Spatial Soft Sweeps in Structured Populations

Mia Zavala Sanborn

May 2026

School of Computer Science
Carnegie Mellon University
Pittsburgh PA 15213

COMMITTEE
Anush Devadhasan
Oana Carja

Submitted in partial fulfillment of the requirements
for the Senior Honors Thesis

Keywords: evolutionary dynamics, population genetics, soft selective sweeps, graph-structured populations

Abstract

Soft selective sweeps, in which multiple adaptive alleles rise to fixation simultaneously within a population, are increasingly recognized as a primary mode of adaptation across many species. Despite growing interest, the influence of population structure on the probability that a soft sweep occurs remains poorly understood, with most theoretical work restricted to well-mixed or simple lattice models. In this thesis, we systematically investigate how the topology of graph-structured populations governs the likelihood of soft sweeps. Using a mutation–migration dynamical model adapted from Paulose et al. (2019), we simulate soft sweeps across a diverse set of graph families (random regular, Watts–Strogatz, random geometric, power-law cluster, bottleneck, preferential attachment, and core–periphery networks), each comprising 1000 demes. Our results reveal that average degree and algebraic connectivity are both strong predictors of soft sweep probability, with lower connectivity and lower degree consistently promoting greater allelic diversity at fixation. However, neither metric alone fully explains variation across graph families: graphs with greater spatial organization, such as Watts–Strogatz and random geometric networks, sustain higher soft sweep probabilities than random regular graphs of equivalent degree. We further explore core–periphery structures, demonstrating that the interplay between a high-degree core and low-degree peripheral clusters produces a characteristic non-monotonic "dip" in soft sweep probability as core size increases. We propose an approximate decomposition of the total soft sweep probability into core, periphery, and cross-cluster contributions, offering a framework for predicting diversity outcomes in heterogeneous population structures. These findings carry implications for understanding the emergence of drug resistance, the maintenance of genetic diversity in endangered species, and the design of intervention strategies in epidemiological contexts.

1 Introduction

While much of previous research on adaptation by natural selection has focused on single mutants fixating and sweeping through a population, in practice this phenomenon, known as a hard selective sweep, is rare. A growing body of evidence suggests that soft selective sweeps, in which multiple independently arising adaptive alleles coexist and rise to high frequency simultaneously, are instead the predominant mode of adaptation in many species [8]. Unlike hard sweeps, which erase genetic diversity around the selected locus, soft sweeps preserve a signature of standing genetic variation and leave behind a more diverse genomic landscape. This distinction is not merely theoretical: genomic data from humans [13] and other organisms provide direct support for the prevalence of soft sweeps when populations must respond rapidly to environmental pressures by drawing upon pre-existing genetic variation.

A particularly striking example of a soft sweep is lactose persistence, the genetic adaptation that allows the continued production of lactase into adulthood. Although lactose persistence was originally thought to have a single European origin, subsequent research by Tishkoff et al. [14] identified the same trait in African populations, arising from entirely different genetic mutations. In fact, at least four distinct single-nucleotide polymorphisms (SNPs) have been found to code for the same phenotype across the global human population, illustrating the hallmark signature of a soft sweep. Other notable examples include the fruit fly's adaptation to pesticide exposure [4] and HIV's rapid evolution of resistance to antiviral drugs [2].

Understanding soft sweeps is important for two complementary reasons. First, it is essential for correctly interpreting genomic data and reconstructing evolutionary lineages: misidentifying a soft sweep as a hard sweep can lead to erroneous conclusions about the demographic and selective history of a population. Second, understanding the mechanisms that promote or suppress soft sweeps has direct practical consequences. In the context of drug resistance, for example, it would be valuable to predict when and how diverse resistance strains will emerge, since the development of multiple resistance alleles can render single-drug therapies ineffective and necessitate costly second-line treatments. Conversely, in conservation biology, understanding these dynamics can help maintain the genetic diversity that is crucial for the long-term fitness and reproductive viability of endangered populations.

We hypothesize that a crucial parameter governing the degree to which a soft sweep occurs within a population is the population's spatial structure. Consider the well-studied example of marine stickleback fish, which have repeatedly and independently adapted from saltwater to freshwater environments across many geographically isolated lakes [1]. This classic case of parallel evolution raises a natural question: would this soft sweep pattern still be observed if all the fish had been placed in a single, well-mixed lake? More broadly, how might changes in the spatial structure of human populations, or the populations of their pathogens, affect genetic diversity and the nature of adaptive responses? Answering such questions requires a clear and quantitative understanding of how spatial structure impacts soft selective sweeps.

Furthermore, in many cases where it would be desirable to manipulate the outcome of a soft sweep in response to an environmental pressure, population structure offers a potential avenue of control. Consider, for example, a tuberculosis outbreak in a hospital setting. How should patients be separated and quarantined to minimize the probability that multiple

strains of drug-resistant tuberculosis arise and reinfect the broader hospital population? To answer this question, one needs a precise understanding of how the connectivity and organization of a population influence the ability of diverse mutations to arise and reach fixation.

In this paper, we represent population structure using graph networks, simulate soft sweeps across a large and diverse set of graph topologies, and distill from the results an understanding of how key structural properties, including average degree, algebraic connectivity, and core–periphery organization, impact the probability of soft sweeps.

2 Background and Prior Work

The concept of soft selective sweeps has its roots in the early population genetics literature, but the phenomenon was first explicitly characterized in the context of modern genomics by Sabeti et al. [12] in 2002. In that study, Sabeti and colleagues used population-genetic methods to identify events in which multiple mutations responded to the same environmental pressure, focusing specifically on two different gene variants conferring adaptation to malaria. Importantly, Sabeti defined the probability of a soft sweep as the probability that two randomly sampled individuals carrying a particular adaptation harbor different underlying genetic mutations.

The computational study of soft sweeps began in earnest with the seminal work of Hermisson and Pennings in 2005, who coined the term "soft sweep" and formally delineated the mechanisms by which they arise [5]. Hermisson and Pennings argued that rapid genetic adaptation is frequently the result of selection acting on genetic variation already present in the population prior to the onset of selective pressure, and that multiple copies of the adaptive allele are often present, leading to a soft sweep rather than a hard one. They supported this claim using a Wright–Fisher model in a well-mixed (panmictic) setting, deriving the probability that more than one allelic copy reaches fixation in response to a selective pressure and validating these predictions with simulations.

