

Fixation of Neutral Alleles in Spatially Structured Populations via Genetic Drift: Describing the spatial structure of faster-than-panmictic configurations

Peter A. Whigham and Grant Dick

Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
Phone: +64 3 479-7391 Fax: +64 3 479-8311
Email: pwhigham@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

This paper considers spatially-structured populations described as a network, and examines the properties of these networks in terms of their affect on fixation of neutral alleles due solely to genetic drift. Individuals are modelled as two allele, one locus haploid, diploid and tetraploid structures. The time to fixation for a variety of network configurations is discovered through simulation. The concept of hyperfixation is introduced, which refers to when time to fixation for a network of n nodes occurs more rapidly than the corresponding panmictic n node structure. A hyperfixation index, h , is developed that attempts to characterise a spatial arrangement such that when $h < 1$ hyperfixation will occur. Issues regarding fixation with ploidy independence, and possible improvements to the described hyperfixation index are discussed.

Keywords and phrases: genetic drift, networks, neutral allele fixation, panmictic populations, hyperfixation

1 Introduction

The mean time to loss of variation (fixation) for a neutral allele at initial frequency p in a randomly mating, unstructured (panmictic) population composed of N monocious individuals with a ploidy of P is given by (Kimura & Ohta 1969):

$$\bar{t}(p, N, P) = -2NP[p \ln(p) + (1-p) \ln(1-p)]$$

where the value of P is one for haploids, two for diploids and so on. This paper considers the effect of imposing a spatial structure on the population thereby creating subpopulations or demes within which random mating occurs. In all cases a two allele, one locus model will be assumed.

The spatial structure for this paper is defined as an undirected graph, where each vertex of the graph represents a location, and an edge or link between vertices implies that they participate in the same deme for mating. In addition, each location is assumed to be part of its own deme. An individual exists at each location, and each generation a new individual is created at each location by randomly mating the individuals defined by the deme of this location. Previous work (Dick & Whigham 2005) has shown that for symmetrical spatial structures, where each location has the same deme structure, the effect of spatial structure on fixation time is additive and independent of ploidy. This paper extends this work by examining the forms of spatial structure that produce a faster-than-panmictic speedup in time to fixation (hyperfixation) and will address the following considerations:

1. What spatial configurations produce hyperfixation?
2. For these configurations, does the additive and ploidy independence relationship still hold?
3. How may a configuration be analysed to determined whether hyperfixation occurs?
4. What are the properties of the faster and slowest spatial configurations?

Previous work on spatially-structured populations from the field of population genetics have focussed on the use of spatial structure to slow the lose of variation in a population (Epperson 2003). These approaches have used a variety of spatial structures, but all have considered explicit migration between demes. The work described here

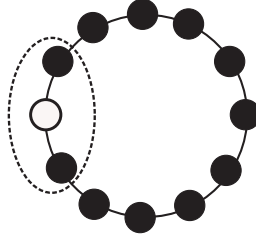


Figure 1: The Ring Spatial Structure for $d = 1$. The deme for the white vertex is outlined.

has implicit migration due to overlapping demes of the network structure used to define the spatial configuration. In general imposing a spatial structure is likely to increase the time to fixation because members of the population become isolated from each other (Crow & Kimura 1970). During the evolution of the population different demes may fixate to different allele values, and therefore the time for genetic drift to propagate a single allele value throughout the population will increase. This has been shown in (Dick & Whigham 2005) for a ring and torus structure.

The remainder of this paper is structured as follows: §2 will give an example for a symmetrical spatial configuration to highlight the properties of the model, §3 will present two spatial configurations where hyperfixation occurs, §4 will develop an index measure for a spatial configuration that predicts when hyperfixation will occur and §5 will discuss the implications of the results and future directions. Finally, §6 will conclude the paper.

2 A Symmetrical Configuration

In this section a simple symmetrical configuration will be described and the properties of time to fixation for one population size and initial allele frequency presented. This will form a baseline for comparison with other spatial configurations presented in subsequent sections of this paper.

