A Framework for Representing Language Acquisition in a Population Setting

Jordan Kodner
University of Pennsylvania
Department of Linguistics,
Dept. of Computer and Info. Science
jkodner@sas.upenn.edu

Christopher M. Cerezo Falco
University of Pennsylvania
Department. of Electrical
and Systems Engineering
ccerez@seas.upenn.edu

Abstract

Language variation and change are driven both by individuals' internal cognitive processes and by the social structures through which language propagates. A wide range of computational frameworks have been proposed to connect these drivers. We compare the strengths and weaknesses of existing approaches and propose a new analytic framework which combines previous network models' ability to capture realistic social structure with practically and more elegant computational properties. The framework privileges the process of language acquisition and embeds learners in a social network but is modular so that population structure can be combined with different acquisition models. We demonstrate two applications for the framework: a test of practical concerns that arise when modeling acquisition in a population setting and an application of the framework to recent work on phonological mergers in progress.

1 Introduction

The process of language change should be thought of as a two-step cycle in which 1) individuals acquire their native languages from their predecessors then 2) pass them on to their successors. Small changes accrue over time this way and create both small-scale interpersonal variation and large-scale typological differences. It is easy to draw a strong analogy here between linguistic evolution and biological evolution. Both feature classic descent with modification, except while phenotypes are transmitted through genes and acted on by natural selection, language is both transmitted through and constrained by the individual (Cavalli-Sforza and Feldman, 1981; Ritt, 2004, etc.).

But while evolution, linguistic or otherwise, is driven by forces acting on the individual, it unfolds on the level of populations (Cavalli-Sforza and Feldman, 1981). The influence of community-level social factors on the path of language change is a major focus of sociolinguistics (Labov, 2001; Milroy and Milroy, 1985; Rogers Everett, 1995). Ideally, one could observe population-level variation unfold in real time while testing out individual factors, but this is impossible because nobody can travel back in time or fit entire natural environments into a lab. Change that has already happened is out of reach, and change in progress is buried in a world of confounds. The classic sociolinguistic method instead approaches the problem by inferring causal factors from patterns discovered in field interviews and corpora (Labov, 1994; Labov et al., 2005, etc.). This is the primary source of empirical data in the field and the only way to look at language change in a naturalistic setting, but it is limited in that it cannot test cause and effect directly. More recently, controlled experimental studies have emerged as a complementary line of research which manipulate causal factors directly (Johnson et al., 1999; Campbell-Kibler, 2009, etc.), but are inherently removed natural time and scale. A third approach, the one we build upon here, relies on computational modeling to simulate how sociolinguistic factors might work together in larger populations (Klein, 1966; Blythe and Croft, 2012; Kauhanen, 2016, etc.).

It has long been argued that language acquisition is the primary cause of language change (Sweet, 1899; Lightfoot, 1979; Niyogi, 1998, etc.). In the last few decades, this connection has been modeled computationally (Gibson and Wexler, 1994; Kirby et al., 2000; Yang, 2000,

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 1149–1159
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

etc.), leading to the strong conclusion that change is the inevitable consequence of mixed linguistic input or finite learning periods (Niyogi and Berwick, 1996), even if children are "perfect" learners. An important result connecting the learner and population emphasizes the need for this line of work: the space of paths of change available in populations is formally larger than the paths available to linear chains of iterated learners. Niyogi and Berwick (2009) prove formally that even *perfectly-mixed* (i.e., uniform and homogeneous social network) populations admit phase transitions in the path of change unavailable to chains of single learners commonly implemented in iterated learning (Kirby et al., 2000). This suggests that small-population experimental studies in sociolinguistics and in child language acquisition do not paint the full picture of language change.

We introduce a new framework for modeling language change in populations. It has an outer loop to represent generational progression, but it replaces the inner loop which calculates randomized interactions between agents with a single formula that is defined generally enough to allow the simulation of a wide range of scenarios. It builds upon the principled formalism described by Niyogi and Berwick (1996, et seq.), privileging the acquisition model and separating it from the population model. The resulting modular framework is described in the following sections. First, Section 1.1 presents a survey of previous simulation work followed by a description of the new population model in Section 2. Next, Section 3 addresses practical concerns relating population size to assumptions about language acquisition. Finally, Section 4 introduces a case study on phonological change which demonstrates the need for appropriate models both of acquisition and populations.

1.1 Related Work

Computational models for the propagation of linguistic variation have been employed with a variety of research goals in mind. Every paper implements its own framework with few exceptions, so comparison across studies is difficult. Additionally, since each model is essentially 'boutique,' it is always possible that models are designed consciously or unconsciously to achieve a specific outcome rather driven by underlying principles. We group these frameworks into three classes according to their implementation, *swarm*, *network*, and *algebraic*, and discusses their strengths and weaknesses.

