Analyzing Cities' Complex Socioeconomic Networks Using Computational Science and Machine Learning

by

Ahmad Alabdulkareem

M.S. in Computer Science, King Abdullah University for Science & Technology (KAUST), 2011
B.S. in Computer Science, King Saud University, 2010
Submitted to the Center for Computational Engineering and to the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational Science & Engineering at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

© Massachusetts Institute of Technology 2018. All rights reserved.

Signature redacted

Center for Computational Engineering

Signature redacted

Alex "Sandy" Pentland
Toshiba Professor of Media Arts and Sciences
Thesis Supervisor

Signature redacted

John R. Williams
Professor of Information Engineering
Thesis Supervisor

Signature redacted

Jesse Kroll
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Signature redacted

Nicolas Hadjicostis
Co-Director, Computational Science and Engineering
Analyzing Cities’ Complex Socioeconomic Networks Using Computational Science and Machine Learning

by

Ahmad Alabdulkareem

Submitted to the Center for Computational Engineering and to the Department of Civil and Environmental Engineering on May 18, 2018, in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational Science & Engineering

Abstract

By 2050, it is expected that 66% of the world population will be living in cities. The urban growth explosion in recent decades has raised many questions concerning the evolutionary advantages of urbanism, with several theories delving into the multitude of benefits of such efficient systems. This thesis focuses on one important aspect of cities: their social dimension, and in particular, the social aspect of their complex socioeconomic fabric (e.g. labor markets and social networks). Economic inequality is one of the greatest challenges facing society today, in tandem with the eminent impact of automation, which can exacerbate this issue. The social dimension plays a significant role in both, with many hypothesizing that social skills will be the last bastion of differentiation between humans and machines, and thus, jobs will become mostly dominated by social skills. Using data-driven tools from network science, machine learning, and computational science, the first question I aim to answer is the following: what role do social skills play in today’s labor markets on both a micro and macro scale (e.g. individuals and cities)? Second, how could the effects of automation lead to various labor dynamics, and what role would social skills play in combating those effects? Specifically, what are social skills’ relation to career mobility? Which would inform strategies to mitigate the negative effects of automation and off-shoring on employment. Third, given the importance of the social dimension in cities, what theoretical model can explain such results, and what are its consequences? Finally, given the vulnerabilities for invading individuals’ privacy, as demonstrated in previous chapters, how does highlighting those results affect people’s interest in privacy preservation, and what are some possible solutions to combat this issue?

Thesis Supervisor: Alex “Sandy” Pentland
Title: Toshiba Professor of Media Arts and Sciences
Thesis Supervisor: John R. Williams
Title: Professor of Information Engineering
Acknowledgments

This is without a doubt the hardest part to write in my entire thesis, and I am fighting the urge to make it longer than the thesis itself, because I know I cannot do justice to everyone who supported and encouraged me. Therefore, I will force myself to restrict this to two pages.

I am grateful to all of the amazing members of my committee. I was extremely lucky to have such an impressive set of professors guiding me. Starting with the person who had the biggest influence on my PhD, I want to thank someone who was more than just an advisor: my friend, Professor Alex “Sandy” Pentland. He not only provided me feedback and assistance in my research, but has also made sure to help me in my personal life and future career. This brings me to the person who introduced me to Sandy many years ago, Professor John R. Williams, who was my guide in both life and academia alike. I still fondly remember our long talks when I first came to Boston almost a decade ago. I was also fortunate to work with Professor Iyad Rahwan, who was kind enough to welcome me into his group. He never asked what was in it for him, and he greatly helped me in my research when the time came. Finally, I want to thank Professor Marta Gonzalez, the ace in my committee. Whenever I ran into a roadblock, Marta was always there to walk me through it, always asking thought-provoking questions to help me advance my work.

This brings me to Professor Anas Alfaris, who I would call my father figure if he did not hate that for making him sound old, since he is not that much older than me. As such, I will stick to calling him my very close friend. Anas was my first real mentor. He taught me the skills I needed to succeed in life, and then some. I can, without any doubt in my mind, say that I would not be where I am today without him.