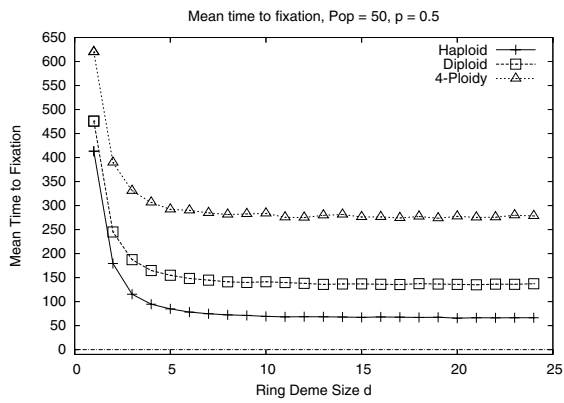
A ring structure of size N , as shown in Figure 1, has a single individual of the population at each vertex of the ring, and a deme (subpopulation) defined by the links to the neighbouring individuals. Each generation a new individual is produced at each vertex by randomly selecting two parents with replacement from the local deme and breeding them to produce a child. This ring structure can be extended by considering the number of steps, d , from a vertex to other vertices of the ring in both directions. The example in Figure 1 has $d = 1$, and the size of each subpopulation is three. For $d = 2$ the subpopulation size increases to five, and so on. Clearly when $d = N/2$ there is no spatial structure (all locations are connected directly to all other locations) and the population is panmictic. Note that the ring defines the complete population structure (there is no migration into or out of the ring).

Breeding of individuals is based on their ploidy. For haploids a single parent from the local deme is randomly selected to replace the individual at a particular location in the next generation. For diploids the rules of Mendelian mating are followed, and for a tetraploid (4-ploidy) two parents are randomly selected, and two randomly chosen genes from each parent used to form the offspring. No mutation occurs during breeding, and there is no selection pressure. The results for a 50 vertex ring structure with various deme sizes is shown in Figure 2. For Figure 2(b) the time to fixation for a panmictic population, based on Eqn. 1, has been subtracted from the fixation time found from the simulation. The average over 10,000 runs for each ploidy level are shown. The key points to note are: the time to fixation is greatest when the deme size is smallest; the values for each time to fixation are independent of ploidy; and the time to fixation due to the spatial configuration is additive.

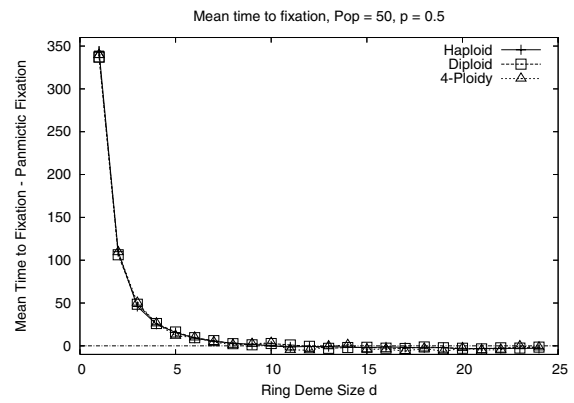
3 Hyperfixation Configurations

This section will explore some general properties of spatial configurations that produce hyperfixation. Since symmetrical configurations produce slower than panmictic fixation times it would be expected that to create configurations with hyperfixation this symmetry needs to be broken. As an introduction, consider the spatial configuration of a generalised line as shown in Figure 3. The initial line structure ($d = 1$) is similar to a ring, however the end points are not connected. For each increase in d an additional link is made between the first vertex (vertex 0), and the d th vertex. Hence in Figure 3 when $d = 5$ there are four additional links back to vertex 0. As d increases this structure has a strong bias towards connections at vertex 0, and therefore the individual at this location has a greater influence over the other individuals on the line.

The resulting time to fixation, averaged over 10000 runs, with increasing connections to vertex 0 is shown in Figure 4. There are several points to note regarding these results: the independence of ploidy does not apply once $d > 25$; as the ploidy level increases the comparative time to fixation decreases; when the line has approximately



(a) Mean time to fixation



(b) Mean time to fixation - panmictic fixation

Figure 2: Mean time to fixation for a generalised ring.

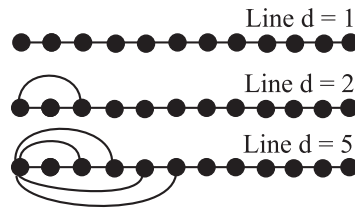
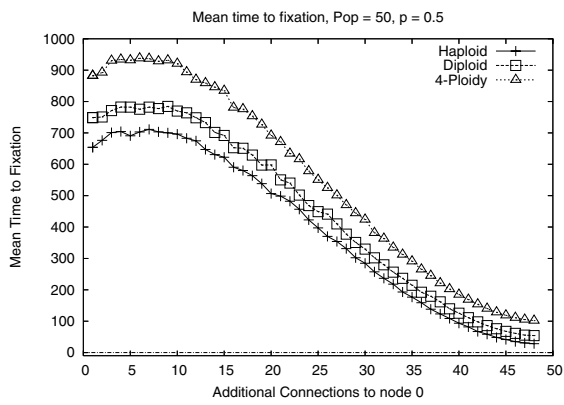
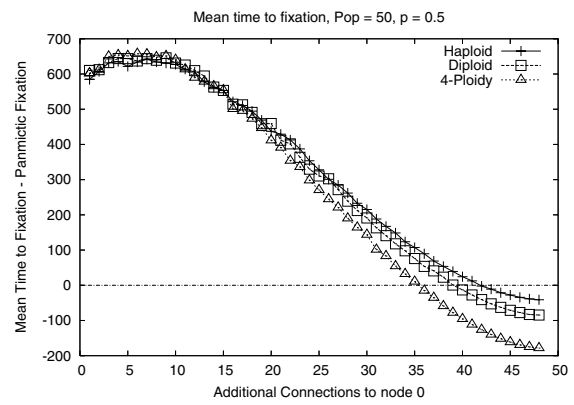


Figure 3: A generalised line structure.