The first class, called *swarm* here, models populations as collections of agents placed on a grid. They "swarm" around randomly according to some movement function, and "interact" when they occupy adjacent grid spaces (Satterfield, 2001; Harrison et al., 2002; Ke et al., 2008; Stanford and Kenny, 2013). This tends toward concrete interpretation, for example, more mobile populations are expressed directly by more mobile agents. They capture Bloomfield (1933)'s "principle of density" which describes the observation that geographically or socially close individuals interact more frequently than those farther away. On the other hand, they provide little control over network structure, relying on series of explicit movement constraints in order to direct their agents, and since each one moves randomly at each iteration, these models have potentially thousands of degrees of freedom. Such simulations should be run many times if any sort of statistically expected results are to be computed.

The second class, *network* frameworks, model speakers as nodes and interaction probabilities as weighted edges on network graphs (Minett and Wang, 2008; Baxter et al., 2009; Fagyal et al., 2010; Blythe and Croft, 2012; Kauhanen, 2016). These frameworks offer precise control over social network structure and can test specific community models from within sociolinguistics. However, implementations usually proceed by some kind of iterative probabilistic node-pair selection process, and in this way suffer from the same statistical pitfalls as swarm frameworks. In contrast to swarm models, interaction is rigidly restricted to immediately connected nodes, so to achieve gradient interaction probabilities, edges must be frequently updated or nearly fully-connected graphs with carefully assigned edge weights would need to be constructed and motivated.

The third class, *algebraic* frameworks, present analytic methods for determining the state of the network at the end of each iteration rather than relying on stochastic simulation of individual agents (Niyogi and Berwick, 1996, 1997; Yang, 2000; Baxter et al., 2006; Minett and Wang, 2008; Niyogi and Berwick, 2009). Removing that inner loop is a more mathematically elegant approach and avoids dealing unnecessarily with statistics behind random trials. Removing that loop speeds up calculation as well, making larger simulations more tractable than with network or swarm frameworks. But this power is achieved by sacrificing the social network. Up to this point, such models have, to our knowledge, only been defined over perfectly-mixed (i.e., no network effects) populations. That assumption is useful for reasoning about the mathematical theory behind language change, but it hinders such models' utility in empirical studies. For example, though Baxter et al. (2006) and Minett and Wang (2008) implement algebraic models for perfectly mixed populations, they fall back on network models to model network effects.

2 Framework for Transmission in Social Networks

Algebraic frameworks have their mathematical advantage, but network frameworks provide a richer model for representing real-world population structures and swarm models capture density effects by default. An ideal framework would combine the benefits of all three of these. Here we do just that. We introduce a framework that instantiates Niyogi and Berwick (1996)'s acquisition-driven formalism where change is handled explicitly as a two-step alternation between individual learners learning and populations interacting. It provides an analytic solution to the state of a network structure over which swarm-like behavior can be modeled.

We begin by conceptualizing the framework in terms of agents traveling probabilistically over a network structure as in Algo. 1 before introducing the analytic solution. There is an individual standing at every node in the graph, and at every iteration, each individual begins at some location and travels along the network's edges, at each step deciding to continue on or to stop and interact with the agent at that node. Any two agents with a non-zero weight path between them could potentially interact, so the overall probability of an interaction is a function of the shape of the network and the decay rate of the step probability. The shorter and higher weighted the path between two agents, the more likely they are to interact. This corresponds to the gradient interaction probabilities of swarm frameworks.

Algorithm 1: One iteration of the propagation model conceptualized on the level of an individual agent

```
for each individual node do
    Begin traveling;
    while traveling do
        Randomly select an outgoing edge by
          weight and follow it OR stop travel;
        increase chance of stopping next time;
    end
    Interact with the individual at the current
      node;
end
```

2.1 Representing the Network

Social networks are typically conceived of as graph structures with individuals as vertices and the social or geographical connections between individuals as edges, and this allows for a great deal of flexibility. If edges are undirected, then all interactions are equal and bidirectional, but if edges are directed, interactions may or may not be. Edges can be weighted to represent likelihood of interaction or some measure of social valuation, and this too can vary over time. Lastly, it is possible to add and remove nodes themselves to capture births, deaths, or migration.

The network structure is represented computationally here as an adjacency matrix $\mathbf{A}$. In a population of n individuals, this is $n \times n$ where each element a_{ij} is the weight of the connection from individual j to individual i. The matrix must be column stochastic (all columns sum to 1 and contain only positive elements) so that edge weights can be interpreted as probabilities. The special case where the matrix is symmetric (every $a_{ij} = a_{ji}$) models undirected edges, and more strongly, the model reduces to perfectly-mixed populations when each $a_{ij} = \frac{1}{n}$.

We define a notion of *communities* over the nodes of the network in order to add the option to categorize groups of individuals. Membership among c communities is identified with an $n \times c$ indicator matrix $\mathbf{C}$. Depending on the problem at hand, it is possible to calculate the average behavior of the learners within each community directly without having to calculate the behavior of each individual member.