The first effort to incorporate population structure beyond the well-mixed assumption was undertaken by Ralph and Coop in 2010 [11]. In their model, well-mixed demes were arranged on a two-dimensional grid, and the probability of a soft sweep was expressed in terms of a single characteristic length scale reflecting the expected distance a spreading allele travels before encountering a different spreading allele. This spatial framework provided the first analytical link between geographic structure and allelic diversity during a sweep.

Building on this foundation, Paulose et al. in 2019 [9] extended the grid model to include long-range dispersal events ("jumps"), analyzing the effect of an initial mutation's expansion from its local core structure to a broader halo structure formed by long-range connections. Their work examined the impact of long-range dispersal on both local and global allelic diversity and, importantly, introduced the modeling framework that we adopt in the present study: demes are occupied by a single allele type at any given time, and once a mutation has fixed within a deme, no further displacement by new mutations is permitted.

Although these studies have advanced the mathematical and computational study of soft sweeps beyond the panmictic setting, the population structures considered have remained limited to regular, symmetric topologies such as grids and lattices. Real biological popu-

lations, however, are characterized by irregular connectivity, heterogeneous degree distributions, clustering, bottlenecks, and hierarchical organization. To bridge this gap by studying soft sweeps on general, heterogeneous graph structures is the central goal of the present work.

In parallel, a substantial body of research has explored evolutionary dynamics on graphs more broadly. Lieberman et al. (2005) established foundational results on how graph topology affects the fixation probability and fixation time of single mutants. Subsequent work has shown that increased degree can decrease time to fixation, and that degree-heterogeneous graphs produce qualitatively different evolutionary outcomes compared to regular networks. More recently, Kuo et al. (2025) [7] examined how population structure influences the maintenance of clonal diversity, focusing primarily on hard sweep dynamics and the time to fixation of a single mutant.

Our work extends this line of inquiry by shifting the focus from single-mutant fixation to the multi-allele dynamics that characterize soft sweeps, and by seeking general principles that relate graph-theoretic properties to the probability of allelic diversity at fixation.

3 Methodology

3.1 Dynamical Model

To examine how population structure impacts the probability of a soft sweep, we simulated mutation and migration events on graph networks using a dynamical system adapted from Paulose et al. (2019) [9]. In our adaptation, we removed the ability of migration events to occur across large distances of the population, limiting migration to occur only between a mutated node and its immediate graph neighbors.

The dynamical system proceeds as follows. Each node (deme) in the graph is occupied by a single clone, which is either of wild type (wt) or of mutant type (mt). At time $t = 0$, all clones are wild type. Each mutant clone additionally carries an allelic identity; all allelic identities confer equivalent fitness and differ only in their labels (represented by distinct colors or numbers).

At each time step, a single clone is randomly selected from the network. This selection is weighted by the parameter $\tilde{\mu} = u/m$ (the ratio of mutation rate to migration rate). All mutant-type nodes receive a selection weight of 1, while all wild-type nodes receive a weight of $\tilde{\mu}$. The probability that a given wild-type node is selected is therefore

$$P(\text{wt selected}) = \frac{\tilde{\mu}}{\tilde{\mu} \cdot N_{\text{wt}} + 1 \cdot N_{\text{mt}}},$$

where N_{wt} and N_{mt} denote the number of wild-type and mutant-type nodes, respectively.

If a mutant-type clone is selected, one of its graph edges is chosen uniformly at random. If the neighboring node at the other end of this edge is wild type, it is converted to a mutant of the same allelic identity as the source node. If the neighbor is already a mutant (of any identity), no change occurs. If a wild-type clone is selected, it is replaced by a mutant of a brand-new allelic identity not previously present in the population.

This dynamical model makes several key simplifying assumptions:

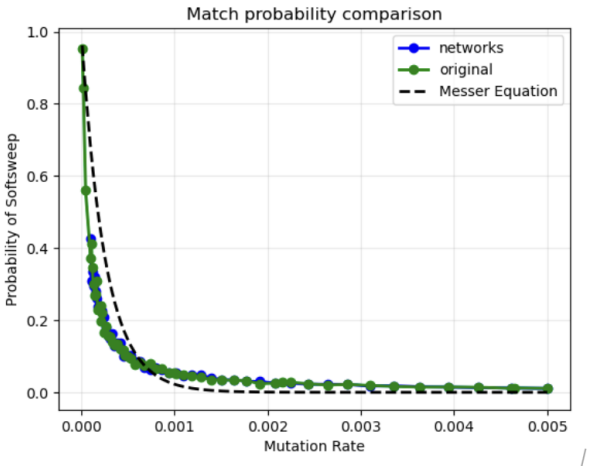
1. **Strong selection:** Mutant types automatically displace wild types upon contact, reflecting an environment in which the adaptive allele confers a large fitness advantage.

2. **Equal fitness among mutants:** All mutant alleles are equally fit, so no mutant type can displace another. This assumption isolates the effect of population structure from the complicating effects of fitness differences among adaptive alleles.

3. **No secondary mutations:** Once a mutant type has fixed within a deme, that deme permanently retains its allelic identity. This simplification ensures that diversity at fixation reflects only the initial stochastic dynamics of mutation and migration.

3.2 Model Validation

To verify the correctness of our implementation, we compared the output of our model on a well-mixed (complete) graph to the results produced by the original code of Paulose et al. (2019). As expected, the probability of a soft sweep in our implementation exactly matches the Paulose et al. results when the population structure is well-mixed.



We further compared our results with the analytical predictions of Pennings and Hermisson (2006) [10], which express the probability of a soft sweep in terms of separate mutation and migration rates rather than the composite parameter $\tilde{\mu}$. At parameter values where the ratio of migration to mutation rate equals the corresponding $\tilde{\mu}$, the results follow a similar curve. Although the two are not identical, owing to the additional assumptions in the Paulose model that are absent from the analytical expression, the agreement in trend confirms that the simplified parameterization captures the essential behavior.