(a) Mean time to fixation



(b) Mean time to fixation - panmictic fixation

Figure 4: Mean time to fixation for a generalised line.

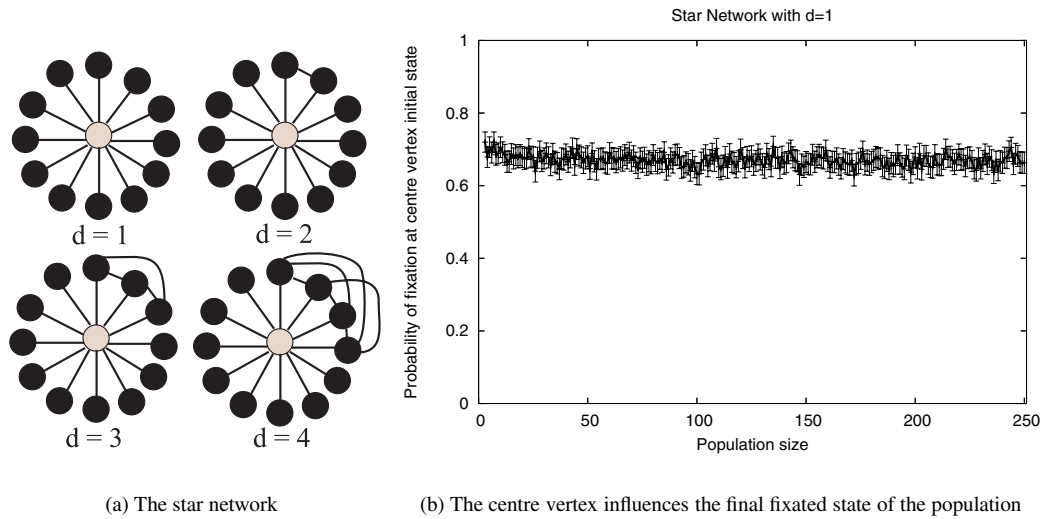


Figure 5: Description and properties of the generalised star network.

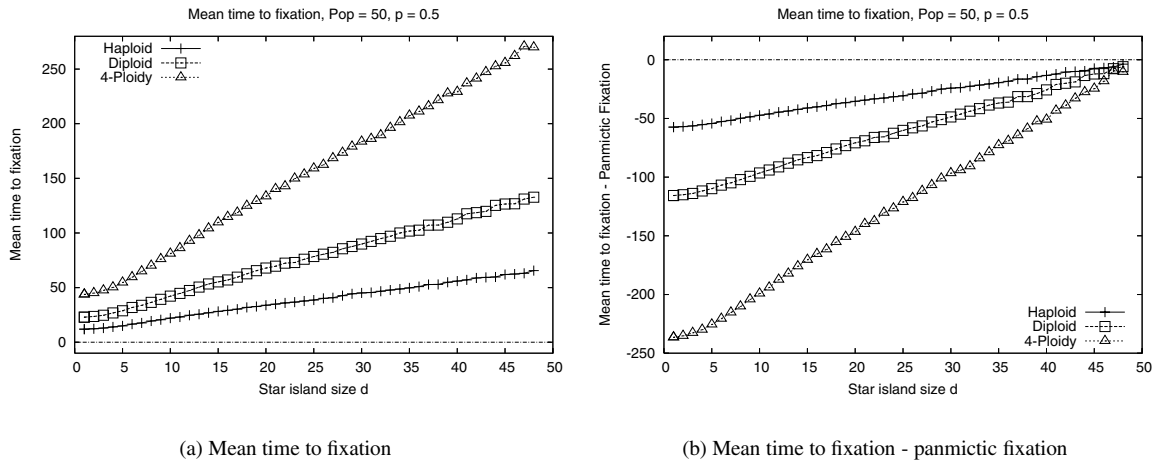


Figure 6: Mean time to fixation for the generalised star.

35 additional links back to vertex 0 hyperfixation commences for the tetraploid; and all ploidy representations hyperfixate after 42 links to vertex 0. This example shows that as a vertex becomes highly connected to other vertices, relative to the overall deme size of the population, the influence of drift on the entire population increases. Hence it is likely that a spatial configuration with a single, dominant vertex over the entire population is likely to have the lowest fixation time.