2.2 Propagation in the Network

In a typical network model, the edge weights between nodes in $\mathbf{A}$ are interpreted directly as interaction probabilities, meaning that individuals only ever interact with their immediate graph neighbors. We take a different approach by allowing the agents to "travel" and potentially interact with any other agent whose node is connected by a path of non-zero edges. If the number of traveling steps were fixed at k, the probability of each pair interacting would be defined as $\mathbf{A}^k$. It is more complicated for us since the number of steps traveled is a random variable. The probability of j interacting with i ($p(ij)$) is the probability of them interacting after k steps times the probability of k for all values of k as in Eqn. 1. Combining this intuition with $\mathbf{A}$ yields the interaction probabilities for all i, j pairs.

$$p(ij) = \sum_k p(ij|k \text{ steps})\; p(k \text{ steps}) \tag{1}$$

The pattern of linguistic variants or grammars (in the formal sense where grammar g is the intensional equivalent of language L_g) within a network unfolds as a dynamical system over the course of many iterations, and learners' positions within the network mediate which ones they eventually acquire. In a system with g grammars and n individuals, a $n \times g$ row-stochastic matrix $\mathbf{G}$ specifies the probability with which each community expresses each grammar. Given this notion of interaction and the specification of grammars expressed within a network, it is possible to compute the distribution of grammars presented to each learner. This is the learners' linguistic environment and is represented by a matrix $\mathbf{E}$ in the same form as $\mathbf{G}^\top$.

An *environment function* $\mathcal{E}_n(\mathbf{G}_t, \mathbf{A}) = \mathbf{E}_{t+1}$ shown in Eqn. 2 calculates $\mathbf{E}$ by first calculating all the interaction probabilities in the network then multiplying those by the grammars which every agent expresses to get the environment $\mathbf{E}$. The α parameter from the geometric distribution[1] defines the travel decay rate. A lower α defines conceptually more mobile agents.

More generally, $\mathcal{E}_n$ is a special case of $\mathcal{E}(\mathbf{G}_t, \mathbf{C}_t, \mathbf{A}_t) = \mathbf{E}_{t+1}$ where the number of communities equals the number of individuals ($c = n$). $\mathbf{C}$ becomes the identity matrix without loss of generality, so the network's initial condition does not have to be defined explicitly. For any other community definition, an initial condition has to be defined as in Eqn. 3 which specifies the starting point in the network that each agent conceptually begins traveling from. The output of $\mathcal{E}$ is a $g \times c$ matrix giving the environment of the average agent in each community.[2]

$$\mathcal{E}_n(\mathbf{G}_t, \mathbf{A}) = \mathbf{G}_t^\top \alpha \left(\mathbf{I} - (1-\alpha)\mathbf{A}\right)^{-1} \tag{2}$$

$$\mathcal{E}(\mathbf{G}_t, \mathbf{C}, \mathbf{A}) = \mathcal{E}_n(\mathbf{G}_t, \mathbf{A})\mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1} \tag{3}$$

The output of $\mathcal{E}$ must be broadcast to $g \times n$, which would result in the loss of some information unless the assumption can be made that each community is internally uniform. However, when that assumption can be made, the $n \times n$ adjacency matrix admits a $c \times c$ equitable partition $\mathbf{A}^\pi$ (Eqn. 4) (Schaub et al., 2016) which permits an alternate environment function $\mathcal{E}_{EP}(\mathbf{G}_t, \mathbf{C}, \mathbf{A})$ shown in Eqn. 5 that is equivalent to the lossless $\mathcal{E}_n$ if $\mathbf{A}$. If $n \gg c$, $\mathcal{E}_{EP}$ is much faster to calculate because it only inverts a small $c \times c$ matrix rather than a large $n \times n$. This makes it feasible to run much larger simulations than what has been done in the past.

$$\mathbf{A}^\pi = (\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top\mathbf{A}\mathbf{C} \tag{4}$$

$$\mathcal{E}_{EP} = \alpha\mathbf{G}^\top\mathbf{C}\left(\mathbf{I} - (1-\alpha)\mathbf{A}^\pi\right)^{-1}(\mathbf{C}^\top\mathbf{C})^{-1} \tag{5}$$

2.3 Learning in the Network

The environment function describes what inputs $\mathbf{E}_{t+1}$ are available to learners given the language expressed by the mature speakers of the previous age cohort with grammars $\mathbf{G_t}$. The second component of the framework describes the learning algorithm $\mathcal{A}(\mathbf{E}_{t+1}) = \mathbf{G}_{t+1}$, how individuals respond to their input environment. The resulting $\mathbf{G}_{t+1}$ describes which grammars those learners will eventually contribute to the subsequent generation's environment $\mathbf{E}_{t+2}$. This back-and-forth between adults' grammars $\mathbf{G}$ and childrens' environment $\mathbf{E}$ is the two-step cycle of language change (Fig. 1).

In *neutral change*, learners would acquire grammars at the rates that they are expressed in their environments, but there is good reason to believe

[1] In this paper, jump probabilities decay according to a geometric distribution, but other distributions including the Poisson have been implemented as well.

[2] $(\mathbf{I} - (1-\alpha)\mathbf{A})^{-1}$ and $\mathbf{C}(\mathbf{C}^\top\mathbf{C})^{-1}$ can be precomputed if network structure does not change over time.

$$\ldots \mathbf{G_t} \rightarrow \mathbf{E_{t+1}} \rightarrow \mathbf{G_{t+1}} \ldots \mathbf{G_{t+i}} \rightarrow \mathbf{E_{t+i+1}} \ldots$$

Figure 1: Language change as an alternation between **G** and **E** matrices

that most language change involves differential fitness between competing variants, and most non-trivial learning algorithms yield some kind of fitness (Kroch, 1989; Yang, 2000; Blythe and Croft, 2012, etc.), so $\mathcal{A}$ is rarely neutral. A neutral and simple advantaged model are both considered in Section 3, and a more complex learning algorithm is described for Section 4.