I appreciate all of the various groups’ very supportive staff, who were always extremely helpful, especially Kate Nelson and Kiley Clapper. To all of my many colleagues and amazing friends at the Center for Complex Engineering Systems, thank you. To my fellowship sponsor, King Abdulaziz City for Science and Technology, I
really appreciate your generous support, especially from CCES's deputy director, Adnan Alsaati, and KACST's president, HH Dr. Turki Al Saud. I am truly fortunate to come from a country that supports education in every way possible.

I was blessed to have made many friends during my PhD. Unfortunately, I cannot list them all, though they are an integral part of who I am. I would like to send a hearty shout-out to all of the members of the Human Dynamics (especially Abdulrahman Alotaibi), Scalable Cooperation (especially Morgan Frank & Edmond Awad), Geonumerics (especially Mohamad Sindi), IDSS (especially Yuan Yuan), and most importantly, thank you to my many close friends at CCES, especially Abdulaziz Alhassan and Fahad Alhasoun, who, along with CCES's program manager, Katherine Paras, organized the best post defense celebration ever. In addition, there are two who deserve a separate acknowledgment: my dear friend Abdullah Alhajri, who was always there when I needed him throughout my PhD, and my lifelong friend and collaborator, Abdullah Almaatouq; thanks for always keeping me thinking.

It takes a village to raise a child; my mother, Badriah Alabdulkareem, was my whole village. She inspired me to become the man I am today, and she was even the reason I pursued a bachelor’s degree in CS, with her stories about her studies. This brings me to the person who made that happen: my father, Dr. Abdulmajeed Alabdulkareem. Words cannot capture the support, encouragement, and guidance that my dad gave me throughout my life. For example, as a surgeon, everyone in my family and friends convinced me to follow in his footsteps, but he was the one who taught me to pursue my passion; thanks, dad. This brings me to my beloved siblings. Each has their own place in my heart, and I have a special relationship with all of them. To my sister Yara and my brothers Feras, Muath, and Abdulrahman, thank you for always being there.

Finally, I reach the person who deserves to have their name listed along mine at the top: my wife, my love, and my soul mate, Sarah Alyahya. She endured all of my endless long hours working, she was always there when I needed anything, even during her own graduate studies, and she never complained. I could not have done it without you. Thank you for being a part of my life; it would not be the same without you.
# Contents

1 Introduction .............................................. 27
   1.1 Introduction ........................................ 27
   1.2 Thesis Outline .................................... 28

2 Methods .................................................. 29
   2.1 Introduction ........................................ 29
   2.2 Methods and Results ................................ 31
   2.3 Quantitative Comparison .............................. 35
   2.4 Conclusion .......................................... 40

3 The Importance of Social Skills in Labor Markets ............................... 41
   3.1 Introduction ........................................ 41
   3.2 Data Sets ............................................. 43
   3.3 The SkillScape: A Skill Complementarity Map ........ 45
   3.4 The SkillScape and Occupations ..................... 52
   3.5 Skill Polarization Using a Bottom-Up Approach ...... 56
   3.6 Moving into Labor Markets ............................ 60
   3.7 The SkillScape and Cities ............................. 62
   3.8 Social Skills and Cities’ Economic Well-Being .......... 63
   3.9 Conclusion .......................................... 65

4 Social Skills for Labor Mobility and Combating Automation .................. 67
   4.1 Introduction ........................................ 67
8.3 How Educational Requirements Relate to Skill Requirements for Occupations ........................................... 130
8.4 Human capital vs. social capital ........................................... 131
List of Figures

2-1 Illustration of the product space connecting 775 products based on their proximity matrix. The color of the node represents the product classification. This is my replication [63, 59, 62] based on 2013 data.  30

2-2 Complexity and Fitness methods’ convergence plots. Plot (A) shows all iterations of the complexity method, while plot (B) only shows the even iterations. Plot (C) however, is the z-score for each even iteration of complexity, making it more closely resemble fitness. Finally, plot (D) shows all Iterations of the fitness method.  34

