring behaviour can be understood in terms of each deme performing a local random walk in tree size with an absorbing boundary. Note for example the white line down the centre of Fig. 5(b), which shows a local deme with at least one large tree, which gradually reduces in size due to the drift associated with the neighbouring demes. In comparison, Fig. 5(c) shows pulses of tree size increasing and decreasing, based on the centre location individual (here drawn on the left hand size of the rectangular grid). When the centre location holds a large tree this rapidly propagates, however once a small tree is produced in the centre this tends to dominate the behaviour of the population and drives the tree sizes towards the lower absorbing boundary.



Fig. 5. Single typical run for panmictic, ring and star populations. A grey-scale is used for size, with white indicating an individual with size > 40 and black indicating size ≤ 3 .

3 Selection and Bloat

This section will examine a simple model of selection and the resulting bloat behaviour for panmictic, ring and star spaces. The selection model is very simple: the initial population is constructed as for the drift model (i.e. each tree is initially size 20), and each tree node is randomly given an allele value of either 0 or 1. For all examples the probability of setting a tree node to zero is 0.9. The fitness of an individual is based on the proportion of ones in the tree versus the size of the tree. For example, a tree with 10 nodes and 3 one-alleles would have a fitness of 0.3. This form of fitness should not excessively promote bloat, since it is possible to have the maximum fitness independent of tree size. However, there is some bias to producing larger trees since adding a subtree composed of one-alleles will increase the fitness of the tree, even though it increases in size. Of course there is also the alternative possibility, where a subtree containing a 0 is replaced by a smaller subtree containing only one-alleles, in which case the fitness will also increase. Hence, although the fitness measure may promote bloat tree size reduction may also have increased fitness. For all experiments a population size of 100 was used, with 1000 generations and proportional fitness selection based on each local deme. In addition, one-point crossover was performed to produce all off-spring.

The evolution of size distribution, averaged over the 100 runs, is shown in Fig. 6, which compares favourably with the drift model. However, a closer comparison with Fig. 2 shows that the rate of increase in small trees for the panmictic space is faster under selection than drift for early generations. The average fitness behaviour for each space is shown in Fig. 7, and shows that average best solution and generations to discover a solution (i.e. a tree with all 1-nodes) are approximately the same for ring and panmictic spaces. However, the average fitness is lower for the ring, suggesting that the



Fig. 6. The evolution of tree size distribution due to selection for each space, averaged over 100 runs.



Fig. 7. Average and average-best fitness, along with the cumulative generation required to solve the selection problem over 100 runs.

diffuse spatial structure is maintaining the diversity of the population. The star, by amplifying selection and drift along with early inbreeding, causes the population to rapidly converge to a non-optimal solution. Hence the star only manages to solve the problem for approximately 20% of the runs, and if this occurs it generally happens in an early generation when tree sizes are generally still quite small.

Of more interest is the average size of the trees in the population and the frequency of small trees produced during evolution. These results are shown in Fig 8 and show the somewhat surprising result that the panmictic average size continues to increase after the point where the majority of solutions have been discovered (generation 400)



Fig. 8. The average tree size for each population, averaged over 100 runs, and frequency of trees of size 3.



Fig. 9. Distribution of largest trees each generation for 100 runs of each space. The upper figures show the distribution up to 20000. The lower figures show the distribution for sizes up to 50.

until about generation 500, at which time the average size behaviour is suggestive of a random walk. There are several possible explanations for this behaviour: the trees in the population at around generation 400 may be large enough that the lower absorbing boundary does not affect the behaviour; the selection pressure for those runs which have not solved the problem by generation 400 begin to select for trees with large one-allele subtrees, thus increasing their fitness but never producing a perfect solution; and the average measure is actually not representative of the true dynamics of the population runs, since the variance of these measures was very high. In comparison the average size for the ring is much lower, and although the size has increased from the initial tree sizes of 20 the increase in tree sizes appears to be much slower. Once again, the star produced the smallest tree sizes, due to the initial inbreeding and rapid convergence of the population. These dynamics for each space are further demonstrated in Fig. 9, which shows that the panmictic space has a trend towards increasing tree sizes as the generations continue. However, although the ring produces some large trees there appears to be a bias towards smaller trees in general.

4 Discussion

The previous work with panmictic, ring and star topologies have shown that changing the spatial structure alters bloat behaviour, and therefore confirms the work of several authors [15–17] for more complex spaces. However, the argument that absorbing boundaries are the main driver of bloat behaviour needs to be reconsidered. For example, although there is clearly a bias towards small trees creating smaller trees, there is also a tendency for tree sizes to expand (large trees making larger trees). If there is no upper absorbing boundary then trees sizes may become quite large. Of course, under

drift, as the number of generations increase the lower boundary will eventually drive the population towards small tree sizes, as shown in Fig. 4.