3 Application: Testing Assumptions

The general nature of the framework described here renders it suitable for reproducing the results of previous works and evaluating their assumptions. To demonstrate this, we reproduce the major result from Kauhanen (2016), which tested the behavior of neutral change in networks of single-grammar learners, in order to dissect two of its primary assumptions. Implemented in a typical network framework, the original setup contains $n = 200$ individuals in probabilistically generated centralized networks in which individuals mature categorically to the single most frequent grammar in their input. The author found that categorical neutral change produced chaotic paths of change regardless of network shape and that periodically "rewiring" some of the network edges smoothed this out. Without commenting on rewiring, we find that the combination of n and choice of categorical learners conspire to create the chaotic results.

We create two communities, both centralized along the lines of the single cluster in Kauhanen (2016), initialize all members of cluster 1 with grammar g_1 and all members of cluster 2 with grammar g_2, and additional edges are added between members of clusters 1 and 2 to allow interaction. **G** is converted to an indicator matrix at the end of each learning iteration by rounding values to 0 and 1 in order to model categorical learners who only internalize the most common grammar in their inputs as in the original model.

In a pair of infinitely large clusters or two clusters where individuals are permitted to learn a probabilistic distribution of grammars, each cluster should homogenize to a 50/50 distribution of g_1 and g_2 after some number of iterations depending on the specifics of the network shape and setting for α creating the red curves in Fig. 2. At $n = 20000$, each of 10 trials roughly follows the path of the predicted curve, but when run at the original $n = 200$ for 10 trials, this produces the type of chaotic behavior which Kauhanen (2016) attempts to repair. The outcome appears to be the result of an assumption made out of convenience ($n = 200$) rather than a principled decision.

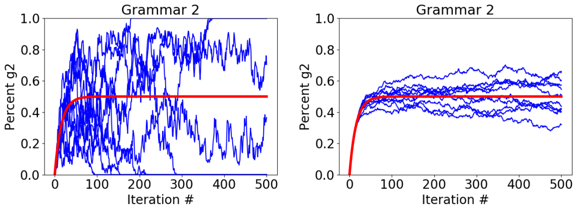

Figure 2: Predicted curve (red); neutral change at $n = 200$ (left; Kauhanen (2016)); neutral change at $n = 20000$ (right)

To further explore the impact of the population size assumption, we experiment on a model of advantaged change, which is typically contrasted with neutral change because of its tendency to produce "well-behaved" S-curve change (Blythe and Croft, 2012; Kauhanen, 2016). This time, only a single cluster is created, and the advantaged grammar is initially assigned to 1% of the population. As seen in Figure 3, results are chaotic for $n = 200$ once again and near predicted for $n = 20000$. This is important because at $n = 200$, advantaged change is chaotic, and most simulations both rise and fall. An experimenter who only studied advantaged change in small population might concluded that it is as ill-behaved as neutral change. While the conclusions that Kauhanen (2016) draws appear valid for $n = 200$, it is not clear to what extent they can be projected onto larger populations. This demonstrates the need for carefully choosing one's modeling assumptions and testing them out when possible.

4 Application: Mergers in Progress

The acquisition of phonological mergers in mixed input settings presents an interesting problem. It appears that mergers have an inherent advantage because they tend to spread at the expense of distinctions, and once they begin, they are rarely reversed (Labov, 1994). Yang (2009)'s acquisition model quantifies this advantage as the relatively

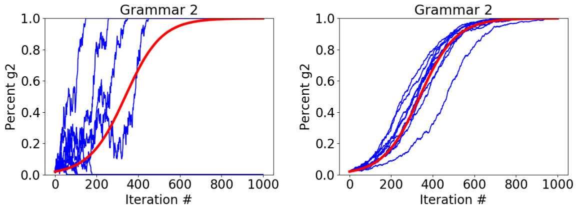

Figure 3: predicted curve (red); advantaged change at $n = 200$ (left; cf. Kauhanen (2016)); advantaged change at $n = 20000$ (right)

lower chance of misinterpretation if a listener assumes the merged grammar instead of the non-merged grammar once a sufficient proportion of the environment is merged. Applied to Johnson (2007)'s detailed population study of the frontier of the COT-CAUGHT merger in the small towns along the border between Rhode Island and Massachusetts, this accurately predicts the ratio of merged input for a child to acquire the merged grammar, however when applied to a perfectly mixed population of learners, it fails to model the spread of the merged grammar in the population. Yang's model is input-driven, so it is conducive to simulation with minimal assumptions past those drawn from the empirical data. We test the behavior of this learning model in a typical population network and demonstrate that it produces a reasonable path of change.