2-3 Adjusted $R^2$ values (y-axis) from regressing GDP for various years over the successive iterations (x-axis) of the two different metrics (complexity and fitness).  36

2-4 The average treatment effect for the treated (the ATT statistical method) on GDP growth (for both the complexity and fitness treatments), after matching and controlling for covariates. The left results are without education, while the results on the right add the covariate of education. It should be noted that the fact that the right plot demonstrates non-significant results is mostly because the coarsened exact matching removed too many data points, see fig.2-6.  38

2-5 The coarsened exact matching scatter plot visual for the two-variate matching, to demonstrate which data points were kept (matched) vs. dropped (unmatched). We see a significant number of remaining data points for statistical analysis.  39
2-6 This figure demonstrates the data points kept by the coarsened exact matching method, with the left figure being identical to fig. 2-5, and the right plot representing the three-variable variant of the matching. Note the limited number of remaining (matched) data points in the three-variable matching.

3-1 (A) Heat-map visualization of the O*NET data matrix, which represents the relationships between occupations in the U.S. labor market and 161 skills “I_{j,s}”. Every row is an occupation, while every column is a skill. (B) Heat-map visualization of the Bureau of Labor Statistics data matrix used, representing the number of employees for different occupations within the various cities. Every row is a city, while every column is an occupation.

3-2 An occupation is identified through the skills of workers of that occupation. The bipartite network connecting occupations to required skills is a result of an underlying tripartite network containing workers as a conduit between occupations and skills. Relationships between skills are determined from their co-occurring importance across occupations.

3-3 The distribution of all co-occurrence proximities between skills. Insets represent counterpart networks from other related works of literature for comparison. Unlike previous applications of co-occurrence networks (insets), the SkillScape contains a bimodal distribution of pairwise skill complementarity.
3-4 We identify two polars of skills by applying Louvain community detection to the complete SkillScape network (i.e. no minimum $\theta$). We notice that this almost perfectly captures the right mode of the bimodal distribution. That is, the links that are within one of the two clusters (intra-polar edges) mostly belong to the strong mode of the distribution, while the weak links within the weak mode almost exclusively belong to edges that are between clusters (i.e. inter-polar edges). The visual on the right only contains $\theta > .6$ edges. The complete list of O*NET skills in each cluster is presented in the SI Appendix.

3-5 A filtered visual for the SkillScape. The SkillScape thresholded according to a minimum skill similarity (i.e. $\theta > 0.6$) visibly reveals two communities of complementary skills and respects expert-derived O*NET categories (colors). Node sizes reflect the PageRank values, while color indicates the O*NET categorization of the skill.

3-6 The SkillScape network respects experts’ skill categorization. For each O*NET skill category, we measure the distribution of $\theta$’s for pairs of skills within a category (blue) and compare it to the distribution of $\theta$’s for each edge connecting a skill within the category to a skill outside of the category (red). The complementarity for skills within a category is significantly stronger according to the KS statistic (title) than the complementarity for skills across categories.

3-7 Four different occupations, showing how they drastically differ in SkillScapes. Nodes of skills that an occupation requires ($e(j, s) = 1$) are maintained (colored in black), while skills that the occupation does not require are ignored (colored in white). It is clear that a job such as (A) nuclear technician requires mostly sensory-physical skills, in contrast to (D) nuclear engineer, for example, which requires mostly SocioCognitive skills.

3-8 Occupations requiring more SocioCognitive skills tend to make higher annual salaries.
3-9 Performed out-of-sample testing for each model in table 3.1 through 1,000 trails of randomly selecting 75% of the occupations as a training data set and measuring the root-mean squared error of the resulting model applied to the remaining 25% of occupations. We represent the resulting model performance as box plots. Medians are represented by a red line, while the mean error is represented by the triangles.

3-10 The two identified clusters/polars of skills using Louvain community detection from figure 3-4, superimposed on the SkillScape visual in figure 3-5. This demonstrates that our categorization of which skills belong to a “SocioCognitive” vs. “Sensory-Physical” cluster was agnostic to the skill labels/categorization from experts, and was reached in a purely data-driven bottom-up approach.