3.3 Simulation Parameters

Because the effects of population size and mutation–migration rate on soft sweep probability have been extensively characterized in prior work, we held these parameters constant throughout our study to isolate the effect of population structure. The composite parameter was fixed at $\tilde{\mu} = 0.000464$, corresponding to a regime in which migration events are approximately 2155 times more likely than mutation events. The population size was fixed at

$N = 1000$ demes, where each deme is assumed to have an internal population large enough to preclude the effects of genetic drift within demes.

Simulations were implemented in C++ adapted from the Paulose et al. original codebase. Each graph configuration was subjected to 100,000 independent simulation runs with distinct random seeds, and results were averaged across runs for each unique graph. The primary quantity of interest is the probability of a soft sweep, defined as the probability that more than one allelic identity is present among the mutant-type nodes at the time when all wild-type nodes have been eliminated:

$$P(\text{soft sweep}) = 1 - \sum_i \left(\frac{n_i}{N}\right)^2,$$

where n_i is the number of demes occupied by allelic identity i at the time of full mutant fixation.

3.4 Graph Families

We studied the following families of graphs, each providing a different lens through which to examine the effect of population structure on soft sweep dynamics.

Complete (well-mixed) graphs. In a complete graph, every pair of nodes is connected by an edge, and no spatial structure exists. These graphs serve as the baseline against which all other topologies are compared. Complete graphs were generated using `nx.complete_graph`.

Random regular graphs. In a regular graph, every node has the same degree k. Random regular graphs are constructed by assigning edges randomly subject to this degree constraint, resulting in graphs with no preferential connectivity or spatial organization. These graphs were generated using `nx.random_regular_graph`, with degrees ranging from 2 to 100. Any disconnected graphs were discarded and regenerated until a connected graph was obtained. For core–periphery experiments, we used degrees of 4, 6, 10, and 20.

Watts–Strogatz graphs. Watts–Strogatz graphs are deterministically constructed regular graphs with explicit spatial organization. Starting from a ring of nodes, each node is connected to its k nearest neighbors, producing a regular lattice with high clustering (transitivity). These graphs were generated using `nx.watts_strogatz_graph` with the rewiring probability set to zero, so that the ring structure was preserved without random edge swaps. Degrees ranged from 2 to 999 (the latter producing a well-mixed graph). For core–periphery experiments, we used average degrees of 4, 6, 10, and 20.

Random geometric graphs. Random geometric graphs are constructed by distributing nodes uniformly within a unit square and connecting all pairs of nodes whose Euclidean distance is at most a specified radius r. These graphs possess explicit spatial embedding and naturally varying degree distributions, though they retain approximately regular properties when the radius is not too small. Graphs were generated using `nx.random_geometric_graph` with radii ranging from 0.05 to 1.0 (the latter yielding a well-mixed graph). Disconnected graphs were discarded and regenerated.

Power-law cluster graphs. Power-law cluster graphs simulate the clustering behavior commonly observed in social networks, where individuals tend to form tightly connected local groups. These graphs are generated using `nx.powerlaw_cluster_graph` with $N = 1000$ nodes and an initial attachment parameter of $m = 10$ edges per new node. A clustering parameter p controls the probability that, after a new edge is added, an additional edge is introduced to close a triangle. By varying p from 0 to 1, we tuned the degree of local clustering while maintaining the power-law degree distribution. For each value of p, 10 distinct graphs were generated, and disconnected graphs were discarded and regenerated.

Bottleneck graphs. Bottleneck networks represent populations composed of tightly connected clusters joined by only a small number of inter-cluster edges. To construct these graphs, we created four well-mixed clusters of 250 nodes each. At each iteration, we selected three nodes from within the clusters and removed their internal edges, then added edges connecting nodes across clusters. This procedure ensured that the average degree remained approximately constant while the number of inter-cluster connections (and hence global connectivity) increased with each iteration.

Island (core–periphery) graphs. Island graphs were constructed by partitioning the $N = 1000$ nodes into a central core cluster of C nodes and 10 peripheral clusters, each containing P nodes, such that $C + 10P = 1000$ (with cluster sizes differing by at most one node). Between the core cluster and each peripheral cluster, either 1, 5, or 10 connection edges were added. We constructed three variants of island graphs:

- *Well-mixed islands:* All clusters are internally well-mixed (complete), and connection nodes are chosen arbitrarily.
- *Regular islands:* The core cluster is a 20-regular random graph, peripheral clusters are 6-regular random graphs, and connection nodes are chosen randomly while avoiding self-loops.
- *Watts–Strogatz islands:* The core cluster is a Watts–Strogatz graph with each node connected to its 20 nearest neighbors, and the peripheral clusters are Watts–Strogatz graphs with each node connected to its 6 nearest neighbors. Connection edges for each peripheral cluster are distributed evenly around the core ring to preserve spatial organization.

3.5 Graph Metrics

To quantify the structural properties of each graph, we relied on two principal metrics.

Average degree. The average degree of a graph is defined as

$$\langle k \rangle = \frac{2|E|}{N},$$

where $|E|$ is the number of edges and N is the number of nodes. This metric provides a basic measure of how densely connected the population is.

Algebraic connectivity. The algebraic connectivity of a graph, first introduced by Fiedler (1973) [3], is defined as the second-smallest eigenvalue of the graph's Laplacian matrix (or, in our case, the normalized Laplacian). This quantity measures how easily the graph can be partitioned into large, weakly connected components: a graph with high algebraic connectivity is tightly integrated and resistant to disconnection, while a graph with low algebraic connectivity possesses bottlenecks or loosely connected substructures. We computed algebraic connectivity using `nx.algebraic_connectivity(G, method="lanczos", normalized=True)`.

4 Results and Discussion

4.1 Relationship Between Average Degree and Probability of Soft Sweep

Initial results across many graph types revealed a strong inverse relationship between the probability of a soft sweep and the average degree of the graph. To isolate and characterize this relationship, we focused on three families of graphs for which the average degree can be systematically varied.

4.1.1 Random Regular Graphs

Random regular graphs are the natural starting point for this analysis because they are the most structurally similar to well-mixed graphs: no node is preferentially connected to any other, and there is no influence of assortativity, clustering, or spatial organization. The key distinction from well-mixed graphs is that not all nodes are connected to all others, so the average degree can be tuned independently of the population size.