As a second example, consider the star network shown in Figure 5(a). Here when $d = 1$ the white vertex is connected to all other nodes in the graph, with the outer vertices only connected to each other via this centre vertex. Hence the centre vertex dominates the interactions of all other nodes in the graph. For increasing d the star network is modified so that the outer nodes gradually become fully connected to each other. For example, when $d = 4$ the outer four vertices are fully connected to each other, as well as to the centre vertex. Obviously as d increases the graph approaches a panmictic configuration. The resulting time to fixation, for increasing d , is shown in Figure 6. Clearly the dominance of the centre vertex produces hyperfixation in the configuration for all d values until the network reaches a fully connected state.

The influence of the centre vertex for the star configuration can also be demonstrated by measuring the probability that the final fixated value for the system is the same as the initial value of this vertex. A haploid population was run to fixation for a variety of population sizes, and the initial allele value of the centre vertex

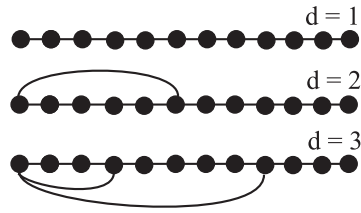


Figure 7: A Small World Line.

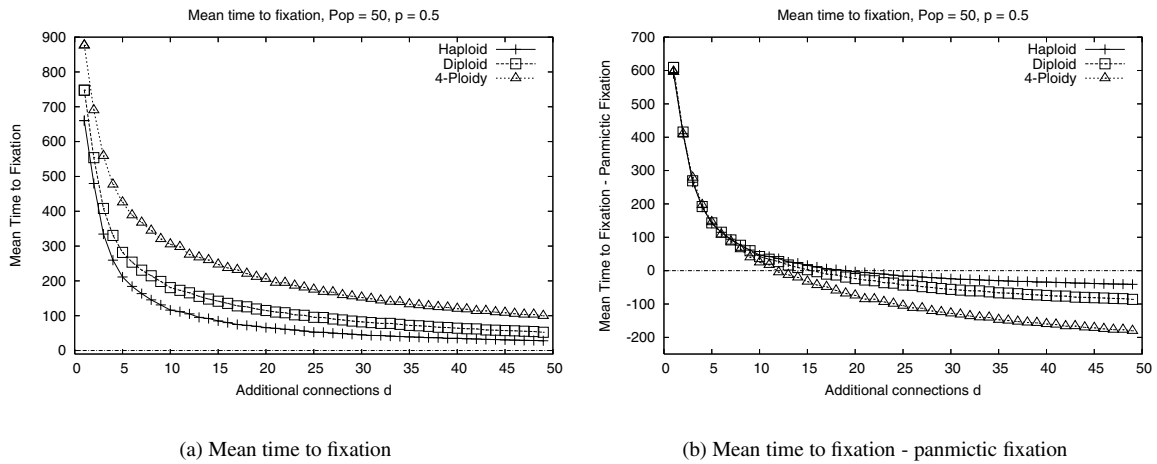


Figure 8: Mean time to fixation for a small world line.

compared with the final fixated value. The average and 95% confidence interval for this model over 1000 replicates is shown in Figure 5(b). Since a node with no influence over the resulting fixation state will have a probability of 0.5, this clearly shows that the final fixation state is influenced by the initial value of the centre vertex. The hypothesis from this experiment is that when a node is more connected than average within a network the resulting fixation of the population will be biased towards the initial value of this node.

The previous examples suggest that hyperfixation occurs when a node dominates the network by being more connected to the network than other nodes. To demonstrate that it is not only the connections, but the way in which the connections relate to the topology of the graph, an example starting with a line can be used to show that the path length from one vertex to another is a key aspect of the fixation rate for a configuration. Figure 7 shows a line for increasing values of d , where each additional link connects vertex 0 half way through the line, one third, one quarter, etc. This produces a network with small world properties, since a single node produces a rapid decrease in the shortest path between any two nodes of the line (via vertex 0). Comparing the results with the generalised line of Figure 4 it is clear that the small world line hyperfixates with fewer additional links. Hence the path length between nodes is a key element in hyperfixation.

4 A Hyperfixation Index

This section will develop an index to determine whether a particular spatial configuration will produce hyperfixation. All examples we have considered are undirected graphs with equal weights for each edge. From §3 it is clear that hyperfixation is related to structures where there is a dissimilarity between connections of the spatial nodes. Since the deme structure for each node determines the subpopulation for mating a natural first measure should use information relating the number of steps from one node to all demes. A node that is closely connected to all demes may influence the outcome of drift and, if this node is more connected to all demes than other nodes, may allow hyperfixation to occur.