3-11 Reliance on SocioCognitive skills predicts increased annual wages. As a baseline, we consider the relative importance of routine labor using routine O*NET variables from [15]. In addition to cognitive skill fraction \( (SocioCognitive_j) \), we calculate the total skill content \( \sum_{s \in S} I(j, s) \) of each occupation. Each educational variable represents the total employment in that occupation whose highest educational degree is a high school diploma, bachelor’s degree, etc. All variables were standardized before regression. Standard errors are reported in parentheses and asterisks indicate the statistical significance of coefficient approximations. We perform out-of-sample testing for each model through 1,000 trails of randomly selecting 75% of the occupations as training data and measuring the root-mean squared error of the resulting model applied to the remaining 25% of occupations. We represent the resulting model performance as box plots. Median error is represented by a red line, while the mean error is represented by the triangles.
3-12 The labor market network connecting jobs based on their coexistence in cities, similar to the one presented by Shutters et al. Panel (A) The job market network, with node colors representing the industry cluster to which an occupation belongs, where red = Production Occupations; blue = Education, Training, and Library Occupations; green = Healthcare Practitioners and Technical Occupations; cyan = Life, Physical, and Social Science Occupations; purple = Office and Administrative Support Occupations; and gray = all other clusters. (B) provides a closer examination of a subpart of the graph and a few examples of jobs that strongly “coexist” together.

3-13 Heat-map visualizations of the data. (A) represents the number of employees for different occupations within the various cities. (B) represents the relationships between occupations in the U.S. labor market and 161 skills “I(j,s)”. (C) shows that by multiplying the BLS city/occupation data set with the occupation/skill data set from O*NET, we obtain a matrix that relates all cities to every skill “I(c,s)”.

3-14 The SkillScapes of different U.S. cities, showing an evolutionary path through different SkillScapes. Nodes of skills in which a city has significant presence, i.e. e(c,s) = 1, are maintained (colored in black), while other skills are ignored (colored in white).

3-15 Larger cities (i.e. in population and number of workers in the labor force) increasingly rely on SocioCognitive skills, leading to economic well-being (e.g. higher median household income and GDP). Example cities are projected onto the SkillScape using black nodes for effectively used skills.

4-1 Frey and Osborne’s [54] automation probabilities superimposed onto our SkillScape.
4-2 (A) Relationship between the predicted impact of automation on cities and the cities' SocioCognitive metric. (B) Relationship between the predicted probability of automation for occupations and their SocioCognitive fraction.

4-3 Skill proximity predicts worker transitions between occupations, skill redefinition of occupations, and skill acquisition in cities. (A) An example demonstrating SkillScape proximity (i.e. $proximity(j, s)$) as a proxy for the connections between effectively used skills and other skills. (B) Skills with high proximity to the effectively used skills of an urban labor market in 2010 are more likely to be effectively used by that workforce in 2015. (C) Skills with high proximity to the effectively used skills of an occupation in 2010 are more likely to be effectively used by that occupation in 2015. (D) The effectively used skills in a worker's occupation in 2015 are more likely to be effectively used in the worker's next occupation in 2016. We provide bar plots including 95% confidence intervals for these probabilities in SI figure 8-7, and we consider an alternative area under the receiver operating characteristic curve (AUROC) analysis in SI figure 8-14.

4-4 The polarized skill network constrains worker mobility. Binning by the SocioCognitive of the worker's occupation in 2014 reveals the (A) expected SocioCognitive change and (B) the expected magnitude of SocioCognitive change when workers change occupations. Random occupation selection is considered as a null model (gray). Standard error bars are provided, but are small. Actual occupation transitions are provided as examples in (A). (C) The national distribution of employment by SocioCognitive with the distribution of individual occupations as an inset.
4-5 The connectivity and embeddedness for each skill category (by averaging the zscores for the PageRanks that each skill possesses in each skill category). The measure corresponds to worker mobility because skill proximity is indicative of skill acquisition. This highlights the importance of social skills for labor mobility. For a detailed result of all skills, see fig.8-2.

