Simulating the Cultural Evolution of Literary Genres

Graham Sack[1], Daniel Wu[2], Benjamin Zusman[3]

[1] Columbia University, English & Comparative Literature Department
gas2117@columbia.edu
[2] Harvard University, Sociology Department
danielwu@fas.harvard.edu
[3] University of Florida, Genetics Research Laboratory
bzusman@gmail.com

Abstract. The purpose of this paper is to explore the evolutionary dynamics of literary genre: the development of the 19th Century British novel is used as a motivating case study. The author constructs an agent-based model in NetLogo consisting of two interacting levels: (I) A genetic algorithm in which cultural forms (e.g., works of literature, pieces of music, etc.) are represented as binary feature strings. Cultural forms evolve across generations via asexual and sexual reproduction. Genres are represented as hierarchical clusters of similar feature strings. (II) Cultural forms are subjected to the selection pressure of consumer preferences. Preferences are heterogeneous: each consumer's tastes are represented by an ideal point in feature space. Preferences are configured in landscapes that vary in their levels of structure, entropy, and diversity. Landscapes are dynamic and may change due to (1) exogenous demographic shifts (e.g., population growth, generational turnover) or (2) endogenous feedback effects (e.g., preference co-evolution, conformity / anti-conformity effects).

Keywords: Cultural Evolution, Literary Theory, Agent Based Modeling

1 Introduction

The evolution of literary form and style is an emerging area of academic research and offers a valuable case study in cultural evolution generally. Several notable papers have appeared recently. In "Quantitative patterns of stylistic influence in the evolution of literature," Krakauer et al (PNAS, 2012) scale up methods traditionally used for authorship attribution to analyze stylistic shifts in the Project Gutenberg literary corpus. In the Genre Evolution Project, Simon and Rabkin at the University of Michigan postulate that literary genre is a complex adaptive system (CAS) and study its properties through the case study of science fiction. Related efforts are underway to 'map the literary genome,' using topic analysis as well as the mining of databases such as Aarne-Thompson-Unter's folktale motif collection (see Dar´anyi et al 2012, "Toward Sequencing "Narrative DNA": Tale Types, Motif Strings and Memetic Pathways").

adfa, p. 1, 2011.

Critic Franco Moretti's essay collection, *Graphs, Maps, Trees* (2005) is a particularly provocative example of literary evolution research. Based on an analysis of 19th Century British novels, Moretti offers the following speculative claims:

1. The growth of the reading public from the 18th to the 19th centuries (due to demographic changes and increasing literacy) resulted in a 'phase change' in the British novel circa 1820: novels became far more heterogeneous and generically differentiated, aimed at specialized niches rather than readers in general (Moretti, 8).
2. The average lifetime of genres is ~25-30 years, the same as a human generation. The reason for this historical rhythm is generational turnover in the reading public.[1]
3. Literary genre evolution is characterized by alternating cycles of divergence and converge – that is, periods of increasing generic diversity and differentiation followed by periods of decreasing diversity and cross-over (Moretti, 80).

Statistician Cosma Shalizi argues in his response, "Graphs, Trees, Materialism, Fishing," that while Moretti identifies provocative historical patterns, he stops short of fully articulating the mechanisms underlying and driving literary genre evolution:

> I don't think Moretti's time series, by itself, is enough to begin to let us decide among these mechanisms (some of which are compatible), but I do think it lets us see that some mechanism is called for… One thing Moretti does not do, anywhere, is construct models linking individual behavior to aggregate patterns (Shalizi, 118).

A similar criticism may be directed at the other papers cited, all of which are *descriptive* and based on *a posteriori* statistical analysis of corpora. The objective of this paper is to take up Shalizi's injunction by building a computational model of possible *generative* mechanisms driving genre evolution. We consider the following questions:

- How do the static characteristics and dynamic behavior of the "reading public" affect literary genre evolution?
  - How is generic diversity affected by reader diversity?
  - Is there a "phase change" in literary genre diversity as the reading public grows?
  - Under what circumstances will the life cycle of literary genres parallel the life cycle of generations?