In a similar fashion to (Botafogo, Rivlin & Shneiderman 1992) define a matrix D such that the element in the i th row and j th column is the shortest path length from node i to node j , and is represented as $d(i, j)$. For our purposes the shortest path length is the measure of the shortest path from node i to any deme that contains node j . Since our graphs are undirected $d(i, j)$ is a metric, and the matrix D is diagonally symmetrical. We also define

$d(i, i) = 1$ so that the minimum path from one node to all demes in a panmictic population is 1. For example, a simple line with 4 nodes and $d = 1$, such as that in Figure 3, has a D matrix of:

$$\mathbf{D} = \begin{pmatrix} 1 & 1 & 2 & 3 \\ 1 & 1 & 1 & 2 \\ 2 & 1 & 1 & 1 \\ 3 & 2 & 1 & 1 \end{pmatrix}$$

The properties of the hyperfixation index, h , will be designed so that a panmictic network structure has $h = 1$. A network where genetic drift is slower than for a panmictic structure should have $h > 1$, and a hyperfixation network will have $h < 1$.

The average deme path length from node i to all nodes, $\bar{p}(i)$, is defined as:

$$\bar{p}(i) = \frac{1}{n} \sum_{j=1}^n d(i, j) \quad (1)$$

The average deme path length for the entire network, \bar{r} , is defined as:

$$\bar{r} = \frac{1}{n} \sum_{j=1}^n \bar{p}(j) \quad (2)$$

Considering how one node influences the overall fixation rate of a network, two nodes both with the same shortest path need to be distinguished. This is done by considering the node degree, k_i , for node i . In the case of two nodes with the same \bar{p} the node which has a larger degree (i.e. is more connected) will have a greater influence on the rate of fixation. Since h needs to be smaller for more influential nodes, we define the isolation of node i over the network, $I(i)$, as:

$$I(i) = \frac{\bar{p}(i)}{k_i} \quad (3)$$

The minimum $I(i)$ over the entire network now represents the node that has the shortest average path of the network, taking into account the degree of i . The hyperfixation index h_1 , for an n node network, is now defined as:

$$h_1 = \frac{n}{\bar{r}} \times \min_{\forall i} (I(i)) \quad (4)$$

The index h_2 is similar, however the degree of the node is not used within the minimisation. and is only used to select between two paths that are equally short. The comparison between two nodes, $I_2(i, j)$ is defined as:

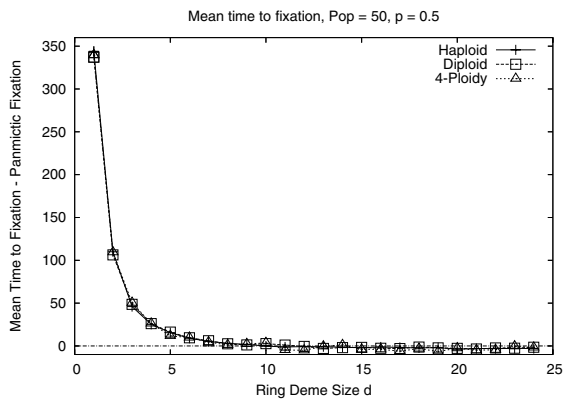
$$I_2(i, j) = \begin{cases} \frac{\bar{p}(i)}{k_i} & \text{if } \bar{p}(i) < \bar{p}(j) \\ \frac{\bar{p}(i)}{k_i} & \text{if } \bar{p}(i) = \bar{p}(j), k_i \geq k_j \\ \frac{\bar{p}(j)}{k_j} & \text{otherwise} \end{cases} \quad (5)$$

The hyperfixation index h_2 , for an n node network, is now defined as:

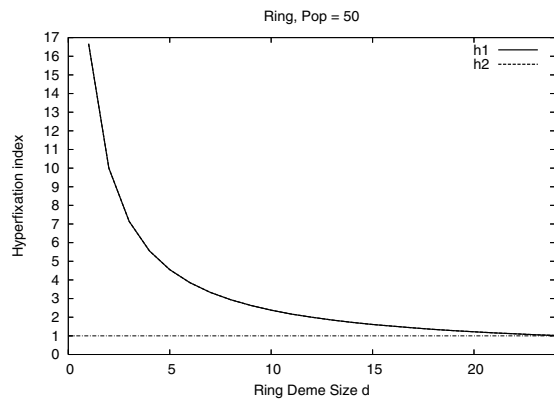
$$h_2 = \frac{n}{\bar{r}} \times \min_{\forall i, j} (I_2(i, j)) \quad (6)$$