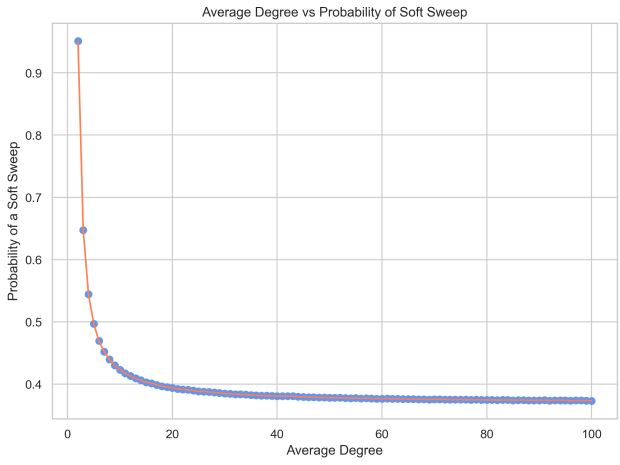


Figure 1: Probability of soft sweep as a function of average degree for 1000-node random regular graphs. For each value of k (ranging from 2 to 100), 10 random graphs were generated and simulated with $\tilde{\mu} = 0.000464$ for 100,000 runs each. Results are averaged across graphs of the same degree.

Our results (Figure 1) show a clear negative correlation between average degree and the probability of a soft sweep. The 2-regular graphs exhibit the highest soft sweep probability at approximately 95%, while the 100-regular graphs converge to approximately 37%, matching the well-mixed baseline. The relationship follows an exponential decay profile, with the soft sweep probability declining steeply over the first 50 units of degree increase before asymptotically approaching the well-mixed limit.

4.1.2 Watts–Strogatz Graphs

Watts–Strogatz graphs share the regularity property of random regular graphs (every node has the same degree) but differ in that their edges are arranged to reflect spatial proximity along a ring. As a consequence, connected nodes are likely to share neighbors, producing elevated transitivity (clustering coefficient) relative to random regular graphs of the same degree. This allows us to test whether the degree–soft sweep relationship observed in random regular graphs persists in the presence of increased spatial organization, and to begin isolating the effect of clustering on allelic diversity.

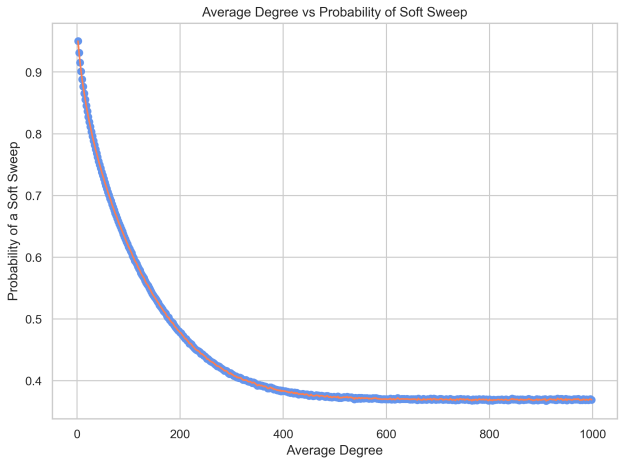


Figure 2: Probability of soft sweep as a function of average degree for 1000-node Watts–Strogatz graphs. Degrees range from 2 to 999 (well-mixed). Simulations were run with $\tilde{\mu} = 0.000464$ for 100,000 runs per graph.

The results (Figure 2) confirm a negative correlation between average degree and the probability of a soft sweep, consistent with the random regular graph results. However, the decay is notably shallower: the soft sweep probability decreases more gradually with increasing degree and approaches the well-mixed limit of 37% much more slowly than for random regular graphs. This observation suggests that the spatial organization inherent in Watts–Strogatz graphs, specifically their elevated transitivity, acts to sustain allelic diversity even as average degree increases, potentially by slowing the spatial spread of any single allele and thereby providing more time for independent mutations to arise and establish.

4.1.3 Random Geometric Graphs

Random geometric graphs offer a more realistic, position-based model of population structure. Because nodes are embedded in a physical space and connected based on proximity, these graphs possess explicit spatial organization while also exhibiting some degree heterogeneity (unlike the strictly regular graphs above). Although the degree distribution of random geometric graphs is relatively narrow when averaged over many realizations, the spatial embedding introduces correlations between node positions and connectivity that are absent in random regular graphs.

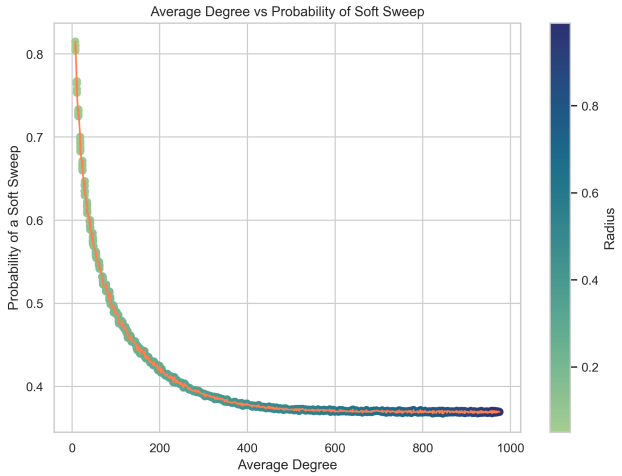


Figure 3: Probability of soft sweep as a function of average degree for 1000-node random geometric graphs. The connection radius ranges from 0.05 to 1.0 (well-mixed). For each radius, 10 graphs were generated and simulated with $\tilde{\mu} = 0.000464$ for 100,000 runs each. Results are averaged across graphs of the same radius.

The results (Figure 3) confirm the negative exponential-decay relationship between average degree and soft sweep probability. The smallest-radius graphs ($r = 0.05$), with an average degree of approximately 7, exhibit a soft sweep probability of about 80%. For comparison, 7-regular random graphs yield a soft sweep probability of approximately 45%, a substantially lower value. This discrepancy reinforces the conclusion that spatial organization, in this case the geometric embedding, promotes allelic diversity beyond what would be predicted by degree alone.