4.1 Background

The COT-CAUGHT *merger*, also called the *low back merger* describes the phenomenon present in varieties of North American English spoken in eastern New England, western Pennsylvania, the American West, and Canada among others where the vowel in words like *cot* and the vowel in words like *caught* have come to be pronounced the same (Labov et al., 2005, pp. 58-65). The geographical extent of the merger is currently expanding, which might be expected if the merger has a cognitive or social advantage associated with it. Johnson (2007)'s study of the merger's frontier on the border Rhode Island and Massachusetts uncovered an interesting social dynamic that illustrates the merger's speed: there are families where the parents and older siblings non-merged, but the younger siblings are. The merger has swept through in only a few years and passed between the siblings.

Yang (2009) seeks to understand why mergers have an advantage from a cognitive perspective, and his model treats the acquisition of mergers as an evolutionary process. Learners who receive both merged (M_+) and non-merged (M_-) input entertain both a merged (g_+) and non-merged (g_-) grammar and reward whichever grammar successfully parses the input. This kind of variational learner (Yang, 2000) is essentially an adaptation of the classic evolutionary Linear Reward Punishment model (Bush and Mosteller, 1953). The fitness of each grammar is the probability in the limit that it will fail to parse any given input, and since it is virtually always the case that this probability is different for both grammars, fitness is virtually always asymmetric. The variational learner is characterized as follows.

Given two grammars and an input token s, The learner parses s with g_1 with probability p and with g_2 with probability $q = 1 - p$. p is rewarded according to whether the choice of g successfully parses s ($g \rightarrow s$) or it fails to ($g \nrightarrow s$), where γ is some small constant.

$$p' = \begin{cases} p + \gamma q, & g \rightarrow s \\ (1-\gamma)p, & g \nrightarrow s \end{cases}$$

Given a specific problem, one can calculate a *penalty probability* C for each g, the proportion of input that would cause $g \nrightarrow s$. The grammar with the lower C has the advantage, so the other one will be driven down in the long run. C can be estimated from type frequencies in a corpus, and the model is non-parametric because these values do not depend on γ.

$$\lim_{t\to\infty} p_t = \frac{C_2}{C_1 + C_2} \qquad \lim_{t\to\infty} q_t = \frac{C_1}{C_1 + C_2}$$

To understand the COT-CAUGHT merger empirically, one must reason about what kind of input would trigger a penalty and then calculate the penalty probabilities of the merged grammar C_+ and non-merged grammar C_- from a corpus. This model considers parsing failure to be the rate of initial misinterpretation, and for a vowel merger, the only inputs that could create an initial misinterpretation are minimal pairs because they become homophones. Examples of COT-CAUGHT minimal pairs include *cot-caught*, *Don-Dawn*, *stock-stalk*, *odd-awed*, *collar-caller*, and so on.

The merged g_+ grammar collapses would-be minimal pairs into homophones, so the penalty rate C_+ comes down to lexical access. Under the observation that more frequent homophones are retrieved first regardless of syntactic context (Caramazza et al., 2001), g_+ listeners only suffer initial misinterpretation when the less frequent member of a pair is uttered regardless of the rate of M_+. If H is the sum token frequency of all minimal pairs and h^i_o, h^i_{oh} are the frequencies of the ith pair's members, then $C+$ is calculated by Eqn. 6.

In contrast, g_- listeners are sensitive to the phonemic distinction, so they misinterpret M_- input at the rate of mishearing one vowel for the other ϵ (Peterson and Barney, 1952) (second half of Eqn. 7). And given M_+ input, they misinterpret whenever they hear the phoneme which g_- does not expect (e.g., a merged speaker pronouncing *cot* with the CAUGHT vowel) times the probability of *not* mishearing that vowel (1-ϵ) plus ϵ times the probability of hearing the right vowel (i.e., the merged speaker pronounces *cot* with the COT vowel but it is misheard anyway) (first half of Eqn. 7). Since g_- misinterpretation rates are a function of the rate of M_+ (p) in the environment, there is a threshold of M_+ speakers above which the merged grammar has a fitness advantage over the non-merged one.

$$C_+ = \frac{1}{H}\sum_i \min(h^i_o, h^i_{oh}) \tag{6}$$

$$\begin{aligned} C_- = \frac{1}{H}\sum_i \big[p_0((1-\epsilon_{oh})h^i_o + \epsilon_{oh}h^i_{oh}) \\ +q_0(\epsilon_{oh}h^i_o + \epsilon_{oh}h^i_{oh})\big] \end{aligned} \tag{7}$$

Calculating this threshold for the frequent minimal pairs that Yang extracts from the Wortschatz project (Biemann et al., 2004) corpus[3] and mishearing rates from Peterson and Barney (1952), the Yang model predicts that a learner exposed to at least $\sim$ 17% COT-CAUGHT-merged input will acquire the merger. This threshold represents a strong advantage for M_+ because it is well under the 50% threshold expected for neutral (non-advantaged) change and it is very close to what was found in Johnson (2007)'s sociolinguistic study. It predicts that younger children may have g_+ while their parents and even older siblings have g_- if the 17% threshold was crossed in **E** after the acquisition period of the older sibling but before that of the younger sibling.