The frequency of small tree sizes has been shown in this paper to occur more rapidly for a ring structure under both drift and selection, compared with a panmictic space. In addition, the average tree size grows more slowly with a ring structure than a panmictic space. Since a ring structure reduces drift, maintains diversity and produces local inbreeding this suggests that the previous work on island models were observing a bloat reduction due to a reduced effective population size. Our conjecture that inbreeding may reduce bloat is given some support through our results for both the ring and star structure, since they promote local convergence which allows the lower absorbing boundary to play a significant role in tree size. Of course, when selection is involved the star structure converges too rapidly to be useful for general problem solving, however the results for the ring suggest that this is one possible mechanism to reduce overall tree size during evolution.

The previous work of Poli [19] showed some promise in reducing bloat by stochastically setting the fitness of above average sized individuals to zero. However, although this may reduce the rate of increase in tree sizes it is unlikely to decrease the average tree size since it will merely place additional selection pressure on other individuals, who will accordingly increase in size. The work presented here suggests an alternative approach when using panmictic populations by explicitly incorporating inbreeding into the selection process. For example, this could be achieved by a form of assortative mating, where individuals are tagged with an identity that produces some preferential parent selection. Alternatively, the use of a ring will clearly reduce overall tree growth, which should reduce the cost of evaluation, and this could be incorporated along with reducing the fitness of above average sized individuals. A combination of these strategies may be one approach to reducing bloat without explicit size constraints being imposed on the population, while still achieving acceptable performance for difficult problems.

5 Conclusion

This paper has explored the results from Skinner et al. [22] in the context of spatiallystructured evolutionary algorithms. The results both confirmed and illuminated several properties of the absorbing boundary hypothesis. In particular, the lower absorbing boundary was shown to be effective at reducing overall tree sizes, however large trees were still able to be occasionally produced.

The use of the ring spatial structure produced small trees and still performed well for a simple selection model, which suggests that the rate of bloat may be reduced by incorporating spatial structure that promotes inbreeding and slow diffusion of solutions. In addition, we have suggested that inbreeding could also be used directly as a form of bloat reduction and could be incorporated into panmictic space along with stochastic reduction in fitness for above average-sized individuals. Bloat appears to be a natural consequence of selection, and if there is no upper absorbing boundary it is likely that tree sizes will continue to grow. However, our results suggest that it may generally be reduced by incorporating certain types of spatial structure, such as a ring. Future work is required to investigate the inbreeding conjecture by a more extensive examination of population size dynamics and through the application of more difficult and deceptive problems for a variety of spatial structures. In addition, the use of assortative mating approaches to GP and their influence on bloat needs to be investigated so that the properties of diffuse spatial structures and bloat reduction can be properly understood.

References

- Tackett, W.: Genetic programming for feature discovery and image discrimination. In Forrest, S., ed.: Proceedings of the Fifth International Conference on Genetic Algorithms, Illinois, USA, Morgan Kaufmann (1993) 303–309
- Angeline, P.: Genetic programming and emergent intelligence. In Kinnear Jr., K., ed.: Advances in Genetic Programming. MIT Press (1994) 75–98
- Langdon, W.: Evolving data structures using genetic programming. In Eshelman, L., ed.: Proceedings of the sixth International Conference on Genetic Algorithms, Pittsburgh, USA, Morgan Kaufmann (1995) 295–302
- Soule, T., Foster, J., Dickinson, J.: Code growth in genetic programming. In Koza, J., Goldberg, D., Fogel, D., Riolo, R., eds.: Genetic Programming 1996: Proceedings of the First Annual Conference, Stanford University, CA, MIT Press (1996) 215–223
- Langdon, W.B., Poli, R.: Fitness causes bloat: Mutation. In Clack, C., Vekaria, K., Zin, N., eds.: ET'97 Theory and Application of Evolutionary Computation, University College London, UK (1997) 59–77
- Banzhaf, W., Langdon, W.: Some considerations on the reason for bloat. Genetic Programming and Evolvable Machines 3(1) (2002) 81–91
- Soule, T., Heckendorn, R.: An analysis of the causes of code growth in genetic programming. Genetic Programming and Evolvable Machines 3(3) (2002) 283–309
- Brameier, M., Banzhaf, W.: Neutral variations cause bloat in linear GP. In Ryan, C., Soule, T., Keijzer, M., Tsang, E., Poli, R., Costa, E., eds.: Genetic Programming: 6th European Conference, EuroGP 2003, Springer Berlin/Heidelberg (2003) 286 – 296
- Langdon, W.B., Poli, R.: Fitness causes bloat. Technical Report CSRP-97-09, University of Birmingham, School of Computer Science, Birmingham, B15 2TT, UK (1997)
- Luke, S.: Modification point depth and genome growth in genetic programming. Evolutionary Computation 11(1) (2003) 67–106
- Koza, J.: Genetic Programming: On the Programming of Computers by Natural Selection. MIT Press, Cambridge MA, USA (1992)
- Zhang, B.T., Mühlenbein, H.: Evolving optimal neural networks using genetic algorithms with Occam's razor. Complex Systems 7 (1993) 199–220
- Jong, E.D., Watson, R., Pollack, J.: Reducing bloat and promoting diversity using multiobjective methods. In Spector, L., Goodman, E., Wu, A., Langdon, W., Voigt, H., Gen, M., Sen, S., Dorigo, M., Pezeshk, S., Garzon, M., Burke, E., eds.: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO-2001, Morgan Kaufmann (2001) 11–18
- Terrio, M., Heywood, M.I.: Directing crossover for reduction of bloat in GP. In Kinsner, W., Seback, A., Ferens, K., eds.: IEEE CCECE 2003: IEEE Canadian Conference on Electrical and Computer Engineering, IEEE Press (2002) 1111–1115
- Rochat, D., Tomassini, M., Vanneschi, L.: Dynamic size populations in distributed genetic programming. In Keijzer, M., Tettamanzi, A., Collet, P., van Hemert, J.I., Tomassini, M., eds.: Proceedings of the 8th European Conference on Genetic Programming. Volume 3447 of Lecture Notes in Computer Science., Lausanne, Switzerland, Springer (2005) 50–61