These results carry intuitive implications for real populations. In the context of infectious disease, for example, they suggest that populations in less densely connected environments, where the average number of contacts per individual is low, are more likely to develop multiple independent resistant strains in response to selective pressure. This is consistent with epidemiological observations that geographically fragmented populations often harbor greater pathogen diversity.

4.1.4 Summary Across Graph Families

Figure 4 plots the soft sweep probability against average degree for all graph families on a single set of axes. Within each family, the exponential decay relationship is clearly evident.

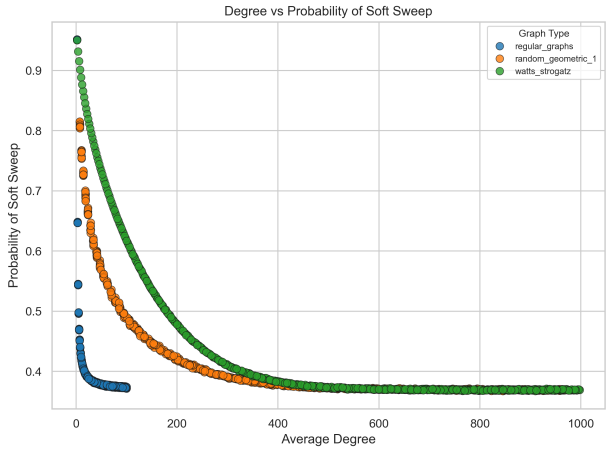


Figure 4: Probability of soft sweep as a function of average degree across all graph families studied. Each point represents the average across multiple graph realizations within a family. The negative trend is consistent within each family, but substantial variation exists between families at the same degree.

However, across families, average degree alone does not fully determine the probability of a soft sweep. For instance, a random geometric graph with a given average degree consistently exhibits a higher soft sweep probability than a random regular graph with the same average degree. This demonstrates that while degree is a strong first-order predictor of soft sweep probability, additional structural features (such as clustering, spatial organization, and connectivity) play an important and independent role.

4.2 Relationship Between Algebraic Connectivity and Probability of Soft Sweep

The observation that average degree alone is insufficient to predict soft sweep probability across graph families motivated us to investigate a second structural metric: algebraic connectivity. Algebraic connectivity, the second-smallest eigenvalue of the normalized graph Laplacian, measures how resistant a graph is to being partitioned into weakly connected components. A graph with high algebraic connectivity is tightly integrated, while a graph with low algebraic connectivity contains bottlenecks or loosely coupled substructures.

4.2.1 Regular Graphs with Tuned Transitivity

To isolate the effect of connectivity from that of degree, we began with 10-regular random graphs whose edge structure was then modified using the triangle-swapping algorithm of Kuo et al. (2024) [6]. This algorithm adjusts the fraction of closed triangles (transitivity) within the graph while preserving both the degree distribution and the regularity of the network. By increasing the fraction of triangles, a greater proportion of each node's connections are allocated to local neighbors rather than distant parts of the graph, which decreases the algebraic connectivity while holding the average degree fixed.

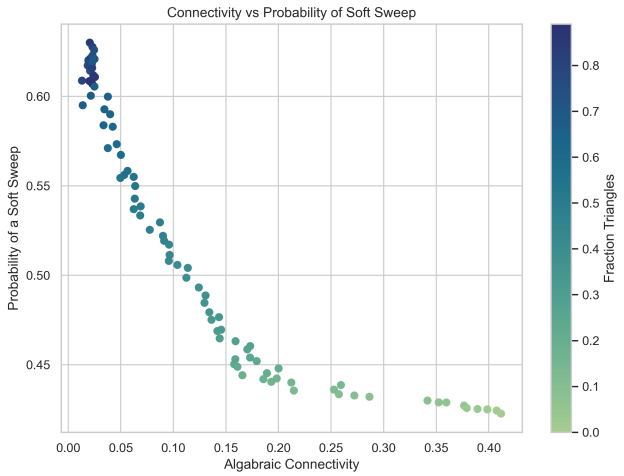


Figure 5: Probability of soft sweep as a function of algebraic connectivity for 1000-node 10-regular graphs with tuned transitivity. The fraction of triangles ranges from 0.01 to 0.89. Simulations were run with $\tilde{\mu} = 0.000464$ for 100,000 runs per graph.

The results (Figure 5) reveal a clear negative relationship between algebraic connectivity and the probability of a soft sweep. The most randomly wired graphs (triangle fraction 0.01, algebraic connectivity ≈ 0.41) exhibit the lowest soft sweep probability, while the most clustered graphs (triangle fraction 0.89, algebraic connectivity ≈ 0.02) exhibit the highest, reaching approximately 0.61. The relationship follows an exponential decay profile similar to that observed for degree, confirming that connectivity, independent of degree, is a powerful predictor of allelic diversity at fixation.

These results are consistent with the findings of Kuo et al. (2025) [7], who demonstrated that increased transitivity accelerates the fixation of new mutants. We hypothesize that the mechanism linking low connectivity to high soft sweep probability operates through fixation speed: in tightly connected graphs, a single mutant allele can rapidly sweep through the entire population before alternative alleles have time to arise and establish, thereby driving the outcome toward a hard sweep. In loosely connected graphs, by contrast, the slower rate of allele spread provides a wider temporal window for independent mutations to arise in distant parts of the network, promoting allelic diversity.

4.2.2 Bottleneck Graphs

To further probe the connectivity–diversity relationship, we examined bottleneck graphs composed of four well-mixed clusters with a variable number of inter-cluster connections. As inter-cluster connections are added (and an equal number of intra-cluster edges are removed to hold degree constant), the algebraic connectivity increases while the degree distribution remains fixed.