4.2 Model Setup

All the mechanics behind the learning model reduce to a simple statement: *learners acquires* g_+ *iff* $> 17\%$ *of their input is* M_+ *and they acquire* g_- *otherwise*. However, this kind of categorical learner in a perfectly-mixed population leads to immediate fixation at either g_- or g_+ in a single iteration, since the proportion of g_+ speakers in the population is equivalent to the proportion of M_+ input in every learner's environment. This is not realistic change. Clearly, social network structure is at least as important as the learning algorithm in modeling the spread of the merger.

We model the change in a non-uniform social network of 100 centralized clusters of 75 individuals each. 75 was chosen as half Dunbar's number, the maximum number of reliable social connections that an adult can maintain (Dunbar, 2010). There are two grammars, g_+ and g_-, and learners internalize one or the other according to the 17% threshold of M_+ in their input. One cluster represents the source of the merger and is initialized at 100% g_+, while the rest begin 100% g_-. Inter-cluster connections are chosen randomly so that some connections are between central members of the clusters and some are between peripheral members. The one merged cluster is connected to half the other clusters representing those at the frontier of the change, and each other cluster is connected to five randomly chosen ones.[4] This network structure echoes work in sociolinguistics, in particular, Milroy and Milroy (1985)'s notion of *strong* and *weak* connections in language change, where weak connections between social clusters are particularly important for propagation of a change.

Propagation of the merged grammar is calculated by $\mathcal{E}_n$ because we are interested in the behavior of individuals without loss of precision and because it cannot be assumed that each cluster is internally uniform.[5] Since the spread of the merger has been rapid enough to detect over a period of a few years, iterations are modeled as short age co-

[3] *Don* (1052) – *Dawn* (736); *collar* (403) – *caller* (23); *knotty* (25) – *naughty* (195); *odd* (830) – *awed* (80); *Otto* (67) – *auto* (260); *tot* (9) – *taught* (1327); *cot* (39) – *caught* (2444); *pond* (258) – *pawned* (31); *hock* (25) – *hawk* (127); *nod* (180) – *gnawed* (53); *sod* (30) – *sawed* (37)

[4] Originally, the clusters were set up as a "stepping-stone" chain with the merged community at one end, and that produced a similar S-curve. The structure presented here is more geographically plausible but not crucial for the results.

[5] $\alpha = 0.45$.

horts rather than full generations in the first experiments by updating only a randomly chosen 10% of nodes at each iteration because only a fraction of the population is learning at any given time. A model where every node is updated is investigated as well.

4.3 Results

The behavior of this simulation is shown graphically in Figure 4. The fine/colored lines indicate the rate of M_+ within each initially non-merged cluster, and the bold/black line shows the average rate across all initially non-merged. The merger spreads from cluster to cluster in succession over the "weak" inter-cluster connections and through each cluster over the 'strong' connections before moving on to the next ones.

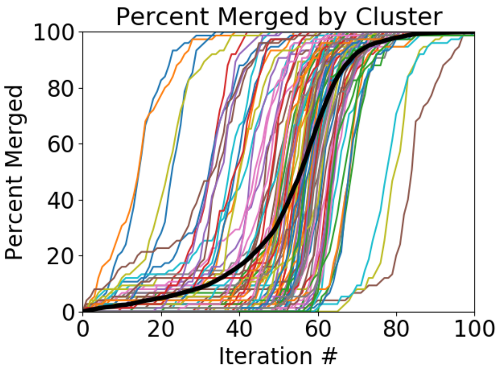

Figure 4: Spread of merger across communities (fine/colored) and population average (bold/black)

Most individual clusters exhibit a period of time in which only a few *early adopter* (Rogers Everett, 1995) members have the merger, a period of rapid diffusion of the merger, then some time where a few *laggards* resist the merger. As a result, most clusters exhibit an S-like shape. A few clusters change rapidly because of their especially well-connected positions in the network, and some lag behind the rest because they are poorly connected to the rest of the network. More interestingly, the population-wide average, the population-level data at the kind of granularity that is often studied, yields a smooth S-curve with a shallower slope than the individual clusters. The fact that it arises naturally here in a network that conforms with typical network shapes but was otherwise randomly generated is encouraging because the experiment was not set up so that it would produce such a curve, and the steep rate of change in individual clusters is what is expected for a change that is rapid enough to affect siblings differently.

In the above simulation, only a fraction of nodes were updated at each iteration in order to model a rapid change. In order to confirm that this choice is not affecting the results and to test a purer implementation of the framework presented here, we remove that constraint and update every node at each iteration. Figure 5 shows what happens over 20 iterations in a network that is otherwise identical but with 2/5 as many inter-cluster connections as the original. A qualitatively similar pattern arises, so the choice to update only a fraction of the population is not crucially affecting the results.

