Discovering Population Structures with Extreme Fixation Rates via Evolutionary Search

Grant Dick and Peter A. Whigham

Spatial Information Research Centre
University of Otago. Dunedin, New Zealand
Phone: +64 3 479-5282 Fax: +64 3 479-8311
Email: gdick@infoscience.otago.ac.nz

Presented at SIRC 2005 - The 17th Annual Colloquium of the Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
November 24th-25th 2005

ABSTRACT

Genetic drift is a well known and important force in directing the evolution of a population. The nature of genetic drift in panmictic populations is well understood, and new research is shedding light on the behaviour of genetic drift. This paper explores the concept of using evolutionary algorithms to search for population structures that exhibit the minimum and maximum conditions for loss of variation via genetic drift. Two spatial structures repeatedly emerge as candidates: a star topology that reduces fixation time to a logarithm of population size, and a "line with islands" topology that can delay fixation via genetic drift to a greater extent than any previously known population structure.

Keywords and phrases: genetic drift, spatially-structured populations, hyperfixation, evolutionary algorithms, multimodal optimisation, niching methods

1 Introduction

Genetic drift is considered an important force in the evolution of populations, with entire branches of population genetics devoted to its study (Kimura 1983). Genetic drift is the effect of sampling errors that are present in the process of mating. Genetic drift is the force applied to a population as a result of errors in the stochastic sampling process. A direct consequence of genetic drift is that a finite population, in the absence of mutation, will always converge to a single allele state (Crow & Kimura 1970).

Space plays an important role in the evolution of species. This has long been known and space was suggested by Darwin as being a pivotal component in promoting speciation (Darwin 1859). This paper examines one particular method of incorporating space into an evolving population. Spatially-structured populations are modelled as individual-based representations of space. A geography is introduced into the population and mating is restricted to networks of closely-located individuals (demes). Each individual can belong to multiple demes, allowing genetic information to propagate throughout space.

The introduction of geography into a population can dramatically change the behaviour of genetic drift. Intuitively, the time to loss or fixation of an allele (drift time) can be greatly increased with the introduction of space (Dick & Whigham 2005). However, certain configurations of space can lead to *hyperfixation*, where drift time is significantly reduced (Whigham & Dick 2005). This paper attempts to address the following questions: Under which configuration(s) of space is hyperfixation at its extreme? Conversely, is there a spatial structure that is most resistant to the effects of genetic drift? This paper uses an evolutionary algorithm to explore the search space of all possible population structures. The results suggest the form of two population structures that appear to be at the minimum and maximum of fixation times for a given population size.

The remainder of this paper is structured as follows: §2 the concept of exploring a search space via evolutionary methods is presented;§3 presents the experimental component of this paper and the results are discussed in §4. Finally, a discussion of the findings of this paper and suggested directions of future work are presented in §5.

2 Evolutionary Search Methods

The search space of this paper is very large, in the order of 2^{N^2} for a given population size N, and as such it is simply too large to search iteratively. An alternative to iterative search is to use an evolutionary algorithm (EA).

EAs explore several points in a search space simultaneously. They use the concept of Darwinian natural selection and inheritance to drive the population towards desirable regions in a search space (Holland 1992, Goldberg 1989).

The search spaces presented in this paper are multimodal in that there are potentially many population structures that have equally desirable traits. Simple EAs has been shown to have difficulty in searching mutimodal problem domains. This has lead to the development of niching evolutionary algorithms, which attempt to evolve subpopulations of individuals around desirable peaks in the search space. The use of niching EAs increase the useful diversity present in the population and helps to reduce the likelihood of the system converging on a suboptimal solution. This paper uses a niching EA that is a variant of the simple genetic algorithm known as deterministic crowding (DC) (Mahfoud 1992). It is an extension of a technique first used by DeJong to help promote diverse populations (DeJong 1975). The method for performing DC is shown in Algorithm 1. The basic premise of DC is that pairs of individuals are recombined to create offspring. These offspring then replace their closest parent if they are of greater fitness.