4.1 Properties of h

This section will describe the properties of h for the previous networks and describe the limitations in the proposed index. Figure 9 shows h for the ring structure introduced in §2. Note that h approaches a value of 1 as the size of the ring deme size approaches a panmictic structure, and that h_1 and h_2 are identical. Figure 10 shows h for the generalised line structure of §3. Here h is less than zero once the number of additional connections increases beyond 34. This coincides with the 4-Ploidy crossing to a hyperfixation state, however this occurs somewhat later for the smaller ploidy levels. Since h was not designed with an explicit model of ploidy, and time to fixation is dependent on ploidy for hyperfixation structures, h cannot distinguish exactly when hyperfixation will occur. One point to note is that h_2 has very different behaviour from h_1 , and seems to be sensitive to identifying when a network structure produces slower fixation rates. Certainly for Figure 11(b) $h < 0$ for all island sizes until a panmictic structure is created, and so this behaviour is correct. However, the small world line h index of Figure 12(b) is more similar to the haploid behavior (Figure 12(a)), and does not capture the early transition below zero of the 4-ploidy after approximately 14 additional connections. A general conclusion of h is that when $h < 0$ the spatial configuration will hyperfixate for some ploidy level, however when $h > 0$ there is no guarantee that for a given level of ploidy that hyperfixation will not occur. However, when h is large (for these experiments, a value of $h > 5$ is sufficient) hyperfixation will not occur.

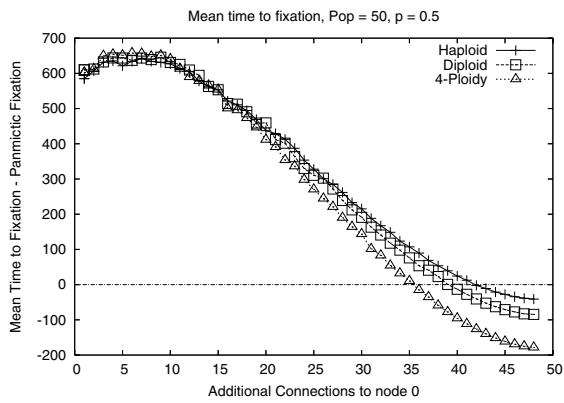


(a) Mean time to fixation - panmictic fixation

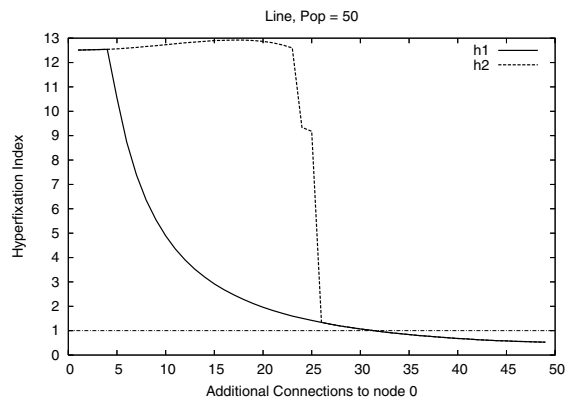


(b) hyperfixation index

Figure 9: Hyperfixation index for a ring.

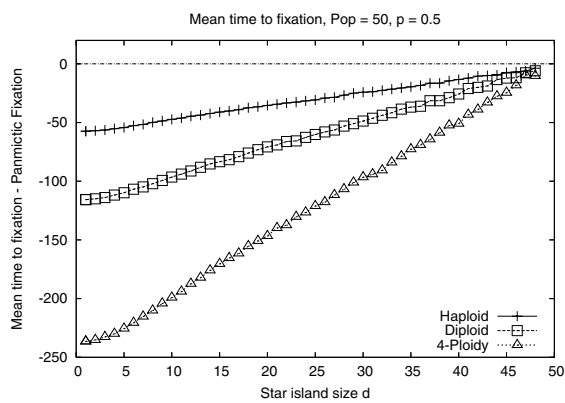


(a) Mean time to fixation - panmictic fixation

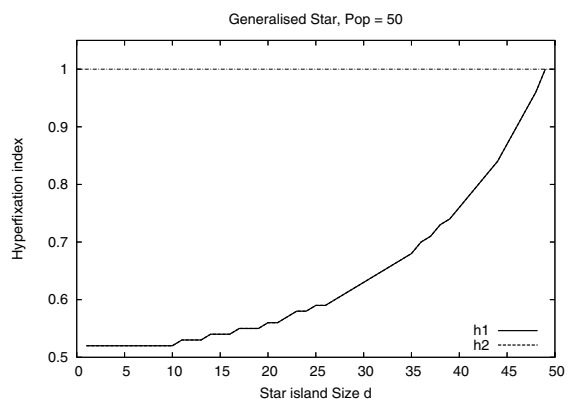


(b) hyperfixation index

Figure 10: Hyperfixation index for a generalised line.