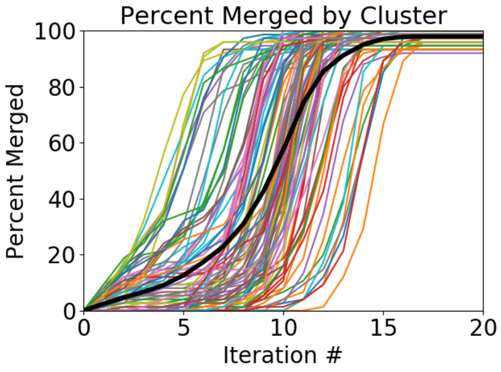

Figure 5: Spread of merger across communities (fine/colored) and population average (bold/black)

In all experiments so far, social connections were fixed at the first iteration even though connections in real populations tend to change over time. To investigate that modeling assumption, we perform another simulation in which connections are randomly updated both within and across clusters at each iteration akin to Kauhanen (2016)'s rewiring. The result as shown in Figure 6 is similar to before, with one major difference. The individual clusters transition more closely in time because no individual cluster remains poorly connected or especially well connected throughout the entire simulation.

Finally, we test our assumptions about population size by repeating the experiments on a smaller network of 40 clusters of 18 individuals. The results are qualitatively similar, but the S-curve appears to be more sensitive to probabilistic connections in the network. To explore this, we present the average network-wide rate of (M_+) across 10 trials, revealing that an S-like curve is formed each time but that its slope varies. A few trials never

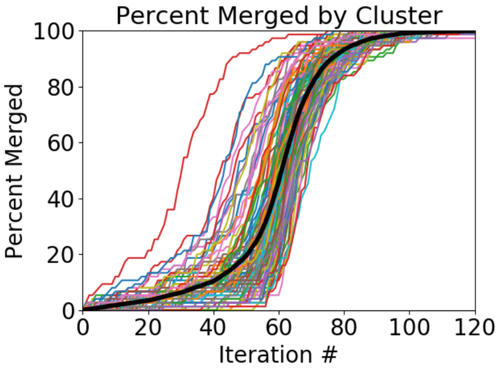

Figure 6: Spread of merger within communities (fine/colored) and as population average (bold/black). Network updated.

reach 100% because some of the clusters are not connected to the innovative one. The slope varies between trials, indicating that the rate of change is a function of both the population structure and the learning algorithm, but the network size does not substantially affect these results.

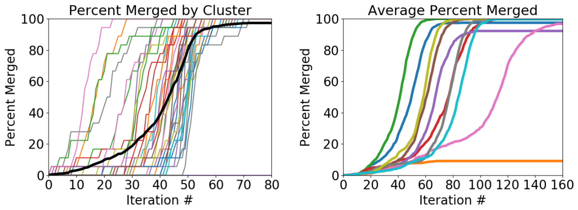

Figure 7: Single small network trial (left); average curves from 10 trials (right)

5 Discussion

The algebraic-network framework for modeling population-level language change presented here has substantial practical and theoretical advantages over previous ones. It is much simpler computationally than previous frameworks because it calculates the statistically expected behavior of each generation analytically and therefore removes the entire inner loop of calculating stochastic inter-agent interactions from the simulation. It follows the Niyogi and Berwick (1996) formalism for language change which presents a clean and modular way of reasoning about the problem and promotes the centrality of language acquisition.

In addition to the core algorithm, the framework offers enough flexibility to represent a wide variety of processes from the highly abstract (e.g., Kauhanen (2016)) to those grounded in sociolinguistic and acquisition research (e.g., Yang (2009)). In our investigation of Kauhanen's basic assumptions, we discover how seemingly innocuous decisions about population size and learning conspire to drive simulation results. If learners are conceived as categorical learners, population size becomes a deciding factor in the path of change. So while the original results are interesting and meaningful, they may only valid for small (on the order of 10^2) populations.

In our simulation of the spread of the COT-CAUGHT merger, we show how a cognitively-motivated model of acquisition requires a network model in order to represent population-level language change. The population is represented as a collection of individual clusters based on sociological work, but the clusters themselves are connected randomly. The fact that S-curves arise naturally from these networks underscores their centrality to language change.

One problem that this line of simulation work has always faced has been the lack of viable comparison between models because every study implements its own learning, network, and interaction models. The modular nature of our framework advances against this trend since it is now possible to hold the population model constant while slotting in various learning models to test them against one another and vice-versa. Finally, since this framework reduces to Niyogi & Berwick's models in perfectly-mixed populations, it can be used to reason about the formal dynamics of language change as well.

Without simulation, it would be difficult or impossible to undercover the interplay between acquisition and social structure on the propagation of language change. Neither factor alone can account for the theoretical or empirically observed patterns. Simulations of this kind which explicitly model both simultaneously is well equipped to provide insights that fieldwork and laboratory work cannot. As such, it is an invaluable complement to those more traditional methodologies.

Acknowledgments

We thank Charles Yang for is input and audiences at FWAV 4 and DiGS 19 for comments on earlier versions of this work. This research was funded by an NDSEG fellowship awarded to the first author by the US Dept. of Defense.

References

Gareth J Baxter, Richard A Blythe, William Croft, and Alan J McKane. 2006. Utterance selection model of language change. *Physical Review E*, 73(4):046118.

Gareth J Baxter, Richard A Blythe, William Croft, and Alan J McKane. 2009. Modeling language change: an evaluation of trudgill's theory of the emergence of new zealand english. *Language Variation and Change*, 21(02):257–296.