input : A population of individuals of size N
output: The same population with reproduced individuals

```
1  Shuffle population;
2  for i ← 1 to (N − 1) step 2 do
3      p1 ← population[i];
4      p2 ← population[i + 1];
5      (c1, c2) ← reproduce(p1, p2);
6      if [ distance(p1, c1) + distance(p2, c2) ] ≤ [ distance(p1, c2) + distance(p2, c1) ]
       then
7          if fitness(c1) > fitness(p1) then p1 ← c1;
8          if fitness(c2) > fitness(p2) then p2 ← c2;
9      else
10         if fitness(c2) > fitness(p1) then p1 ← c2;
11         if fitness(c1) > fitness(p2) then p2 ← c1;
12     end
13 end
14 return population;
```

Algorithm 1: The Deterministic Crowding Algorithm

3 Methodology

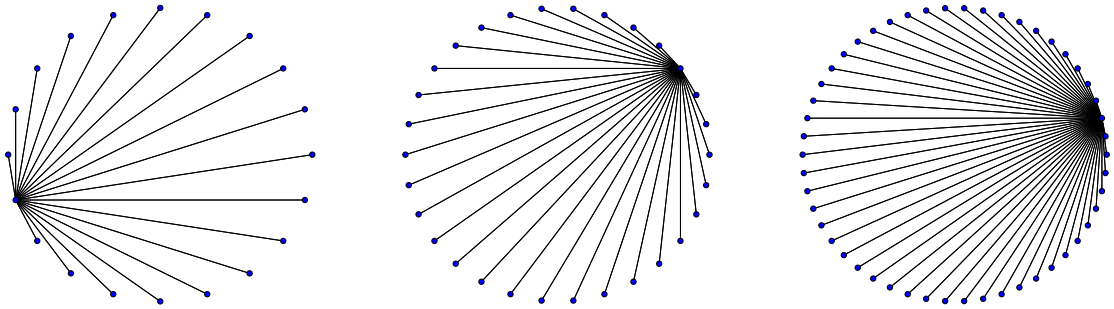
The spatial structures considered in this paper are undirected graphs. Each vertex in the graph represents a location within the population at which an individual can live. Deme participation is represented by edges between two vertices. For a set of N vertices, there are $N^2 - N$ possible graphs[1].

The EA represents graph connectivity as a bitstring. Each bit in the string represents an edge (or lack thereof) between two vertices. For an N-vertex graph, each bitstring contains $N^2 - N$ bits. Two-point crossover was used with bit-flipping mutation. An additional macro-mutation operator was also used. This replaced all the edges from one vertex in a graph with a randomly generated set. This operator was applied with probability of $1/(EA\ population\ size)$.

The fitness function for the EA was the resultant drift time for the graph that resulted from the decoded individual. When trying to find the fastest graph in terms of fixation time, this became a minimisation problem. Conversely, to find the slowest graph required maximisation of fitness. Genetic drift is a stochastic process, which means there is noise present in the evaluation of an individual. To overcome this noise, multiple trials of genetic drift need to be performed. The fitness of an individual is therefore the mean result of fixation via genetic drift over 1000 independent runs. This naturally leads to a computationally expensive evaluation process. As a consequence, only graph sizes of $20, 30$ and 50 were considered. As will be shown, these population sizes are still sufficient to provide some useful insights into the nature of genetic drift in spatially structured populations.

The EA used a population size of 100 for the 20 and 30-vertex graphs. A larger population size of 400 was used for the larger graph. The population was constructed with the following ratios: 50% of the population was initialised randomly, 25% of the population was initialised as fully-connected, panmictic graphs and the final 25%

[1]The number of feasible graphs is actually lower than this as only fully connected graphs are considered here. Any disconnected graphs produced by the system will be discarded automatically and replaced by their closest parent as per the deterministic crowding algorithm.