The bottleneck graphs (Figure 6) display the same pattern: soft sweep probability decreases with increasing algebraic connectivity, even within the relatively narrow range of connectivity values accessible to this graph family. This result confirms that the connectivity–diversity relationship is not an artifact of the particular graph construction method used for

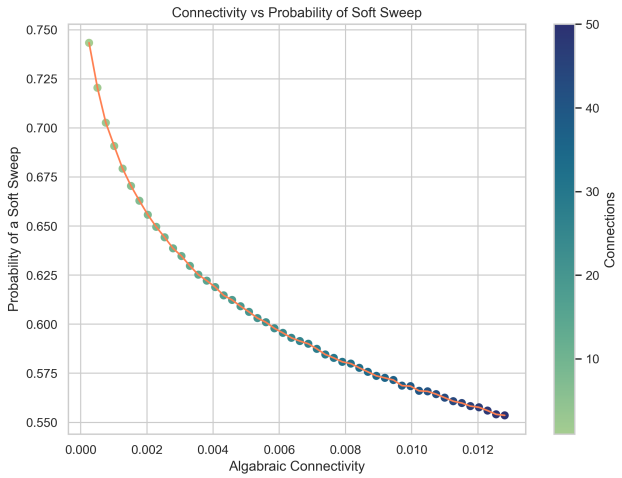


Figure 6: Probability of soft sweep as a function of algebraic connectivity for 1000-node bottleneck graphs composed of four well-mixed clusters with increasing inter-cluster connectivity. Average degree is held approximately constant. Simulations were run with $\tilde{\mu} = 0.000464$ for 100,000 runs per graph.

the regular-graph experiments.

4.2.3 Power-Law Cluster Graphs

Finally, we examined the connectivity–diversity relationship in power-law cluster graphs, which represent a more realistic and heterogeneous population structure. By varying the clustering parameter p from 0 to 1, we generated networks spanning a range of algebraic connectivities while preserving the power-law degree distribution characteristic of many real-world social and biological networks.

The same inverse relationship between algebraic connectivity and soft sweep probability is observed (Figure 7), indicating that our findings generalize beyond regular and symmetric graph structures to the scale-free networks that are common in biology and epidemiology.

When all graph families are re-plotted with algebraic connectivity on the horizontal axis (rather than degree), the data points collapse more tightly, suggesting that algebraic connectivity is a better single predictor of soft sweep probability than average degree. Nevertheless, residual variation between graph families persists, indicating that no single scalar metric fully captures the effect of population structure on allelic diversity.

4.3 Soft Sweep Probability in Core–Periphery Structures

Having established that both average degree and algebraic connectivity influence soft sweep probability in relatively homogeneous graphs, we turned to more complex, heterogeneous topologies. Core–periphery structures, in which a densely connected core cluster is surrounded by multiple sparsely connected peripheral clusters, are of particular interest because they are a common feature of real population architectures, from hospital wards connected by shared staff to metropolitan areas linked by transportation networks.

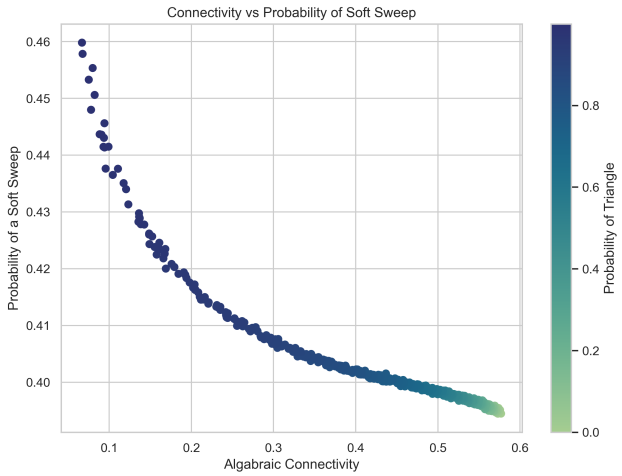


Figure 7: Probability of soft sweep as a function of algebraic connectivity for 1000-node power-law cluster graphs. The clustering parameter p ranges from 0 to 1, with 10 graphs generated for each value. Simulations were run with $\tilde{\mu} = 0.000464$ for 100,000 runs per graph.

4.3.1 Watts–Strogatz Islands

We first explored core–periphery structures built from Watts–Strogatz subgraphs, with the core having a higher degree (20) than the peripheral clusters (degree 6). As the core size increases, the total number of high-degree nodes rises and the peripheral clusters shrink, creating a tension between the core's tendency to promote hard sweeps (due to its high connectivity) and the peripheries' tendency to sustain diversity (due to their low connectivity and relative isolation).

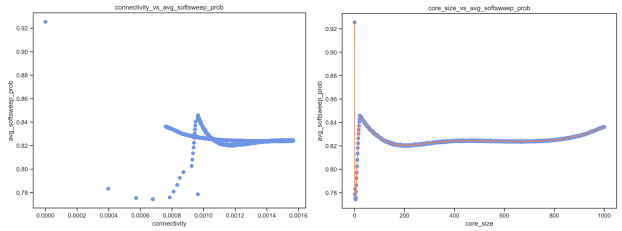


Figure 8: Probability of soft sweep as a function of core size for 1000-node Watts–Strogatz island graphs (core degree 20, peripheral degree 6, 1 connection per peripheral cluster). Simulations were run with $\tilde{\mu} = 0.000464$ for 100,000 runs per graph.

The results (Figure 8) reveal a striking non-monotonic pattern. Rather than a simple decrease in soft sweep probability as the (higher-degree) core grows, we observe a characteristic "dip": the soft sweep probability initially decreases, reaches a minimum at an intermediate core size, and then recovers. This behavior defies the simple predictions of the degree–diversity and connectivity–diversity relationships established above and suggests a more complex interplay between subpopulation structure and global diversity.

We interpret this dip as follows. When the core is small, it can be rapidly colonized by

a single allelic identity, which then has a high probability of spreading to the peripheries through the inter-cluster connections. At the same time, the peripheries are large enough that their collective mutant population is substantial, reducing the effective mutation rate for the overall system. The combination of these effects (a small, fast-sweeping core feeding a uniform allele to large, slow-diversifying peripheries) minimizes the probability of a soft sweep. As the core grows further, it eventually becomes large enough to sustain its own internal diversity, and the peripheries become too small to be easily dominated by the core's output, causing the soft sweep probability to recover.