Christian Biemann, Stefan Bordag, Gerhard Heyer, Uwe Quasthoff, and Christian Wolff. 2004. Language-independent methods for compiling monolingual lexical data. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 217–228. Springer.

Leonard Bloomfield. 1933. *Language history: from Language (1933 ed.).* Holt, Rinehart and Winston.

Richard A Blythe and William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language*, 88(2):269–304.

Robert R Bush and Frederick Mosteller. 1953. A mathematical model for simple learning. In *Selected Papers of Frederick Mosteller*, pages 221–234. Springer.

Kathryn Campbell-Kibler. 2009. The nature of sociolinguistic perception. *Language Variation and Change*, 21(1):135–156.

Alfonso Caramazza, Albert Costa, Michele Miozzo, and Yanchao Bi. 2001. The specific-word frequency effect: implications for the representation of homophones in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(6):1430.

Luigi Luca Cavalli-Sforza and Marcus W Feldman. 1981. *Cultural transmission and evolution: a quantitative approach.* 16. Princeton University Press.

Robin Dunbar. 2010. *How many friends does one person need?: Dunbar's number and other evolutionary quirks.* Faber & Faber.

Zsuzsanna Fagyal, Samarth Swarup, Anna María Escobar, Les Gasser, and Kiran Lakkaraju. 2010. Centers and peripheries: Network roles in language change. *Lingua*, 120(8):2061–2079.

Edward Gibson and Kenneth Wexler. 1994. Triggers. *Linguistic inquiry*, 25(3):407–454.

K David Harrison, Mark Dras, Berk Kapicioglu, et al. 2002. Agent-based modeling of the evolution of vowel harmony. In *PROCEEDINGS-NELS*, 32; VOL 1, pages 217–236.

Daniel E Johnson. 2007. Stability and change along a dialect boundary: The low vowel mergers of southeastern new england. *University of Pennsylvania Working Papers in Linguistics*, 13(1):7.

Keith Johnson, Elizabeth A Strand, and Mariapaola D'Imperio. 1999. Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics*, 27(4):359–384.

Henri Kauhanen. 2016. Neutral change. *Journal of Linguistics*, pages 1–32.

Jinyun Ke, Tao Gong, and William SY Wang. 2008. Language change and social networks. *Communications in Computational Physics*, 3(4):935–949.

ed. Knight Chris Kirby, Simon, Michael Studdert-Kennedy, and James Hurford. 2000. *The evolutionary emergence of language: social function and the origins of linguistic form.* Cambridge University Press.

Sheldon Klein. 1966. Historical change in language using monte carlo techniques. *Mechanical Translation and Computational Linguistics*, 9(3):67–81.

Anthony S Kroch. 1989. Reflexes of grammar in patterns of language change. *Language variation and change*, 1(03):199–244.

William Labov. 1994. Principles of language change: Internal factors.

William Labov. 2001. Principles of language change: Social factors. *Malden, MA: Blackwell.*

William Labov, Sharon Ash, and Charles Boberg. 2005. *The atlas of North American English: Phonetics, phonology and sound change.* Walter de Gruyter.

David W Lightfoot. 1979. Principles of diachronic syntax. *Cambridge Studies in Linguistics London*, 23.

James Milroy and Lesley Milroy. 1985. Linguistic change, social network and speaker innovation. *Journal of linguistics*, 21(02):339–384.

James W Minett and William SY Wang. 2008. Modelling endangered languages: The effects of bilingualism and social structure. *Lingua*, 118(1):19–45.

Partha Niyogi. 1998. The logical problem of language change. In *The Informational Complexity of Learning*, pages 173–205. Springer.

Partha Niyogi and Robert C Berwick. 1996. A language learning model for finite parameter spaces. *Cognition*, 61(1):161–193.

Partha Niyogi and Robert C Berwick. 1997. A dynamical systems model for language change. *Complex Systems*, 11(3):161–204.

Partha Niyogi and Robert C Berwick. 2009. The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences*, 106(25):10124–10129.

Gordon E Peterson and Harold L Barney. 1952. Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, 24(2):175–184.

Nikolaus Ritt. 2004. *Selfish sounds and linguistic evolution: A Darwinian approach to language change*. Cambridge University Press.

M Rogers Everett. 1995. Diffusion of innovations. *New York*, 12.

Teresa Satterfield. 2001. Toward a sociogenetic solution: Examining language formation processes through swarm modeling. *Social Science Computer Review*, 19(3):281–295.

Michael T Schaub, Neave O'Clery, Yazan N Billeh, Jean-Charles Delvenne, Renaud Lambiotte, and Mauricio Barahona. 2016. Graph partitions and cluster synchronization in networks of oscillators. *Chaos: An Interdisciplinary Journal of Nonlinear Science*.

James N Stanford and Laurence A Kenny. 2013. Revisiting transmission and diffusion: An agent-based model of vowel chain shifts across large communities. *Language Variation and Change*, 25(2):119.

Henry Sweet. 1899. *The practical study of languages*. London: Dent.

Charles Yang. 2009. Population structure and language change. *Ms., University of Pennsylvania*.

Charles D Yang. 2000. Internal and external forces in language change. *Language variation and change*, 12(03):231–250.