A graph theory model for second generation immigrant language proficiency

Andrew Wang

March 2025

1 Introduction

In the world of globalization, immigration is at an all time high. Countries are becoming increasingly diverse, and immigrants seem to be able to assimilate relatively well in many cases, often times building a life and starting a family. Although first generation immigrants typically retain fluency in their heritage language, this does not always hold true for their children. In fact, it's very common and well established that second generation immigrants encounter challenges in learning their heritage language. Often times as a child's proficiency in the host language grows, it will squash their proficiency and improvement in the heritage language. Regardless, even with a limited improvement in the heritage language, their proficiency is usually not totally stagnant [4].

This report contains my attempt to quantify language proficiency of second generation immigrants. In my approach to mathematically model this phenomena, I'll rely heavily on my own personal experience to motivate behavior and assumptions. Furthermore, I'm interested in the early stages of an individuals life, and whether or not they reach conversational fluency rather than absolute fluency (this might entail understanding of technical and formal language or slang).

Although I didn't find many attempts to make a mathematical model for language proficiency, there are models that take a graph theory approach to model culture dynamics [2]. This paper's topic is slightly different, but they do share the similarity of leveraging graphs. Additionally, their use of a game theory utility function to determine edge weight provided me with inspiration for the function I'll use to determine an individuals family interactions.

2 Model Setup

I'll be using a graph for my model, in particular, an Erdős–Rényi[3] graph:

$$G(n, p)$$

where:

- $n =$ Number of nodes (agents)
- $p =$ Probability of a connection between any two nodes
- Nodes are assigned a random position in a 2D space:

$$(x, y) \in [0, 1] \times [0, 1]$$

- $\forall n \in V(G)$ represents a person
- $\forall e \in E(G)$ represents a social interaction
- Nodes are connected if their Euclidean distance is below a threshold d:

$$\|n_i - n_j\| < d$$

  - These edges represent proximity interactions

This is not the typical choice to model social networks. More often, researchers will choose a network where the vast majority of nodes have a relatively low number of neighbors, and a select few nodes have a very high number of neighbors [1]. This structure works very well at modeling phenomena such as spread of information or disease, but is not well suited for my applications. In my model, I do not want the presence of these highly connected nodes, as I am more interested in average case behavior. Thus, I chose the Erdős–Rényi graph for its simplicity and efficacy in modeling my topic.

2.1 Proficiency Variables

Let:

- $P(t) =$ Heritage proficiency of an agent at time t
- $\bar{P}(t) =$ Host proficiency of an agent at time t

Proficiency values are bounded in the range $[0, 1]$, where:

- $P(t) = 0 \implies$ No ability to communicate in the heritage language.
- $P(t) = 1 \implies$ Full conversational fluency in the heritage language.

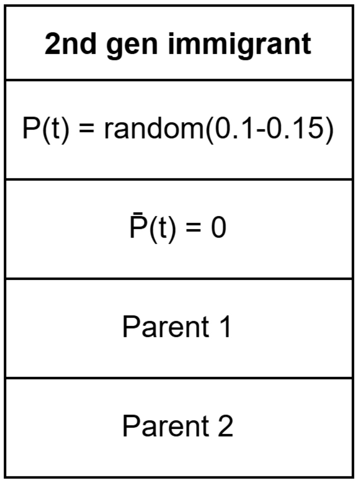
2.2 Node Types

The model categorizes people into three distinct groups:

- Host
- Second generation immigrant
  - Note that these nodes are initialized with a P(t) ¿ 0, which represents a propensity for their heritage language as a result of family influence

    - Assigned two parent objects:
        - Parent 1: Always first-gen
        - Parent 2: 50% chance first-gen, 50% chance host.