2 Methodology

2.1 Feature Strings and Preference Strings

The model contains two basic agent populations: (1) *cultural forms* (e.g., books); and (2) *cultural consumers* (e.g., readers). The key attribute of agents in each population

[1] "Normal literature remains in place for twenty-five years or so… but where does this rhythm come from?... The causal mechanism must be external to the genres and common to all: like a sudden, total change of their ecosystem. Which is to say: a change of their audience. Books survive if they are read and disappear if they aren't: and when an entire generic system vanishes at once, the likeliest explanation is that its readers vanished at once. This, then, is where those 25-30 years come from: generations." (Moretti, 21)

is a bit string of user-specified length. For cultural forms, this bit-string represents the morphological features of the work (e.g., for literature, bits may represent attributes such as authorial style, length, plot, and theme). For cultural consumers, the bit-string represents an individual's ideal preference. Each consumer has a *tolerance* for variation from this ideal represented as an acceptable hamming distance. For example, if a particular reader has a preference string [1,1,1,1] and a tolerance of hamming distance 1, then he would be willing to consume cultural works with feature strings [1,1,1,1], [0,1,1,1], [1,0,1,1], [1,1,0,1], or [1,1,1,0].

2.2 Preference Landscapes

Individual cultural consumers are in turn organized into larger *preference landscapes*, which vary in their levels of structure, entropy, and reader diversity:

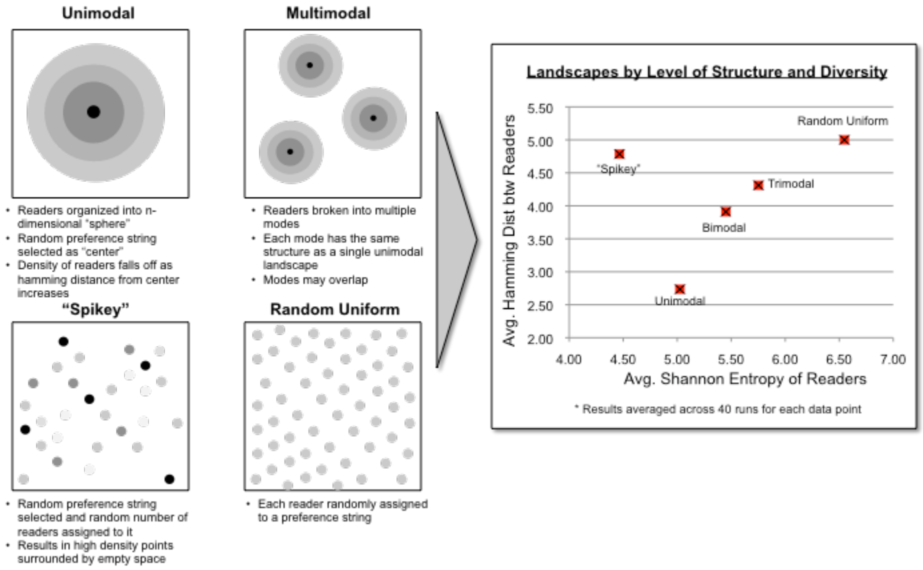


Fig. 1. Preference Landscape Types

Unimodal: This landscape represents circumstances in which reader preferences are homogeneous. Preferences are represented as an n-dimensional "sphere" of variable reader population density. The user specifies the radius of the sphere (expressed as a hamming distance) as well as the size of the reader population to distribute over it. The sphere is divided into segments – a randomly-assigned center, a layer of preferences at hamming distance of 1 away, a layer at hamming distance of 2 away, etc. with the density of readers decreasing as one moves radially outward.

Multimodal: This landscape represents circumstances in which reader preferences are clustered around two or more contrasting poles. The user specifies the number of modes. Each individual mode has the same structure as a unimodal sphere of preferences, but the modes vary randomly in the number of readers assigned to them.

Spikey: The limiting case of the multimodal landscape. Preferences are organized into isolated "spikes," with high reading population densities concentrated on single bit-strings surrounded by empty space. This landscape represents a case in which reader preferences are sharply divided into separate niches with no overlap and in which the regions of high preference density are totally uncorrelated with one another. The user specifies only the size of the reading population. The algorithm selects a random number between 0 and the number of unassigned readers and assigns them to a random bit-string. This process iterates until all readers are assigned.

Random Uniform: Readers are uniformly distributed over the preference space. This represents a case in which reader preferences are highly diffuse and have minimal organization and structure.