4-6 The growth rate exponents for the two skill clusters (SocioCognitive vs. sensory-physical). This demonstrates that the occupations that have SocioCognitive skills grow superlinearly with city size and therefore, the presence of such skills in cities also grows superlinearly with city size.

5-1 Figure recreated from [84]. Gumbel probability distribution of imperfect transmission, based on Henrich [61]. This demonstrates how transmission and learning, which are facilitated by social ties, are prone to error (i.e. the distribution is attempting to mimic the transmitted skill value represented by the dash line). Some of those "distortions" are beneficial and are actually innovations that increase the value of the "skill". Those innovations and advancements are facilitated by the increased connectivity of social ties (in quantity or diversity).

5-2 Figure recreated from [6], where they studied communication effectiveness. This plot demonstrates the results of their empirical analysis and findings where they highlight the fact that optimal case is a balance between heterophily and lack thereof (i.e. homophily).

5-3 Summary of the steps in this chapter.
5-4 This figure demonstrates the results of using the feature vectors produced by the proposed model vs DeepWalk for predicting the occupation and gender of individuals in the TMDB network. Even though our model was tested on a variety of data sets, this results highlights both homophily and heterophily at the same time. Since our proposed model, which considers both drivers, outperforms DeepWalk in predicting occupations (reasonably assumed to be more driven by heterophily), while it underperforms in Gender prediction which is assumed to be more driven by homophily (which DeepWalk is purely focused on).

6-1 The strength of different indicators for the prediction of anomalous behavior. Results demonstrate that social ties are better predictors than “tweet” or profile contents.

6-2 The three messages shown to the three different groups, Control, Self Information Leakage, and Social Information Leakage Group, respectively.

6-3 The decision analysis results, which show the differences between each group in terms of users’ response to the SIL message, in addition to the media time taken to make the decision. The results show that the third group behaves differently than the other two groups.

6-4 Snapshot of the social network of the Friends and Family data set. We can see many features, such as stronger and more bidirectional ties within cliques than across the clusters that were identified through the clustering algorithm (which is stronger than random, and visually apparent).
6-5 Probability distribution for 10,000 two-dimensional points from three different one-dimensional laplacian noise generations at different locations ($\epsilon = .3$ (blue), $\epsilon = .4$ (green), & $\epsilon = .5$ (red)). The middle inset plot is a scatter plot of the generated noise points, while the left plot represents many probability density functions for the noise points in two dimensions, and the right plot is a two-dimensional heat-map for the noise points.

6-6 Architecture for the behavioral privacy algorithm.

6-7 Approach for comparing utilities of the state-of-the-art vs. the proposed behavioral privacy method.

6-8 The state-of-the-art method (red) vs. the behavioral privacy variant (blue), using different privacy parameters (x-axis) and then applying a linear SVM on the resulting distorted data sets. We show the resulting utility measures (AUROC values) on the y-axis. Inset plots represent the kernel density estimator (KDE) and cumulative distribution function (CDF) for the y-axis when aggregating over all of the x-axis.

6-9 Aggregated resulting p-values for all individuals in the data set using the various machine learning algorithms demonstrated. Each machine learning method is the aggregation of the various kinds of results (e.g., Call/Bluetooth -> Stress/Happiness) for all users. The 344 data sets represent around 86 individuals, each having four possible combinations for the pair of (Call vs BT) and (Stress/Happiness).
7-1 A summary for the flow of the thesis, starting from the initial chapters that transformed matrices of individuals’ job features to construct the skill network, which demonstrated the importance of the social dimension in the labor force (for both economic well-being and labor mobility). Then, using a city’s social network, later chapters employed a mechanistic model to capture the value of social ties for social exchange by reconstructing the individuals’ features embedded in the social network. ................................. 110