- First generation immigrant
    - Not represented explicitly as a node in the graph
    - Stored as pointers of second generation nodes



2nd gen immigrant

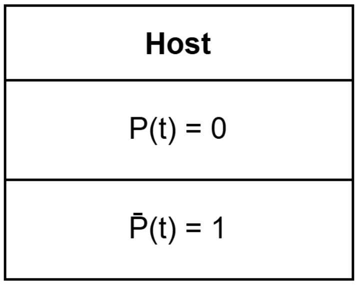
P(t) = random(0.1-0.15)

P̄(t) = 0

Parent 1

Parent 2

Figure 1: 2nd gen node

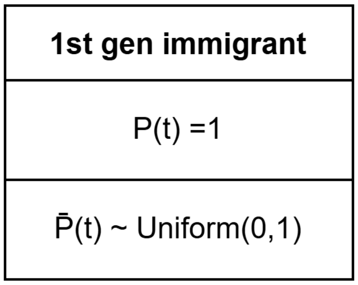


Host

P(t) = 0

P̄(t) = 1

Figure 2: Host node



1st gen immigrant

P(t) =1

P̄(t) ~ Uniform(0,1)

Figure 3: 1st gen object

3 Proficiency Update Logic

3.1 Proficiency Update Algorithm

```
for each time step $t$ do
    for each second gen node $n$ in $G$ do
        Update $\bar{P}(t)$ through proximity and social interactions
        Update $P(t)$ and $\bar{P}(t)$ through family interactions
    end for
end for
```

3.2 Social and Proximity Update

The host proficiency of a second generation node will first be updated by social and proximity interactions by the following rule:

$$\text{host interaction weight} = \frac{\sum_{j \in N_{\text{phys}}(n)} \frac{1}{\text{dist}(n,j)} + W_{\max} \cdot |N_{\text{soc}}(n)|}{W_{\max} \cdot |N_{\text{soc}}(n)| + |N_{\text{phys}}(n)|}$$

$$\text{where: } W_{\max} = \max \sum_{j \in N_{\text{phys}}(n)} \frac{1}{\text{dist}(i,j)}$$

$W_{\max}$ is assumed to be the weight of each social interaction (since social interactions are likely at least as strong as the strongest proximity interaction).

The host proficiency update is:

$$\bar{P}(t) = \bar{P}(t) + \alpha \cdot \text{host interaction weight}$$

3.3 Family Interaction Update

The update rule for family interactions is a game theory inspired function, where the goal is for a family to maximize communication, while minimizing effort

1. **Parental Bias Toward Heritage:**

   let $P_1, \bar{P}_1$ = heritage and host proficiency for parent 1 let $P_2, \bar{P}_2$ = heritage and host proficiency for parent 2

   $$f_1 = \frac{P_1}{P_1 + \bar{P}_1}, \quad f_2 = \frac{P_2}{P_2 + \bar{P}_2}$$

   Combined parental bias:

   $$f_P = \frac{f_1 + f_2}{2}$$

2. **Child's Bias Toward Heritage:**

   $$f_A = \frac{P(t)}{P(t) + \bar{P}(t)}$$

3. **Time Spent Speaking Heritage Language:**

   $$T = f_P \cdot f_A$$

4. **Time Spent Speaking Host Language:**

   $$\bar{T} = 1 - T$$

5. **Proficiency Update:**

   $$P(t+1) = \sigma(P(t) + \Delta_t \cdot T)$$
   $$\bar{P}(t+1) = \sigma(\bar{P}(t) + \Delta_t \cdot \bar{T})$$

   $$\text{where } \sigma(x) = \frac{1}{1+e^{-k(x-x_0)}}$$

4 Simulation Parameters

Parameter	Value
Number of agents	1000
Fraction host/second-gen	80% / 20%
Edge probability in ER graph	0.02
Physical threshold	0.1
Number of time steps	20

5 Results

5.1 Population-Level Outcomes

An initial analysis of the simulation shows that second generation nodes reach a relatively high heritage proficiency, and almost always become fluent in the host language. The proficiency distributions serve as a sanity check for expected behavior — second-generation nodes achieve full host fluency, and a significant amount nodes attain heritage fluency.