- de Vega, F.F., Gil, G.G., Pulido, J.A.G., Guisado, J.L.: Control of bloat in genetic programming by means of the island model. In Yao, X., Burke, E., Lozano, J.A., Smith, J., Merelo-Guervós, J.J., Bullinaria, J.A., Rowe, J., Kabán, P.T.A., Schwefel, H.P., eds.: Parallel Problem Solving from Nature - PPSN VIII. Volume 3242 of LNCS., Birmingham, UK, Springer-Verlag (2004) 263–271
- Fernandez, F., Galeano, G., Gomez, J.A.: Comparing synchronous and asynchronous parallel and distributed GP models. In Foster, J.A., Lutton, E., Miller, J., Ryan, C., Tettamanzi, A.G.B., eds.: Genetic Programming, Proceedings of the 5th European Conference, EuroGP 2002. Volume 2278 of LNCS., Kinsale, Ireland, Springer-Verlag (2002) 326–335
- Whigham, P., Dick, G.: Fixation of neutral alleles in spatially structured populations via genetic drift: Describing the spatial structures of faster-than-panmictic configurations. In Whigham, P., ed.: Proceedings of the 17th Annual Colloquium of the Spatial Information Research Centre (SIRC), Dunedin, New Zealand, Otago University Press (2005) 81–90
- Poli, R.: A simple but theoretically-motivated method to control bloat in genetic programming. In Ryan, C., Soule, T., Keijzer, M., Tsang, E., Poli, R., Costa, E., eds.: Genetic Programming, Proceedings of EuroGP'2003. Volume 2610 of LNCS., Essex, Springer-Verlag (2003) 204–217
- Stringer, H., Wu, A.: Bloat is unnatural: An analysis of changes in variable chromosome length absent selection pressure. Technical Report CS-TR-04-01, University of Central Florida, University of Central Florida (2004)
- Stringer, H., Wu, A.: Winnowing wheat from chaff:the chunking GA. In Deb, K., Poli, R., Banzhaf, W., Beyer, H.G., Burke, E., Darwen, P., Dasgupta, D., Floreano, D., Foster, J., Harman, M., Holland, O., Lanzi, P.L., Spector, L., Tettamanzi, A., Thierens, D., Tyrrell, A., eds.: Genetic and Evolutionary Computation – GECCO-2004, Part II. Volume 3103 of Lecture Notes in Computer Science., Seattle, WA, USA, Springer-Verlag (2004) 198–209
- Skinner, C., Riddle, P.J., Triggs, C.: Mathematics prevents bloat. In Corne, D., Michalewicz, Z., Dorigo, M., Eiben, G., Fogel, D., Fonseca, C., Greenwood, G., Chen, T.K., Raidl, G., Zalzala, A., Lucas, S., Paechter, B., Willies, J., Guervos, J.J.M., Eberbach, E., McKay, B., Channon, A., Tiwari, A., Volkert, L.G., Ashlock, D., Schoenauer, M., eds.: Proceedings of the 2005 IEEE Congress on Evolutionary Computation. Volume 1., Edinburgh, UK, IEEE Press (2005) 390–395
- Dick, G., Whigham, P.: The behaviour of genetic drift in a spatially-structured evolutionary algorithm. In: 2005 IEEE Congress on Evolutionary Computation, IEEE Press (2005) 1855– 1860
- 24. Wright, S.: The interpretation of population structure by f-statistics with special regard to systems of mating. Evolution **19** (1965) 395–420