7-2 Cities with various sizes are at different stages of evolution for their skill portfolio. That is, smaller cities, on the left, mostly possess Sensory-Physical skills, while larger cities, on the right, possess mostly SocioCognitive skills. ................................. 111

7-3 (A) An example current status of acquired skills in the SkillScape. This could be an individual, occupation, or city. (B) demonstrates the weighing of costs and benefits for various skill acquisitions. ............. 111

8-1 Transforming raw O*NET data with RCA. The first plot is the raw occupation-skill matrix, $I(j, s)$, the middle plot is the RCA occupation-skill matrix, $rca(j, s)$, and the final plot is the thresholded RCA job-skill matrix, $e(j, s)$, for 2014. Here, $e(j, s) = 1$ if and only if $rca(j, s) > 1$. Occupations (y-axis) are ordered by the sum of threshold RCA skill values, and skills (x-axis) are ordered by the correlation of their thresholded RCA values across occupations to the occupational sums. 113

8-2 Full list of skills for figure 4-5 in the main text. PageRank scores for every individual Skill (node in the network). That is, the connectivity & embeddedness of each skill. Color represents skill category. ........ 115

8-3 Continuing PageRank of the SkillScape Skills in figure 8-2. ........ 116
Instead of PageRanks in figure 4-5 (which is an n-th order calculation), this is the first order calculation for comparison. This figure demonstrates complementarity scores for every skill category. That is, the average Z-score of each categories' node strengths (sum of it's edges, or "complementarity weights" $\theta$). Color represents skill category. We can see Social Skills still achieve the highest score.

Complementarity scores for every individual Skill (node in the network). That is, the Z-score of each node's strength (sum of it's edges, or "complementarity weights" $\theta$). Color represents skill category.

Continuing Node Strengths of the SkillScape Skills.

Instead of the interpolated plots in the main text (figure 4-3), here we provide bar plots with the associated error bars.

Out of sample testing of model performance from Table 8.2. For each model, 1,000 trials are run where 75% of the data is randomly selected as training data and the remaining 25% of data is used as validation. The distribution root-mean-square errors for each model is reported. Medians are represented by a red line, while the mean error is represented by the green square.

Out of sample testing of model performance from Table 8.3. For each model, 1,000 trials are run where 75% of the data is randomly selected as training data and the remaining 25% of data is used as validation. The distribution root-mean-square errors for each model is reported. Medians are represented by a red line, while the mean error is represented by the green square.

Out of sample testing of model performance from Table 8.4. For each model, 1,000 trials are run where 75% of the data is randomly selected as training data and the remaining 25% of data is used as validation. The distribution root-mean-square errors for each model is reported. Medians are represented by a red line, while the mean error is represented by the green square.
8-11 Out of sample testing of model performance from Table 8.5. For each model, 1,000 trials are run where 75% of the data is randomly selected as training data and the remaining 25% of data is used as validation. The distribution root-mean-square errors for each model is reported. Medians are represented by a red line, while the mean error is represented by the green square.

8-12 A cartoon example of AUROC calculation.

8-13 The top figures represent the fraction of instances (for cities or occupations) that have a change (i.e. \( \text{Acquired}_{t_1, t_2}^{\lambda_1, \lambda_2} \)) with the x and y axis representing the respective \( \lambda_1 \) and \( \lambda_2 \) values. While the dotted boxes in each of the top figures represent the zoomed area that the AUROC values will be studied in the bottom figures. The bottom figures represent comparison of AUROC distribution for all of the various cities of occupations. That is, for each \( \lambda_1 \) and \( \lambda_2 \) value, there are many instances of cities or occupations that have such a change, and we study the AUROC results for predicting such jumps using the different indicators (raw onet, RCA, or SkillScape’s proximity metric).

8-14 The varying averages of AUROC achieved by combining the different variables with varying degrees (two network based, and one raw data based), creating this Dirichlet triangle. SkillScape is a network features, while \( I \& LQ \) are none network features.

8-15 The skill requirements of an occupation indicate the education required. In each panel, we plot the SkillScape network thresholding edges with \( \theta > 0.6 \). Nodes (or skills) are colored according to the Pearson correlation between \( \text{onet}(j, s) \) and the proportion of workers of each occupation with a given degree (title).