(a) $N = 20$ (b) $N = 30$ (c) $N = 50$

Figure 1: Examples of the fastest population structures discovered by the GA.

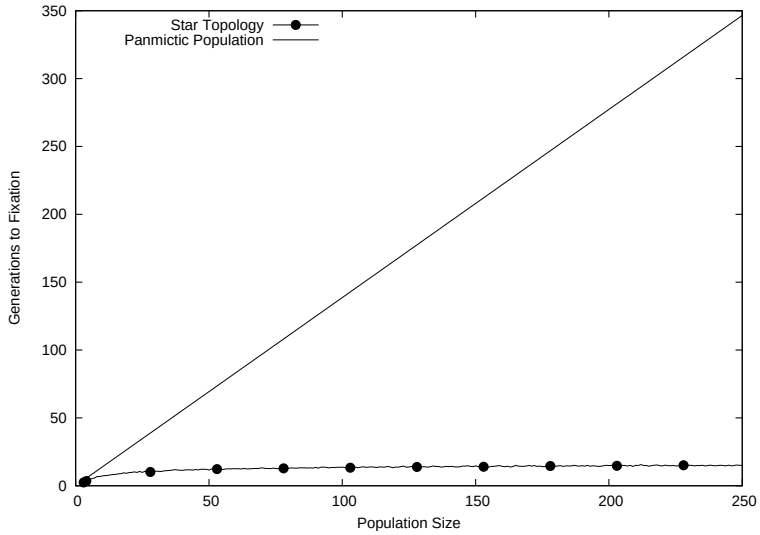


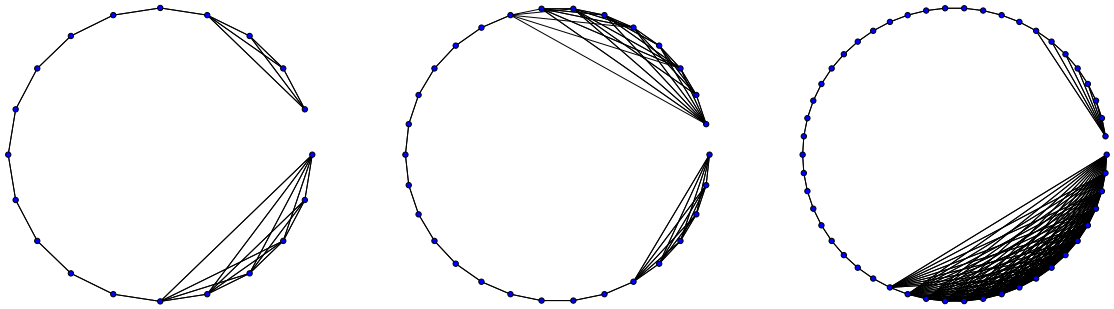
Figure 2: Comparison of "Star" population structure with existing population structures.

was initialised from "lines", where each vertex is arranged in one dimensional space and connected to each of its immediate neighbours. Each run of the EA performed 1000 generations. At the end of the run, the EA saved the best graph to disk for later inspection.

4 Results

Several runs of the EA were performed on $20, 30$ and 50-vertex graphs to find the fastest population structure for genetic drift. Examples of the resultant graph for the three population sizes are shown in Figure 1. These graphs all demonstrate a common trend; each has a single vertex which is connected to every other vertex in the graph, and the remaining vertices in the graph share no additional edges. This is equivalent to the star topology demonstrated in other research (Whigham & Dick 2005). The majority of runs by the EA converged to this topology, which adds weight to the case for it being the fastest possible structure for fixation via genetic drift. As a comparison, Figure 2 shows the fixation time for a population structured with a star topology versus the equivalent panmictic population for increasing population sizes. In contrast to the linear increase in fixation time for panmixia, the time for loss of variation via genetic drift in a star topology increases logarithmically with respect to population size.