- Average heritage proficiency (2nd gen) = 0.647
- Average host proficiency (2nd gen) = 0.993

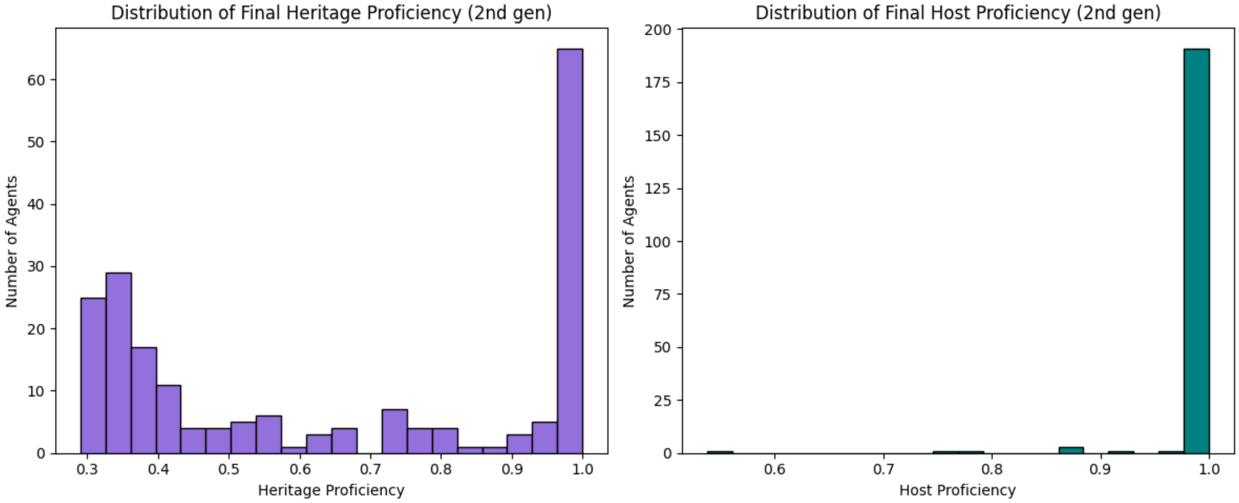


Figure 4: Proficiency Distributions (2nd gen nodes)

Nodes that achieved full fluency had:

- Average heritage proficiency of parents = 1.000
- Average host proficiency of parents = 0.324

The nodes that didn't achieve full fluency in the heritage language had:

- Average heritage proficiency of parents = 0.656
- Average host proficiency of parents = 0.727

These averages provide us with several insights

1. Fully fluent nodes always have parents fully fluent in the heritage language, who also tend to have poor proficiency in the host language
2. Non fully fluent nodes still have parents that are fairly fluent in the heritage language, however they have a much higher host proficiency

We can isolate average heritage proficiency of the second generation nodes by the number of number of first generation parents they have

- 2 first gen parents $\Rightarrow \ P(T) = 0.850$
- 1 first gen parent $\Rightarrow \bar{P(t)} = 0.357$