2.3 Genetic Algorithm

Once the preference landscape has been constructed at set-up, a genetic algorithm is run on the cultural forms in order to simulate evolution. The genetic encoding, as explained above, is a bit-string representing the features of each book. Each reader can consume a user-specified maximum number of books in each time period. The fitness of each book is measured by the number of readers it receives in that period. High fitness books are selected by tournament and are more likely to survive and reproduce, increasing their influence on the genetic content of the next generation. Three reproductive mechanisms are used:

Survival: books carry over from generation T to T+1 with no genetic change

Asexual: individual bit-strings from generation T are copied with a user-specified probability of mutation to create a new generation of books at T+1

Sexual: pairs of bit-strings from generation T are spliced in order to create a new generation of books at T+1

Each of these reproductive strategies has an intuitive interpretation in the context of cultural production. *Survival* corresponds to the case in which successful books are simply reprinted. *Asexual reproduction* corresponds to the case in which successful books spawn similar works with slight variation: that is, authors copy and modify the template provided by recently successful works. Sexual reproduction corresponds to what we might call "genre-crossing": authors take the features of two successful works and synthesize them in order to produce a new work.[2] The relative proportions of these reproductive strategies are user-specified. Users also specify the *mutation rate*, which is the probability that any bit will be mutated during either reproduction

[2] The current trend of "mash-up" literature provides a salient example. Best-sellers such as *Abraham Lincoln: Vampire Hunter* splice the features of already-successful genres (e.g., historical biography and gothic). Lest we dismiss such works as gimmicks, it is worth recognizing that many high-prestige genres emerged through hybridization. Modernist works such as James Joyce's *Ulysses* self-consciously combined the features of the realist novel with those of the classical epic. *Pastiche*, *bricolage*, and the combination of high and low art were central to postmodern literature, epitomized by William Burrough's "cut-up" novels. Recombination thus seems to be an widely-used mechanism in literary production.

process. The mutation rate also has an intuitive interpretation in the context of cultural production: it characterizes the inherent *creative experimentalism* of the cultural field, that is, how far authors are generally willing to depart from established models.

2.4 Clustering Algorithm

The focus of this study is not individual literary works, but rather the aggregate genres into which they are organized. “Genre,” however, is an ambiguous concept. It may be conceptualized in two ways: top-down or bottom-up. In the top-down case, genre functions as a generalized formula or set of conventions that *precedes* literary works and according to which they are constructed or judged. Alternatively, genre may function as a bottom-up phenomena: genres, in this case, are constructed after the fact as category labels for works that have similar characteristics.

For the sake of this paper, we are concerned only with genre in the bottom-up sense. To simulate this, we cluster books based on the statistical similarity of their feature-strings. A hierarchical clustering algorithm is employed to group books based on the average hamming distance between them. The algorithm works as follows: (1) the user specifies a minimum hamming distance to be used as a cut-off[3] (clusters that are separated by hamming distances above this cut-off will not be merged); (2) each book is initially placed in its own cluster; (3) each iteration, the algorithm merges the clusters separated by the minimum pairwise average hamming distance provided it is less than the cut-off; if not, the algorithm halts.

To measure generic diversity, several metrics are used. These include the number of clusters as well as the following:

$$\text{Shannon entropy} = - \sum p_i \log(p_i) \tag{1}$$

$$\text{Diversity} = \text{Inverse of Herfindahl-Hirschman Index} = 1 / (\sum p_i^2) \tag{2}$$

These metrics are calculated based on the probability distribution of books and are helpful for differentiating cases in which there are equal numbers of clusters, but books are distributed unevenly across them.[4] The Shannon entropy of clusters can be interpreted as the average number of questions one would have to ask to determine what cluster a book is in. Shannon entropy decreases as (1) the number of clusters decreases, (2) the frequency distribution of books across clusters becomes more skewed (and therefore more predictable).

2.5 Dynamic Landscapes

Thus far, we’ve assumed one-way causality: the preference landscape is (1) static and (2) formed in a vacuum. This is unrealistic: preferences change over time.

[3] A user-specified cut-off is preferred to specifying the number of clusters (as is required by K-means), since this would defeat the purpose of cluster statistics as a metric for genre diversity.

[4] For example, 100 books may be distributed across 4 clusters as {25, 25, 25, 25} or as {97, 1, 1, 1}. Both cases have the same number of clusters, but the Shannon entropy in the first case will be 2 bits, while in the later case it will be ~1.