(a) Mean time to fixation - panmictic fixation



(b) hyperfixation index

Figure 11: Hyperfixation index for a generalised star.

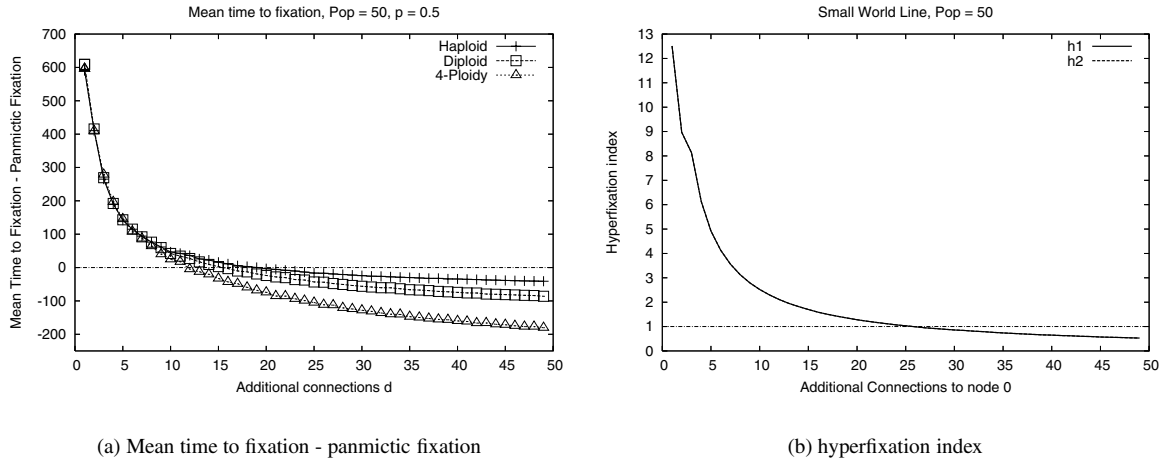


Figure 12: Hyperfixation index for a small world line.

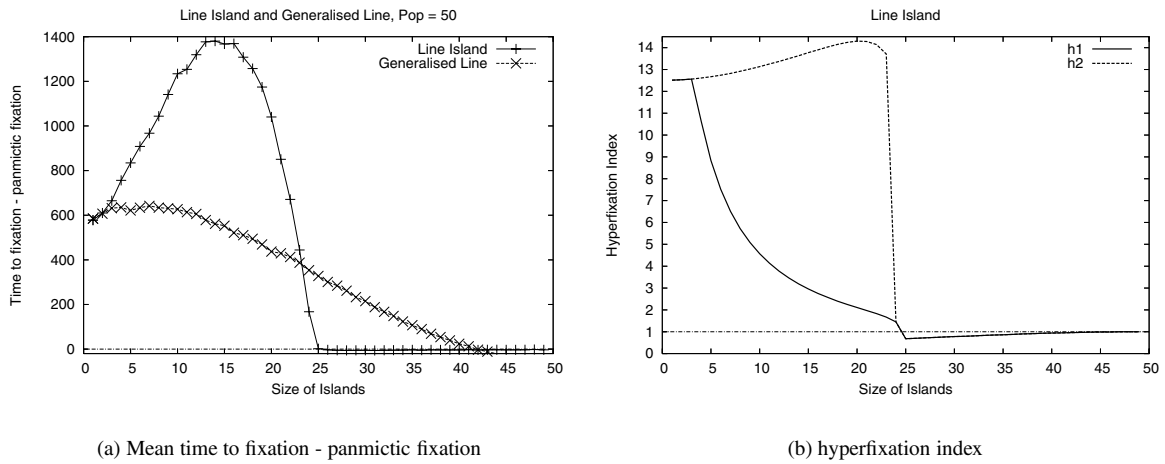


Figure 13: The Line Island: A slow fixation configuration.

4.2 Implications of h for the fastest and slowest configurations

Examining Equation 4 suggests that to produce the fastest fixation network requires a node that minimises isolation in relation to the average path length of the entire network. Considering Equation 3 the minimum value will be achieved by producing the smallest average deme path length $\bar{p}(i)$ and by making the degree of node i as large as possible. For an n node network this minimum is reached when node i is fully connected to all other nodes, so that $\bar{p}(i) = 1$ and therefore $k_i = n$. This implies that node i is connected as shown for the star configuration of Figure 5. In addition, to minimise h we need to make \bar{r} as large as possible. If we start with the star configuration we now want to increase the average path length between all other nodes. However, since these nodes are all connected to node i , any additional links between nodes will decrease the average path length. Therefore based on minimising h it is hypothesised that the fastest hyperfixation structure is the star network.