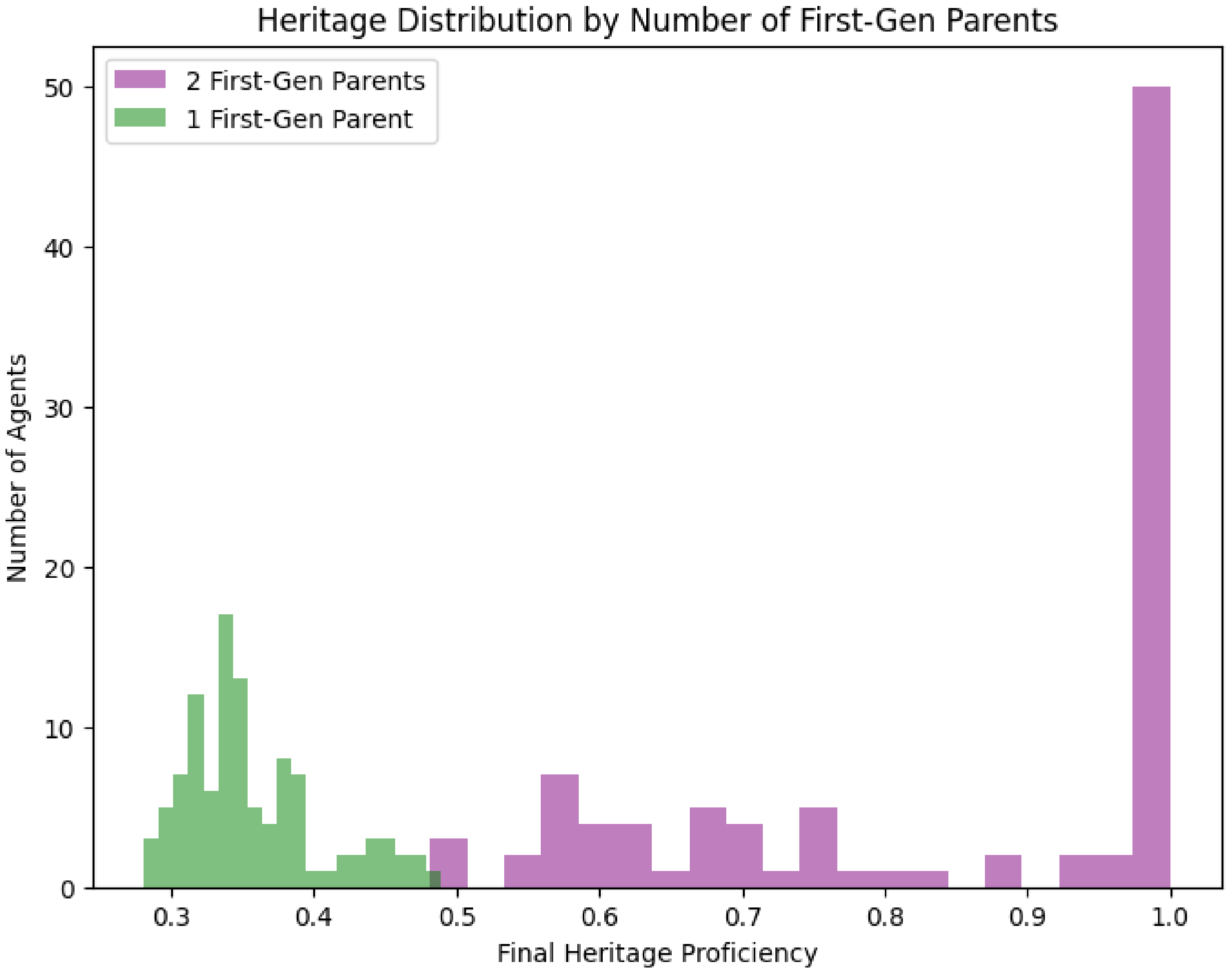


Figure 5: Proficiency Distribution Categorized by Parent Status

According to this model, number of parents is a huge predictor of heritage proficiency, but even among nodes with two first generation parents, many nodes don't achieve full fluency

We can further restrict our analysis to only nodes with two first generation parents and, categorize them by fluent vs non fluent.

First we'll look at the parent host proficiencies of each group (no need to analyze heritage proficiency since all these parents are 1st generation):

- Fully fluent nodes ⇒ parent host proficiency = 0.32
- Non fully fluent nodes ⇒ parent host proficiency = 0.69

Now we'll plot their average proficiency curves of each category over time.