Moreover, the prevailing culture shapes preferences just as preferences shape the culture. To address this, the model incorporates several landscape updating processes.

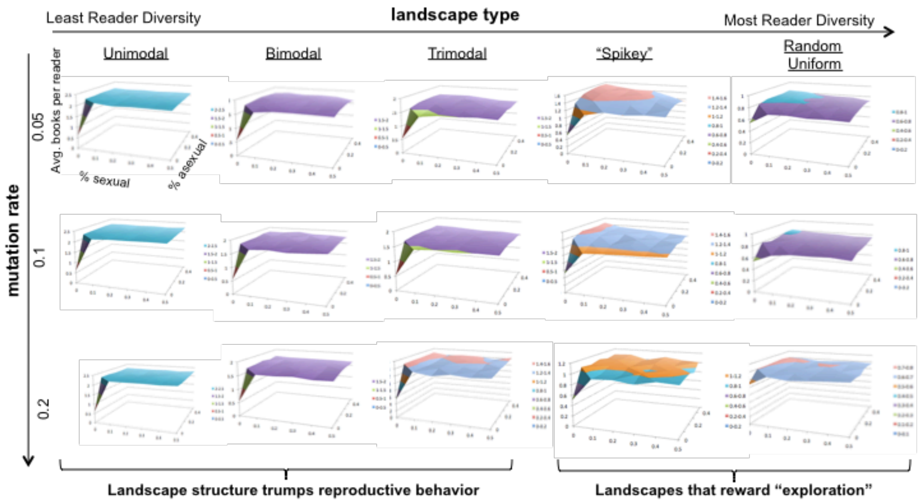
First, we allow for demographic changes. These include *growth* as well as reduction in size of the "reading public." We also include *generational effects*: sub-sections of the population update preferences synchronously at discrete time intervals.

Second, we allow for feedback effects. These include *co-evolution* -- whereby consumers who cannot find cultural forms that are within a tolerable distance of their preferences exit the market – as well as *conformity / non-conformity effects*. In the latter, popular cultural forms influence preferences. We assume the population is divided into conformers and non-conformers. Conformers update their preferences *towards* forms that are currently popular. Non-conformers update their preferences *away from* forms that are currently popular.

3 Results

3.1 Static Landscapes

Figures 2 and 3 show the results from evolving the cultural forms on different static preference landscapes. Figure 2 displays a five-dimensional graph in which each miniature surface shows the average number of books per reader vs. crossover rate vs. asexual reproduction rate. The miniature graphs, in turn, are organized along two axes: (1) landscape type and (2) mutation rate.[5]



* Each (x,y,z) coordinate on each surface is an average of 30 model runs

Fig. 2. Average Books per Reader vs. Landscape Types and Reproductive Behaviors

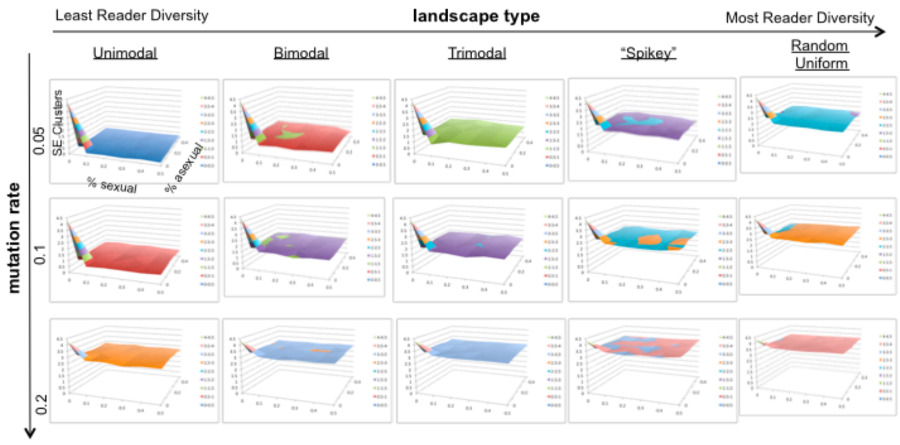
[5] The following values are held constant: # readers = 100; # books = 50; bit-string-length = 10; tolerance = 1 hamming distance; maximum # books a reader can consume in a period = 3.