4.3.2 Effect of Inter-Cluster Connections

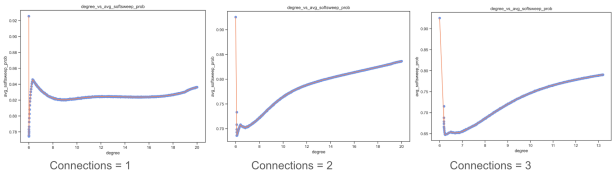


Figure 9: Probability of soft sweep as a function of core size for 1000-node Watts–Strogatz island graphs with varying numbers of inter-cluster connections (1, 5, and 10 connections per peripheral cluster). Core degree 20, peripheral degree 6. Simulations were run with $\tilde{\mu} = 0.000464$ for 100,000 runs per graph.

To test this interpretation, we varied the number of connections between the core and each peripheral cluster (Figure 9). Increasing the number of connections deepens the dip, consistent with our hypothesis: more connections increase the probability that the core's dominant allele can spread to the peripheries, strengthening the homogenizing effect and further suppressing diversity at the critical core size.

4.3.3 Regular and Well-Mixed Islands

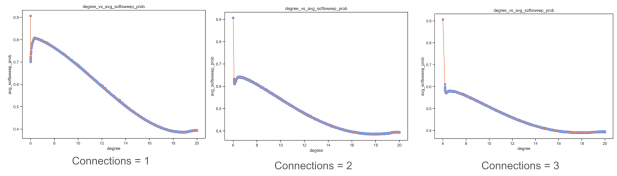


Figure 10: Probability of soft sweep as a function of core size for 1000-node regular island graphs (core degree 20, peripheral degree 6) with varying numbers of inter-cluster connections. Simulations were run with $\tilde{\mu} = 0.000464$ for 100,000 runs per graph.

The same non-monotonic pattern is observed in regular island graphs (Figure 10), though the dip is shallower, consistent with the earlier finding that random regular graphs exhibit a steeper degree–diversity decay than Watts–Strogatz graphs. Increasing the number of inter-cluster connections again deepens the dip.

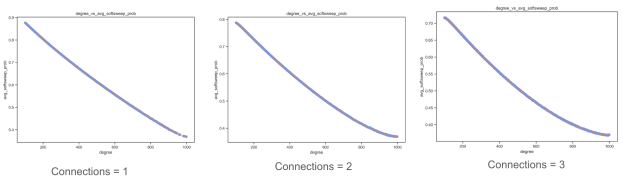


Figure 11: Probability of soft sweep as a function of core size for 1000-node well-mixed island graphs with varying numbers of inter-cluster connections. Simulations were run with $\tilde{\mu} = 0.000464$ for 100,000 runs per graph.

In well-mixed island graphs (Figure 11), however, the dip disappears entirely. Instead, the soft sweep probability decreases monotonically (though more slowly than a simple exponential) as the core grows. We attribute this to the fact that well-mixed clusters allow mutants to spread so rapidly within each cluster that the core cannot sustain any internal diversity long enough to create the conditions for the dip. The system effectively behaves as a smoothly interpolated version of a single well-mixed population of increasing size.

4.3.4 Approximate Decomposition of Soft Sweep Probability

To formalize the intuition developed above, we propose an approximate decomposition of the total soft sweep probability into contributions from the core, the periphery, and cross-cluster interactions:

$$P_{\text{total}} = \left(\frac{n}{N}\right)^2 P_{\text{core}} + \left(\frac{P_{\text{periph}}}{N}\right)^2 P_{\text{periphery}} + 2 \cdot \frac{n}{N} \cdot \frac{P_{\text{periph}}}{N} \cdot P_{\text{cross}},$$

where n is the number of core nodes, $P_{\text{periph}} = N - n$ is the total number of peripheral nodes, P_{core} and $P_{\text{periphery}}$ are the soft sweep probabilities of the core and peripheral clusters in isolation (estimated from our earlier simulations), and P_{cross} is a fitted parameter representing the probability that a randomly chosen core node and a randomly chosen peripheral node carry different allelic identities.

Figures 12 and 13 show the results of fitting this decomposition to the well-mixed and Watts–Strogatz island data, respectively. The approximation captures the qualitative shape of the soft sweep probability curves, including the dip in the Watts–Strogatz case. The fitted value of P_{cross} decreases as the number of inter-cluster connections increases, consistent with our earlier finding that greater connectivity between subpopulations reduces diversity by facilitating the spread of a dominant allele from the core to the periphery.

The decomposition does not perfectly fit the data in all cases, likely because P_{cross} is treated as a constant when it should in principle vary with the relative sizes of the core and periphery. Nevertheless, this framework provides a useful starting point for predicting soft sweep probabilities in heterogeneous population structures from the properties of their constituent subpopulations.

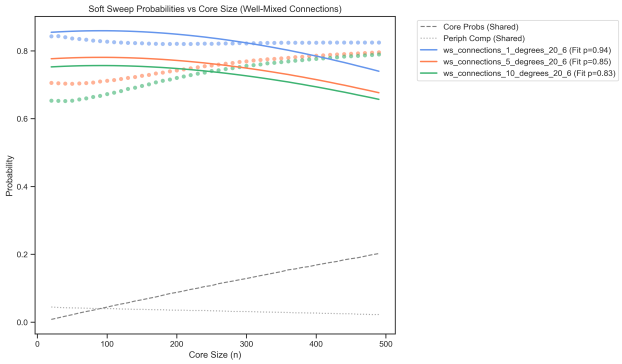


Figure 12: Soft sweep probability for well-mixed core–periphery graphs (points) compared with the approximate decomposition (curves). The fitted values of P_{cross} are given in the legend. Dashed black lines indicate the soft sweep probabilities of the core and peripheral clusters in isolation, estimated from independent simulations.

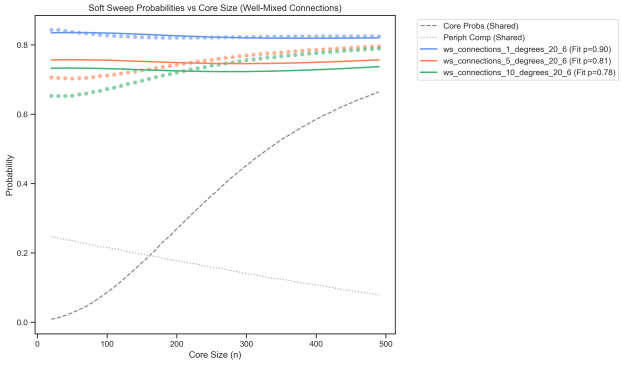


Figure 13: Soft sweep probability for Watts–Strogatz core–periphery graphs (points) compared with the approximate decomposition (curves). The fitted values of P_{cross} are given in the legend. Dashed black lines indicate the soft sweep probabilities of the core and peripheral clusters in isolation, estimated from independent simulations.