The problem of discovering the slowest fixating population structure displayed similar characteristics for each population size. Examples graphs for this case are shown in Figure 3. In each case, it appears that two fully connected islands form at opposite ends of a line of individuals. This structure would provide the population with two largely independent demes. While both demes are still connected, and hence must eventually converge to the same state, one can imagine the senario whereby the two large demes are effectively trying to converge to two different values. The population as a whole must wait for one deme to become fixated to one state, then



(a) $N = 20$ (b) $N = 30$ (c) $N = 50$

Figure 3: Examples of the slowest population structures discovered by the GA.

wait again for that allele value to take over the second deme. This sounds like an ideal case for delaying fixation globally. To test this, four populations of sizes $20, 30, 50$ and 100 were initialised as lines. The larger population size of 100 was included as a test to see if the given population structure would hold for a population size that was not investigated by the EA. The end vertices of each line were then fully connected to form panmictic islands separated by a line structure. The size of these islands was increased until the population was fully-connected. The mean drift time for each line-island configuration over 1000 independent trials was measured and compared with a line topology. The results are shown in Figure 4. Clearly, the "line with islands" model takes far longer to converge, with the ratio of island size to global population size that results in maximum fixation time being about one third. A smaller ratio means that the line component of the population dominates the fixation time. A larger ratio means that two larger islands are separated by a smaller line, which in turn reduces the distance between the islands making them more connected and rapidly reducing the fixation time.

5 Conclusion and Future Work

The nature of genetic drift in spatially-structured populations is an interesting and important area of research. Until recently, there has been an implied notion that all spatial structures increase the amount of time required for genetic drift to strip genetic variation from a population. Recent work has presented population structures that hyperfixate, in which the force of genetic drift requires less time to cause fixation than in an equivalent panmictic population (Whigham & Dick 2005). While this research gives some indication into the nature of genetic drift, it leaves one question unanswered; are there population structures that, for a given population size, will cause genetic drift to have the greatest (or equally, the least) effect in terms of fixation time? This paper has attempted to answer this question by searching the enitre space of populations via an evolutionary algorithm.

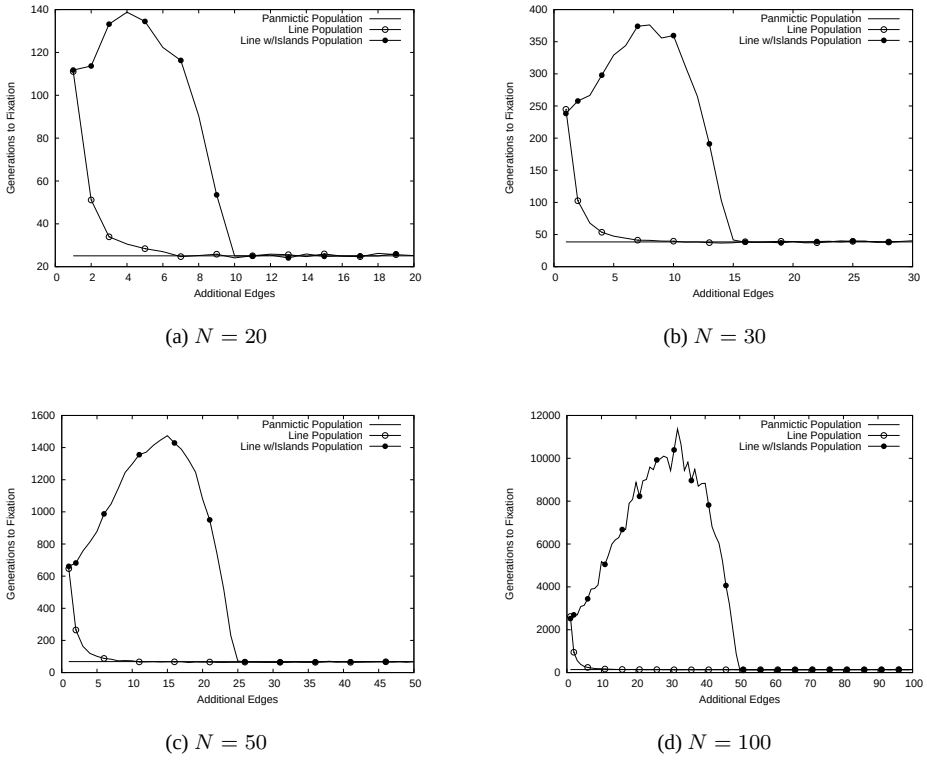
In the process of evolutionary search, two population structures have emerged. The first, a star topology, can reduce the time required for loss of genetic variation to a fraction of the equivalent panmictic population. Indeed, as the population size grows, the time for a star-structured population to fixate grows logarithmically. The second population structure consists of two fully-connected islands separated by a long line of individuals. Given a population size of 100, with islands in the region of 30 individuals, this topology can increase the time to fixation via genetic drift by a factor of 80 (compared to the equivalent panmictic population).