Readership is a scarce resource in this model: each reader can consume at most 3 books per period. The *average number of books per reader* measures the net effectiveness of the cultural forms at extracting readership from the landscape. Figure 2 shows that *average readership decreases as reader diversity increases*. That is, the average books per reader falls as the landscape becomes more fragmented and reader preferences become more heterogeneous. Why is this so? Because even when readers are dispersed in preference space, a book's best survival strategy is to focus on regions of relatively higher reader density, leaving readers at the periphery with few if any options tolerably near their preferences. This is consistent with the behavior of real cultural markets: consumers with preferences close to the mainstream usually have many cultural products to choose from (though those products may be very similar to one other) while readers with preferences far from the mainstream have limited options and may not find anything that suits their idiosyncratic tastes.

Moreover, figure 2 shows that readership is maximized on highly structured preference landscapes (i.e., unimodal, bimodal, and trimodal) *regardless of reproductive strategy or mutation rate*. That is, the mechanisms in the model that are associated with creative activity (variation, genre-crossing, creative experimentation) have *minimal payoff in highly structured preference landscapes*. For any positive value of the mutation rate and the asexual or sexual reproduction rates, the cultural forms will do equally well on average in extracting the maximal readership from the landscape. High-risk creative strategies do no better than low-risk ones on a population-wide basis. This dynamic comports with the risk-averse creative behavior observed in mass markets with many homogeneous consumers. Writers and artists targeting these markets tend to be less experimental, perhaps because, as this model suggests, experimentation conveys no incremental benefit on average. Diffuse, fragmented, and niche markets (such as spikey and random uniform), on the other hand, do reward exploration: because pockets of reader density are dispersed in preference space, increasing asexual reproduction (which increases exploration of the preference landscape) raises average books for reader. This is also consistent with real cultural markets: in markets where consumer preferences are fragmented and therefore less well-known, cultural producers on average benefit from greater exploration of the product space.

Figure 3 displays the Shannon entropy of clusters and shows that generic diversity is unaffected by the mix of reproductive strategies – that is, *how* authors create new works (e.g., by modifying vs. splicing) has little impact on the ultimate level of generic diversity. However, *both the shape of the preference landscape and the mutation rate do have a significant impact on the generic diversity.* As reader preferences become more diverse, generic diversity correspondingly increases. This makes intuitive sense: a small set of genres is adequate to meet the tastes of a homogeneous mass audience while a diverse set of genres is required if there are many separate niches.

Generic diversity likewise increases with the mutation rate. This result is less obvious. The mutation rate, as noted above, captures the "experimentalism" of a particular creative market and can be thought of as the predilection of writers or artists to adhere closely to established / successful models or to depart from them significantly. What figure 3 suggests is that *highly experimental creative markets will produce greater generic diversity regardless of the structure of consumer preferences*. For example, a mutation rate of 0.2 on the unimodal landscape produces the same level of generic diversity as a mutation rate of 0.1 on a random uniform distribution (Shannon entropy of 2.5 – 3). This illustrates that increasing creative experimentation in an environment of homogeneous preferences has an equivalent effect on generic diversity as fragmenting the preference landscape. This means that *generic diversity cannot be explained solely in terms of consumer preferences – in order to explain the levels of generic diversity that we observe in cultural forms we also need to account for the artistic process*, in particular, how experimental vs. conservative it is. This is a point under-emphasized by Moretti, who explains generic diversity exclusively in terms of readership. The model shows that reader preferences are at best a partial mechanism and that the level of creative experimentation in a cultural form at a given historical moment is also a crucial input to explaining generic diversity.



* Each (x,y,z) coordinate on each surface is an average of 30 model runs

Fig. 3. Shannon Entropy of Clusters vs. Landscape Types and Reproductive Behaviors

3.2 Demographic Changes

The prior section concerned genre evolution on static preference landscapes. We next consider genre evolution on landscapes that shift over time due to demographic changes. Figures 4 and 5 show the impact of growth in the reading population. Seven different growth scenarios are considered for each of three different initial landscapes (unimodal, spikey, and random uniform). The scenarios differ based on two parameters: (1) the rate of population growth (no growth, 1%, 3% or 5%); (2) how preferences are assigned to new readers (new preferences can be mutations of existing preferences or can be generated randomly).