5 Comparison with Prior Work

Our findings both extend and complement the existing literature on soft sweeps in structured populations. The earliest computational models of soft sweeps, by Hermisson and Pennings [5], established the probability of a soft sweep in a well-mixed population as a function of the mutation rate and population size. Our well-mixed simulation results reproduce these predictions exactly, providing a validated baseline from which to explore the effects of structure.

Ralph and Coop [11] introduced spatial structure through a two-dimensional grid of demes, showing that allelic diversity at fixation depends on a characteristic length scale determined by the rate of allele spread relative to the rate of new mutation. Our results are consistent with this framework in the specific case of lattice-like structures (e.g., Watts–Strogatz graphs), but extend it substantially by demonstrating that the relationship between connectivity and diversity holds across a much wider range of graph topologies, including

those lacking the regularity and symmetry of a grid.

Paulose et al. [9] extended the grid model to incorporate long-range dispersal and analyzed the resulting effects on diversity. Our work builds directly on the Paulose et al. dynamical model but replaces the grid-plus-jumps framework with arbitrary graph topologies. By doing so, we are able to identify algebraic connectivity and average degree as general predictors of soft sweep probability across diverse graph families, a result that could not have been obtained within the grid-based framework.

Finally, Kuo et al. [7] examined how population structure affects clonal diversity and the time to fixation of a single mutant, showing that increased transitivity accelerates fixation. Our results on regular graphs with tuned transitivity are directly consistent with this finding and extend it to the multi-allele setting: we show that the same structural features that accelerate single-mutant fixation also suppress soft sweeps, providing a mechanistic link between fixation dynamics and allelic diversity.

6 Future Work

Several promising directions emerge from this work. First, a key open challenge is to identify an overarching network parameter, or a combination of parameters and interaction rules, that fully determines the probability of a soft sweep in an arbitrary structured population. While algebraic connectivity and average degree are strong predictors, they do not capture all of the relevant structural variation, and a more complete characterization may require incorporating higher-order topological features such as community structure, degree heterogeneity, or spectral properties beyond the Fiedler eigenvalue.

Second, the approximate decomposition of soft sweep probability into core, periphery, and cross-cluster components offers a promising framework, but the cross-cluster probability P_{cross} requires further investigation. In particular, deriving a functional relationship between P_{cross} and the number of inter-cluster connections, the relative sizes of the core and periphery, and the internal structure of each cluster would enable fully predictive (rather than fitted) estimates of soft sweep probability in composite networks.

Third, we aim to apply our framework to estimate the probability of a soft sweep in real-world populations by mapping empirical contact networks or population structures onto graph representations and using the relationships established here to predict diversity outcomes. Such applications could inform public health strategies for managing drug resistance or designing quarantine protocols.

Finally, it will be essential to validate our computational predictions against empirical genomic data. Comparing the allelic diversity observed in natural populations with the predictions of our models for population structures matching the known spatial organization of those populations would provide a powerful test of the theory and guide its refinement.

References

[1] Pamela F Colosimo, Kim E Hosemann, Sarita Balabhadra, Guadalupe Villarreal Jr, Mark Dickson, Jane Grimwood, Jeremy Schmutz, Richard M Myers, Dolph Schluter, and David M Kingsley. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *science*, 307(5717):1928–1933, 2005.

[2] Alison F Feder, Pleuni S Pennings, and Dmitri A Petrov. The clarifying role of time series data in the population genetics of hiv. *PLoS genetics*, 17(1):e1009050, 2021.

[3] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.

[4] Nandita R Garud, Philipp W Messer, Erkan O Buzbas, and Dmitri A Petrov. Recent selective sweeps in north american drosophila melanogaster show signatures of soft sweeps. *PLoS genetics*, 11(2):e1005004, 2015.

[5] Joachim Hermisson and Pleuni S Pennings. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4):2335–2352, 2005.

[6] Yang Ping Kuo and Oana Carja. Evolutionary graph theory beyond pairwise interactions: higher-order network motifs shape times to fixation in structured populations. *PLOS Computational Biology*, 20(3):e1011905, 2024.

[7] Yang Ping Kuo, Jiewen Hu, and Oana Carja. Clonal interference, genetic variation and the speed of evolution in structured populations. *bioRxiv*, 2025.

[8] Philipp W Messer and Dmitri A Petrov. Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution*, 28(11):659–669, 2013.

[9] Jayson Paulose, Joachim Hermisson, and Oskar Hallatschek. Spatial soft sweeps: patterns of adaptation in populations with long-range dispersal. *PLoS genetics*, 15(2):e1007936, 2019.

[10] Pleuni S Pennings and Joachim Hermisson. Soft sweeps ii—molecular population genetics of adaptation from recurrent mutation or migration. *Molecular biology and evolution*, 23(5):1076–1084, 2006.

[11] Peter Ralph and Graham Coop. Parallel adaptation: one or many waves of advance of an advantageous allele? *Genetics*, 186(2):647–668, 2010.

[12] Pardis C Sabeti, David E Reich, John M Higgins, Haninah ZP Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J McDonald, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002.

[13] Daniel R Schrider and Andrew D Kern. Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular biology and evolution*, 34(8):1863–1877, 2017.

[14] Sarah A Tishkoff, Floyd A Reed, Alessia Ranciaro, Benjamin F Voight, Courtney C Babbitt, Jesse S Silverman, Kweli Powell, Holly M Mortensen, Jibril B Hirbo, Maha Osman, et al. Convergent adaptation of human lactase persistence in africa and europe. *Nature genetics*, 39(1):31–40, 2007.

A Additional Data and Code

All code used for simulations, analysis, and graph generation is available at `https://github.com/mzsanborn/honors_thesis`.