What would Equation 4 suggest is the structure of the slowest configuration? It might be expected that the larger the value of h the slower will be the fixation rate, however this is shown to be false by comparing h for the ring (Figure 9) and generalised line (Figure 10). Although the ring structure is faster than the generalised line, the relationship of h for each structure is the converse ($h > 16$ for the ring versus $h < 13$ for the line when $d = 1$). Clearly there is not a direct relationship between the magnitude of h and the time to fixation. However, if we assume maximising h may produce the slowest configuration, what form of spatial structure would result? In this case we need to maximise the shortest path length in relation to \bar{r} , and reduce the node degree. Since the minimum node degree occurs when a node is connected to only one other node, this implies that a line-like structure is a

good candidate, where each node has minimum connection to other nodes. However, since h uses the minimum isolation value of $I(i)$ the selected node would be in the middle of the line.

There is a second, conflicting, tradeoff that needs to be considered: minimising \bar{r} . This occurs as the nodes become more connected and hence there are two interacting and conflicting requirements to maximise h . One possible solution is to construct a line and form an island at both ends. This would reduce \bar{r} since many nodes would have a reduced overall deme length, however the islands could both be separated by a line. Hence the minimum node for $I(i)$ would be in the middle of the line, but the average path length would be reduced due to the islands. The behaviour of this type of spatial structure is shown in Figure 13(a), where the time to fixation minus panmictic fixation is shown for haploid individuals. The time to fixation for a generalised line is also shown for comparison. Here it is clear that this form of structure is significantly slower than the generalised line when the islands at each end of the line contain approximately one-third of the total number of nodes in the network. The hyperfixation index for this structure is shown in Figure 13(b). The value of h_1 always decreases with increasing island size, although the time to fixation increases until the island size is ≈ 15 . Hence although the discussion on maximising h suggested that this form of structure may be slow, the value of h_1 does not follow the overall behaviour of drift for this network. However, h_2 once again shows a sensitivity to structures that slow fixation, since in Figure 13(b) h_2 increases until the island size is approximately 22, which coincides with the approximate peak for fixation time. Note also that $h < 1$ once the island size ≥ 25 that corresponds to the line island entering hyperfixation (although not clearly seen in Figure 13(a) the fixation rate is just below zero).

5 Discussion and Future Directions

The concept of hyperfixation for network structures has not been previously examined in the literature, mainly because this form of spatial representation is not commonly used in population genetics. This paper has shown that many spatial structures result in this behaviour, and that this occurs in general when one or more nodes have a greater than average influence (in terms of connectivity) over the network.

The hyperfixation index h that has been introduced appears to capture the basic properties of a network, although it is not guaranteed to detect exactly when hyperfixation will occur. There are several possible reasons for this weakness:

- h takes into account only the most influential node in the network, and does not consider the distribution of all nodes. Although a comparison against the average shortest path \bar{r} takes into account the general network properties it may well be that the distribution of the nodes may better characterise hyperfixation;
- h does not have an explicit model of ploidy, nor does it model the behaviour of genetic drift. Although this is an attractive property of h , some consideration of how drift behaves within a network could be considered as part of the index measure;
- The distance matrix $d(i, j)$ is currently a linear measure between demes, although it may well be the case that a different step analysis (such as distance increasing as the square or as the logarithm of the number of steps) may be more suitable.

Other future work includes a formal analysis of the properties of h_1 and h_2 , and a detailed examination of the network properties at the point when the independence of ploidy property is disrupted. Since this happens well before hyperfixation some other characteristic of the network needs to be formulated, and understanding this property may well help to produce a better version of h . In addition, this work has considered genetic drift without selection. How the previous spatial configurations behave under a variety of selection pressures should also allow a more thorough understanding of the relationship between structure and behaviour for these simple genetic models.

6 Conclusion

This paper has introduced the notion of hyperfixation for spatial configurations described as undirected networks with genetic drift. A number of example structures that produce hyperfixation have been analysed by simulation, and an index to estimate when hyperfixation will occur has been developed. The resulting index has been used to suggest the characteristics of the fastest and slowest fixation structures, and has been shown to have properties that approximate configurations that lead to hyperfixation for some ploidy level.

This work is an introduction to exploring the relationship between time to fixation for genetic drift and spatial configurations. There are many directions for future work in this area that complement the current work in geographical population genetics.

References

- Botafogo, R., Rivlin, E. & Shneiderman, B. (1992). "Structural analysis of hypertexts: identifying hierarchies and useful metrics" *ACM transactions on Information Systems*. **10**: 142–180.
- Crow, J. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper and Row, New York, Evanston and London.
- 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.
- Epperson, B. (2003). *Geographical Genetics*. Princeton, New Jersey.
- Kimura, M. & Ohta, T. (1969). "The average number of generations until fixation of a mutant gene in a finite population" *Genetics*. pp. 763–771.