Figure 4 shows the effect of each growth scenario on reader diversity—measured as 1 / ($\sum p_i^2$). Of the initial landscapes, unimodal has the least diversity (~13), followed by spikey (~19), and random uniform (~90). If new readers have randomly assigned preferences, then all three landscapes respond the same way:

reader diversity increases exponentially in time regardless of starting conditions. This is not so if new preferences are mutations of those composing the original landscape. In this case, the landscapes respond differently. For spikey and random uniform landscapes, reader diversity actually *decreases* as new readers are added.[6] For unimodal, the effect of adding new readers is uncertain: it may increase or decrease reader diversity.

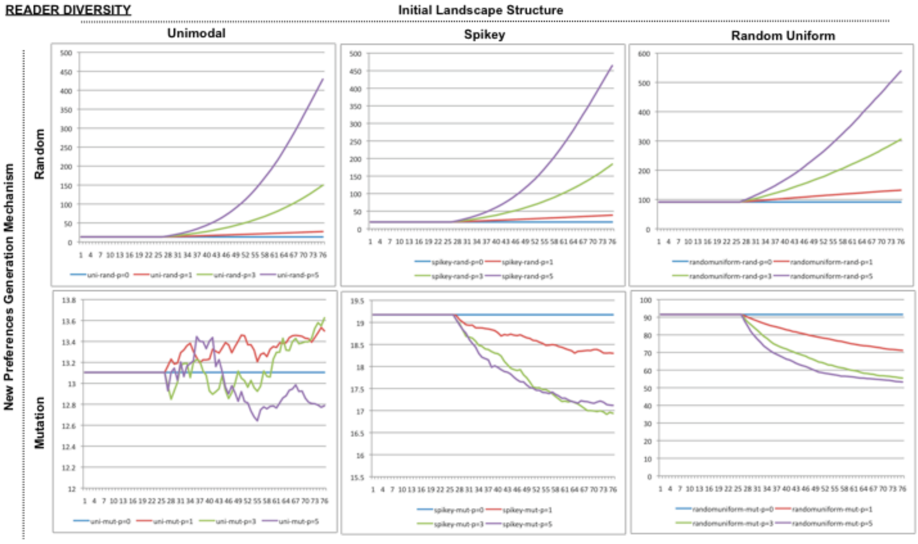


Fig. 4. Reader Diversity vs. Time for Various Population Growth Scenarios[7]

Figure 5 shows the effect of each growth scenario on generic diversity as measured by the Shannon entropy of clusters. We see that, in general, generic diversity tends to increase as time passes. This is indicated by the light blue lines on each graph, which show the no-population-growth scenarios. Introducing population growth can either increase or decrease generic diversity relative to the zero-growth scenario. Consistent with figure 4, if population growth increases reader diversity, then generic diversity increases *faster* than in the zero-growth case; while if population growth decreases reader diversity, generic diversity increases *slower* than the zero-growth case.

These graphs elucidate a gap in Moretti's claim that "the growth of the market creates all sorts of niches for 'specialist' readers and genres" (Moretti, 8). *Market growth does not in and of itself guarantee an increase in either reader diversity or generic diversity.* In fact, market growth may actually *reduce* reader and generic diversity under certain conditions. Whether the growth of a literary market increases diversity or not depends crucially on (1) whether the initial condition of the market was homo-

[6] The intuition is as follows. For spikey and random uniform landscapes, preferences are initially highly diverse. Adding mutations of existing preferences to the landscape means adding readers that are similar to those that exist, making these landscapes more homogeneous.

[7] Each data point in figures 4 and 5 is a composite averaged over 40 runs of the model.

geneous vs. diverse and (2) whether new readers have preferences that are similar to or different from the readers who already populate that market.