8-16 The correlation between social status and degree distribution in the Andorra dataset.

8-17 The correlation between human capital and social capital in TMDB dataset.
List of Tables

3.1 We consider the SocioCognitive skill fraction \( (SocioCognitive_j) \) and the total embeddedness \( (\sum_{s \in S} I(j,s)) \) of each occupation in addition to educational variables in linear regression models. Each educational variable represents total employment in that occupation whose highest educational degree is a high school diploma, bachelor’s degree, etc. All variables were standardized before regression. Models for occupational wages are improved and less susceptible to over-fitting when accounting for the SocioCognitive skill fraction of occupations, as can also be seen in fig.3-9. ................................................................. 55

8.1 Descriptions of each occupation type indicator variable used in regression models. For each occupation, the indicator variable is 1 if and only if the occupation SOC code belongs to that occupation category. Each occupation belongs to exactly one occupation category. ........ 121

8.2 Linear regression using standardized \( SocioCognitive_j \) for each occupation and occupation type indicator variables. ................. 121

8.3 Linear regression using \( SocioCognitive_j \) and employment in each occupation with a bachelor’s degree (denoted B.D. Employment) and without a bachelor’s degree (denoted No B.D. Employment). Each variable has been standardized. Employment by level of education for each occupation is taken from onet data. ..................... 122
8.4 Linear regression using standardized $SocioCognitive_c$ for each city and employment in that city of each occupation type. All variables have been standardized.

8.5 Linear regression using $SocioCognitive_c$ and education variables. Education variables represent the employment in each city by highest educational degree attainment. All variables have been standardized.
Chapter 1

Introduction

1.1 Introduction

By 2050, it is expected that 66% of the world population will be living in cities [85]. The urban growth explosion in recent decades has raised many questions concerning the evolutionary advantages of urbanism, with several theories delving into the multitude of benefits of such efficient systems. This thesis focuses on one important aspect of cities: their social dimension, and in particular, the social aspect of their complex socioeconomic fabric (e.g. labor markets and social networks). Economic inequality is one of the greatest challenges facing society today, in tandem with the eminent impact of automation, which can exacerbate this issue. The social dimension plays a significant role in both, with many hypothesizing that social skills will be the last bastion of differentiation between humans and machines, and thus, jobs will become mostly dominated by social skills.

Using data-driven tools from network science, machine learning, and computational science, the first question I aim to answer is the following: what role do social skills play in today’s labor markets on both a micro and macro scale (e.g. individuals and cities)? Second, how could the effects of automation lead to various labor dynamics, and what role would social skills play in combating those effects? Specifically, what are social skills’ relation to career mobility? Which would inform strategies to mitigate the negative effects of automation and off-shoring on employment. Third,
given the importance of the social dimension in cities, what theoretical model can explain such results, and what are its consequences? Finally, given the vulnerabilities for invading individuals’ privacy, as demonstrated in previous chapters, how does highlighting those results affect people’s interest in privacy preservation, and what are some possible solutions to combat this issue?

1.2 Thesis Outline

I begin by demonstrating a network science method applied more than once in my thesis. I utilize causal inference to illustrate its value in one instance, in addition to comparing its variants. In the next chapter, I apply that method in one case study on the occupations of individuals. The aim is to highlight the value of social skills in occupations, their relation to education, and annual wages. Subsequently, I extend the approach using a few adaptations to fit the larger scale of labor markets, which demonstrates the importance of social skills for cities and their economic well-being. The following chapter then deals with the critical issue of the negative impact of automation on cities and individuals. The chapter discusses how the approaches and structures proposed in the previous chapter can be utilized to combat those negative effects using individually optimized retraining and educating programs. In addition, the chapter also presents a website tool created for such an objective. Next, the following chapter delves into a theoretical explanation of the previous results and observations and proposes a computational model for social tie formation that would highlight the benefit for social exchange and trade. This chapter then applies the proposed model on a few small- and large-scale empirical datasets to test the validity of the approach and demonstrate some of its capabilities. Finally, the last chapter addresses some of the privacy concerns that the previous chapters might raise. It first presents an experiment conducted to examine whether people would in fact have increased concerns, after which it proposes an approach to tackle the issue by introducing the concept of decoupling privacy from utility, followed by proposing an algorithm.
Chapter 2

