Creating Abundance: Immigrants and the Birth of American Entrepreneurship

Costas Arkolakis* Michael Peters† Sun Kyoung Lee‡

April 27, 2017

This proposal is a major endeavor to assess the contribution of immigrants for the process of American economic growth and technological progress since the mid 19th century. A focal point of our analysis will be the role of immigrants as entrepreneurs. We plan to merge existing databases with novel sources to create an unprecedented historical database on the regional allocation of immigrants, their occupational employment patterns and the spatial distribution of economic activity across US counties and cities.

While there is an extensive literature on the short-run consequences of immigration, the long-run effects of immigration have so far received less attention. This project aims to complement the literature by analyzing the long-run impacts of one the biggest migration episodes in modern time – the era of Mass Migration (1850-1913). We focus on the role of businesses created by immigrants for the transmission of knowledge and aggreagte economic growth. In particular, we aim to analyze whether immigrants from the European continent brought novel ideas and technologies to the US and then used such ideas to start their own companies. The answer to this question is not only of historical interests but we also hope to shed light on the extent to which technologies and knowledge are spatially mobile and wheter entrepreneurs play an important role in the diffusion process. To measure the relationship between immigrants' knowledge, the extent of entrepreneurship and local productivity, we propose a big data approach by digitizing and linking three types of historical data sources:

1. Immigration Database: The Hamburg Passenger List (1850-1934) & Castle Garden Database (1820-1892)

2. The US complete-count demographic census data (1850 - 1940, every decade except for 1890)

3. The Historical Census of Manufacturers (1860-1940, every decade)

*Yale University and NBER
†Yale University and NBER
‡Columbia University

Our starting point are the official immigration records from the Hamburg and Castle Garden Passenger Passenger Lists. These contain detailed information on millions of immigrants, who arrived in the US in the 19th and early 20th century. Importantly, these records contain information on the immigrants' occupations and sector of employment *prior* to their arrival in the US and their previous residence in Europe. We plan to use this information, together with existing information on spatial economic development in Europe, to construct empirical measures of the quality of immigrants' existing knowledge. Intuitively, immigrants with work experience in a particular industry stemming from a region in Europe, who specialized in that industry, were arguably more exposed to the existing frontier technology in that particular sector relative to other immigrants.

Using the information on immigrants' names and demographics, we then match these records to the US complete count census using machine learning techniques. Because the US complete count census contains detailed micro information on the place of residence and economic characteristics, we can directly measure immigrants' mobility patterns - both spatially and socio-economically. In particular, we can measure to what extent immigrants were able to start their own businesses in their pre-migration industries and hence were able to employ their existing knowledge productively. We have already completed part of the micro-data matching of individuals across different datasets with very satisfactory results. Currently we are in the process of matching immigrants in the US census to the immigration records to get good information on their occupation before and after the migration to the United States.

To finally measure the resulting productivity effects, we merge this data to newly digitized data from the Historical Census of Manufacturers at the county and city level. The published sources from the Census of Manufacturers contain detailed information on employment, value added and the number of plants for a variety of industries for all 3000 counties of the US. By harmonizing the industrial classification from the Census of Manufacturers and the US complete count census, we can study whether the inflow of skilled immigrants in the 19th century led to economic prosperity at the local level. Intuitively: if knowledge is spatially mobile in that it can be employed easily in different locations, one would expect regions who received more immigrants with prior experience in a particular industry to experience growth in that particular industry.

The episode of Mass Migration is uniquely suited for this analysis. On the one hand, it is an episode where detailed micro data, such as the Hamburg Passenger List (1850-1934) is, in principle, available. On the other hand, it is a point in time where information flows were more restricted than today and information about economically promising techniques were tied to the mobility of people. Finally, the historical environment suggests various empirical strategies to generate exogenous variation in the spatial allocation of immigrants. Not only can we exploit pre-existing settlement patterns of immigrants that arrived before 1850, but we will also explore particular events like the passing of the 1917 Immigration Act or the Irish famine of 1879 as aggregate immigration supply shifters. Hence, this setting is a rare opportunity to empirically measure the spatial flow of ideas and to quantify the economic importance of the transmission of knowledge for economic growth.

Budget Justification

Translation Costs: Support is requested to translate records written in foreign languages to English as some immigration records were recorded in foreign languages. For example, The Hamburg Passenger List records are in German, and information such as immigrants' occupations need to be translated into English. Additionally, we need to index various German county-level historical statistical tables, as information such as industrial classifications and regional information needs to be translated from German to English. The coauthors will hire undergraduate students who are proficient in both languages at an hourly rate of $20. Translating data for both the Hamburg Passenger Records and the German Census of Manufacturers will take about 250 hours. We therefore expect a total cost of $5,000 ($20/hr * 250 hours) for translation services.

Data Transcription: The coauthors are in the process of converting published statistics tables from published booklets into machine-readable forms. As these tables were mostly printed materials in the nineteenth and early twentieth century, running the images through Optimal Character Recognition software for data conversion is not optimal (as the outputs can be highly inaccurate). Thus, the coauthors have been transcribing these contents into machine-readable database by working with a professional data transcription service provider. The data transcription vendor charges approximately $3 per page, and each decadal records have on average 400 pages. Therefore, approximately $1,200 budget is expected for data transcription per decade and given the whole span of data starting from 1850 to 1940, the total cost of $12,000 is requested for the entire transcription of US Census of Manufacturers records (relevant tables only).

Additionally, the coauthors need to transcribe historical economic census of immigrants' countries of origin to measure the extent of economic development prior to their arrival in the US. We plan to start with the historical German Census of Manufacturers. Given the overall volume of materials for manual transcription, the total cost is $2,000 for the data transcription.

Professional Services of Map Georeferencing and Shapefile Creation: The coauthors plan to digitize and georeference historical maps to reflect the cities' boundary change. Although the National Historic Geographic Information System (hereafter, NHGIS) provides excellent resources for identifying county and state-level boundaries, typically this does *not* cover historical *city boundaries*. This posits a major barrier, as most of Census of Manufacturers records were calculated and published at the city-level since the 1880s. In order to make use of this rich city-based data, the coauthors need to create city-boundary shapefiles for 200 major cities in the United States from the nineteenth century to the first half of twentieth century. The map georeferencing company charges $8.50 per image for georeferencing. Given the total number of city boundary images is about 2000 (200 cities, and 10 images per city as there are 10 decadal city-boundary maps per city), this amounts to $17,000.

Once the relevant maps are geo-referenced, attributing polygons and boundaries in shapefile formats are needed. Given that the service vendor charges an hourly rate of $50 for the service and that it typically takes about one hour to generate five images, we expect to spend about $20,000 to

process and attribute the 2000 images ($50/hour * 400hours).

Overall, the total estimated budget is therefore $56,000. The coauthors plan to use the NBER Entrepreneurship grants (if approved) primarily for data transcription. The remaining funds will be used for creating the city-boundary files.