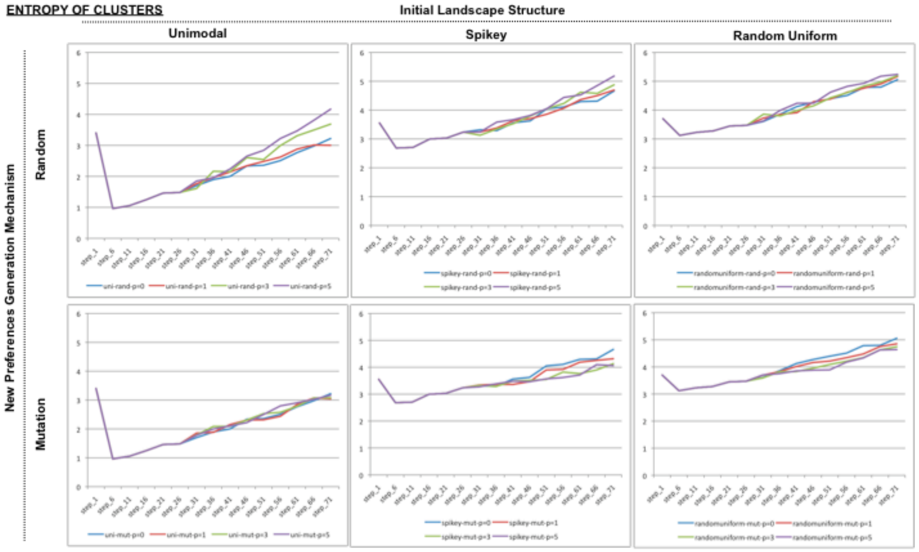


Fig. 5. Shannon Entropy of Clusters vs. Time for Various Population Growth Scenarios

3.3 Feedback Effects

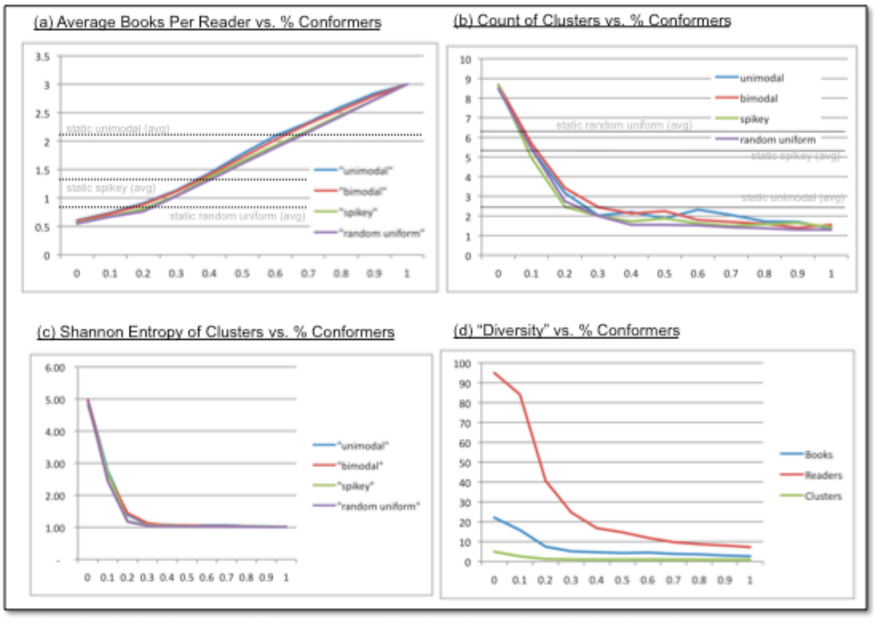
The previous sections assume one-way causality *from* consumer preferences *to* cultural forms. In this section, we consider scenarios in which cultural forms evolve on a dynamic landscape with two-way causality: preferences affect cultural forms, but cultural forms also affect preferences. We divide the reading population into two segments, (1) *conformers* and (2) *non-conformers*. Conformers update their preferences *towards popular books*, while non-conformers update their preferences *away from popular books*.[8]

Graph (a), (b), and (c) show how key outputs -- average books per reader, number of clusters, Shannon entropy of clusters – vary as the percentage of conformers is increased from 0% (everyone is a non-conformer) to 100% (everyone is a conformer). The first salient result is that feedback effects extremize the outcomes. With a high percentage of conformers, the average number of readers per book converges to 3, the maximum possible of 3. The reason is clear: when 100% of the readers are conform-

[8] Each tick, conformers check if the most popular book is within a tolerable distance of their preferences (meaning they are willing to consume it). If so, they do nothing. If not, a conformer compares his preference-string to the most popular book and identifies bits where they differ. He then flips one of his bits, shifting his preference-string one hamming distance closer to what is popular. Non-conformers do the opposite: they compare their preference-strings to the most popular book and update so as to move one hamming distance away.

ers, their preferences eventually all collapse to the feature set of the most popular book. The opposite is true for high percentages of non-conformers.