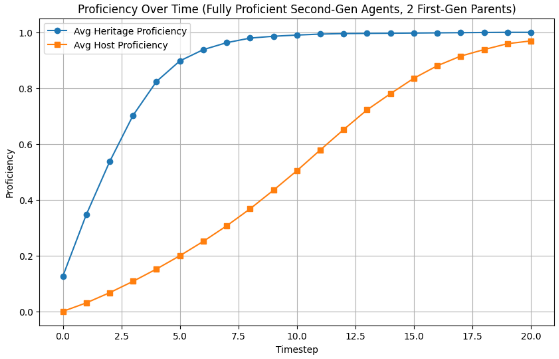


Figure 6: Average proficiency curve of fluent nodes

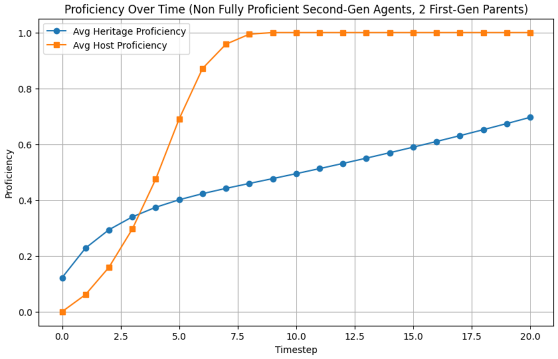


Figure 7: Average proficiency curve of non fluent nodes

Fluent nodes have a consistent, steady improvement of their heritage over time. We know they have parents fluent in the heritage language, with poor proficiency in the host language. Thus, these nodes likely receive regular exposure to the heritage language at home, resulting in the node reaching full fluency. The graph also shows delayed fluency in the host language. This result is congruent with my model design and may be true to some degree in real world settings, but it's unrealistic that host proficiency is delayed by such a large margin. A more likely scenario would be a slight delay at most in the host proficiency, which would rapidly improve upon entering the education system.

Non fluent nodes start with a higher heritage proficiency, but because their parents have a much higher host proficiency, these nodes will converse at home primarily in the host language. Their exposure to the heritage language is limited, and thus they don't reach fluency.

6 Conclusion

The analysis of my simulation provides a few valuable generalizations about language proficiency of second generation immigrants.

1. Families with two parents of the same heritage are much more likely to pass on fluency
2. Second generation immigrants are much more likely to obtain fluency in the heritage language if they have some motivation to speak the language at home
   (a) We observed that second-generation nodes reaching fluency depended heavily on their parents' host proficiency. However, this is likely an oversimplification, as there are many cases where children become fluent in their heritage language even when their parents are fluent in the host language. Rather than claiming definitively that parent fluency in the host language determines heritage language proficiency, we can generalize this result and suggest that fluency in the heritage language depends on a strong incentive to use the heritage language over the host language.

It's worth nothing that the results of this model are at risk of a logical fallacy. The assumptions that motivated the decisions and structure of the model are more or less synonymous with the results. However, I don't think that this admission makes the model less meaningful to me. To me, this model was a quantitative expression of my qualitative experiences. It was not designed with the intention of offering scientific value, but rather a way for myself and others to better understand their experience with language.

7 Future Work and Improvements

7.1 Parameter Testing

This is likely the most practical next step; to try different permutations of parameters and compare the results to some ground truth data in order to determine the accuracy of the model, which can then be used predictively.

7.2 Game Theory Function

Another interesting direction would be to experiment with different functions for determining how much the heritage language is spoken at home. The current function probably does a fair job of achieving the desired behavior, but further experimentation seems to be a valuable improvement.

7.3 Node ”Strategies”

One of the limitations of this model was to assume behavior was uniform across nodes. Introducing nodes that follow certain rules or behaviors would provide insights into how specific scenarios affect language proficiency.

References

[1] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[2] Yao-Li Chuang, Tom Chou, and Maria R. D'Orsogna. A network model of immigration: Enclave formation vs. cultural integration. *Networks amp; Heterogeneous Media*, 14(1):53–77, 2019.

[3] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.

[4] Mila Schwartz. Exploring the relationship between family language policy and heritage language knowledge among second generation russian–jewish immigrants in israel. *Journal of Multilingual and Multicultural Development*, 29(5):400–418, 2008.