5.1 Future Work

There are several possible extensions that could be made to this work. The most obvious extension is to apply the techniques to larger population sizes. This will require significant computational effort, so future work should investigate more intelligent methods of reproduction and fitness evaluation in order to make searching of larger problem sizes feasible.

Reproduction in the EA used in this paper had a non-zero probability of producing a dicsonnected graph, resulting in an invalid, or "lethal" individual. When this happened, it was automatically replaced by one of its parents. There is some overhead in managing this. One interesting area of future work would be to incorporate a repair mechanism into the process to create valid offspring. Alternatively, the crossover operators could be replaced with new mechanisms that ensured that all offspring were viable.

Evolutionary algorithms do not always need a complete picture of fitness to work. Instead, what is more important is the relative fitness between individuals; the EA does not need know the extent to which an individual



(a) $N = 20$ (b) $N = 30$

(c) $N = 50$ (d) $N = 100$

Figure 4: Comparison of "Line-Island" population structure with existing population structures.

is fitter than another, only that the fitness of one exceeds the other. Currently, each individual is evaluated independently of the remaining population members. An alternative measure would be to have individuals "compete" with each other for fitness evaluation. Two individuals could perform genetic drift on their structures in parallel and when one fixates, a winner can be determined and the second simulation stopped. This process could repeat until the difference between the number of wins for each individual differs by a significant margin. This has the potential for significant savings in computational effort.

References

Crow, J. F. & Kimura, M. (1970). *Introduction to Population Genetics Theory*. Harper and Row. New York, Evanston and London.

Darwin, C. (1859). *On The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray. London.

DeJong, K. A. (1975). *An Analysis of the Behavior of a Class of Genetic Adaptive Systems*. PhD thesis University of Michigan Ann Arbor, MI. Dissertation Abstracts International 36(10), 5140B, University Microfilms Number 76-9381.

Dick, G. & Whigham, P. (2005). "The Behaviour of Genetic Drift in a Spatially-Structured Evolutionary Algorithm" *2005 IEEE Congress on Evolutionary Computation*. IEEE Press. pp. 1855–1860.

Goldberg, D. E. (1989). *Genetic Algorithms in Search Optimization and Machine Learning*. Addison-Wesley.

Holland, J. H. (1992). *Adaptation in Natural and Artifical Systems*. MIT Press. Cambridge, MA.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Mahfoud, S. W. (1992). "Crowding and preselection revisited" In R. Männer & B. Manderick (eds), *Parallel problem solving from nature 2*. North-Holland Amsterdam pp. 27–36.

Whigham, P. A. & Dick, G. (2005). "Fixation of neutral alleles in spatially-structured populations via genetic drift: Describing the spatial structure of faster-than-panmictic configurations" *The 17th Annual Colloquium of the Spatial Information Research Centre*. Dunedin, New Zealand pp. 81–90.