Second, the behavior of each landscape is effectively identical. Unlike with population growth, in which the dynamics differ between landscapes, for conformity-effects the dynamics overwhelm differences between the initial conditions: each conformer/non-conformer combination, effectively, has its own asymptotic landscape to which the preferences converge regardless of their original configuration.


(a) Average Books Per Reader vs. % Conformers

(b) Count of Clusters vs. % Conformers

(c) Shannon Entropy of Clusters vs. % Conformers

(d) "Diversity" vs. % Conformers

Inputs held constant: mutation rate = 0.05; sexual reproduction rate = .4; asexual reproduction rate = 0.4;
# books = 50; # readers = 100; dendrogram cut-off = .299; mode radius = 3

Fig. 6. Conformer / Non-Conformer Effects

Third, consistent with intuition, as the level of conformity increases, generic diversity decreases. However, it takes only a small percentage of conformers for generic diversity to collapse to its minimum. *Both the number of clusters and the Shannon entropy of clusters have a pronounced L-shape indicative of a phase change*: the generic diversity falls steeply as the percentage of conformers increases from 0 to 30%, but then levels off. Generic diversity barely changes as the percentage of conformers is raised from a minority of 30% to a majority of 100%. One explanation is that conformers are a consistent population, while non-conformers are transient. Once they converge on a popular books, conformer preferences do not to change any further due to positive feedback. For non-conformers, on the other hand, there is a negative feedback effect. Even if a book manages to capture non-conformer readers for a period, its popularity will cause those non-conformers to change their preferences shortly thereafter, so the book eventually loses its readers. Because the non-conformer population cannot support a consistent book population, stable genre-clusters develop around the conformer population, even if that population is relatively small. This has interesting implications for cultural markets: it suggests that the preferences of a large segment of the population may have no effect on the diversity of cultural products. Product diversity is instead determined by a small but tractable population of conformers.

4 Conclusion

The majority of agent-based models of cultural evolution are focused on what might be termed "internal culture"—e.g., belief propagation or the emergence of behavioral norms such as pro-sociality. Comparatively few models explore what might be termed "external culture"—that is, the artifacts, forms of expression, tools, and technologies that humans make. Whereas the former class generally understand cultural evolution as a process of social learning, the latter understand cultural evolution as morphological change in physicalized cultural artifacts over time. This approach calls for different methods. Whereas most models of internal culture use a single population of agents who hold particular beliefs or operate according to particular norms, modeling of external culture more naturally calls for two populations: cultural forms that evolve under selective pressure and human agents that consume and/or create them. This study is intended as an initial effort to develop such a model.

The results suggest a number of preliminary but valuable insights about plausible mechanisms driving literary evolution. In particular, we find that a number of critic Franco Moretti's claims about genre evolution require refinement. First, generic diversity cannot be explained solely in terms of the characteristics of the reading public: we also need to account for the characteristics of the creative process, in particular, the level of experimentation in the cultural market at a given historical moment (represented in this model by the mutation rate). Second, we show that growth in the reading public is not sufficient to guarantee an increase in either reader diversity or generic diversity. In fact, market growth may actually reduce diversity under certain conditions. To determine whether the effect that the growth of a cultural market will have, we need to know whether the preference landscape was initially homogeneous vs. diverse and whether new readers have preferences that are similar to or different from the readers who already populate that market.

Acknowledgements

This model was developed during the authors' attendance at the Santa Fe Institute's Complex Systems Summer School. We wish to thank the SFI faculty, administration, and fellows for their guidance and support.

References

1. Moretti, Franco. *Graphs, Maps, Trees*. Verso: New York, 2005.
2. Shalizi, Cosmo. “Graphs, Trees, Materialism, Fishing.” *Reading Graphs, Maps, Trees: Responses to Franco Moretti*. South Carolina: Parlor Press, 2011.
3. Michel, J.B et al. “Quantitative Analysis of Millions of Books.” *Science*. January, 2011.
4. Krakauer, D. et al. “Quantitative patterns of stylistic influence in the evolution of literature.” PNAS, May 2012.
5. Rabkin, Eric and Simon, Carl.
6. S. Dar´anyi, P. Wittek, L. Forr´o. “Toward Sequencing “Narrative DNA”: Tale Types, Motif Strings and Memetic Pathways.” Proceedings Computational Models of Narrative. LREC. Istanbul, 2012.