Methods

This chapter presents one of the methods used multiple times in my thesis, and applies causal inference to show its value in one instance. This method is the “product space”, which has been applied to countries and products, in contrast to our application in this thesis being labor markets. The product space provided a new vantage point for understanding the ecosystem of product creation through exports and its possible future changes, which is key when evaluating countries’ economic strength. The literature [63, 62, 59, 103, 41, 42] has shown many methods of using network science to evaluate the underlaying structure of product creation, in addition to providing the ability to predict future product exports [32]. The aim of this section is to analyze these methods, evaluate them, and compare their power for predicting economic evolutions. The goal is to adapt these methods for our own use in later sections.

2.1 Introduction

In “The Product Space Conditions the Development of Nations” [62], Hidalgo et al. took a unique network science approach (fig.2-1) to analyze production ability; then, in “The Building Blocks of Economic Complexity” [63] they used their outputs to evaluate a country’s level of economic complexity and predict its future economic growth. Furthermore, Haussman et al. [59] showed how network science could be
Figure 2-1: Illustration of the product space connecting 775 products based on their proximity matrix. The color of the node represents the product classification. This is my replication [63, 59, 62] based on 2013 data.

used to evaluate a country’s competitiveness in products and therefore the knowledge it possesses to engineer those products. One of the most critical outputs of Hidalgo and Hausmann’s work was providing an original approach to measure the economic complexities of countries based on a network view of international trade data (i.e. the product space) using the method of reflection (MR). While this method proved to be useful for their analysis and the results that followed [63, 59, 62], later publications, first Tacchella et al.’s “A New Metrics for Countries’ Fitness” [103] and later Cristelli et al.’s work [41, 42], have argued for other methods, converging on a method named “fitness”, which has a different mathematical formulation than complexity. They argued for their metric’s stability, and stated that it assimilated more of the theory. Though extensive work has been published on the shortcomings of one method vs. the other, and though quantitative comparisons have been made [41, 42], no causal inference comparison between the methods has been published, either individually or pair wise.
2.2 Methods and Results

The main resource used in this work is the World Bank, which reports data about the world development indicators (WDI) for most countries. However, for data on per capita GDP at Purchasing Power Parity (PPP), the International Monetary Fund (IMF) is used. The main years used in the analysis are 1985, 1990, 1995, 2000, and 2005.

First, we discuss and analyze the parts of the methods on which both approaches agree, before discussing the areas where they diverge. To analyze the bipartite network of countries and products, we first define whether or not we consider a country to be connected to a product:

\[
M_{cp} = \begin{cases} 
1 & \text{if } \frac{X_{cp}}{\sum_{cp'} X_{cp'}} > \frac{\sum_{ct} X_{ctp}}{\sum_{ct} \sum_{cp'} X_{ctp}} \\
0 & \text{otherwise} 
\end{cases}
\]  

(2.1)

Where “c” represents a country and “p” represents some product. \(X_{cp}\) is the value of exports of product p from country c, while \(M_{cp}\) uses \(X_{cp}\) to decide whether or not country c has a comparative advantage in product p. If the fraction of product p exports out of country c is higher than the fraction of product p in the global export trade of the world, then country c indeed exports product p with a comparative advantage, otherwise it is not considered to be an exporter of p. For example, if oil represents 1% fraction of the world export trade, then a country needs more than 1% of its exports to be oil for it to be significant enough to be considered as an exporter of oil with a comparative advantage.

At this point, each method diverges in how it evaluates each node (country or product) in the network. In building the economic complexity index, Hidalgo et al. [62] characterized countries and products by introducing a set of variables that capture the structure of the network defined by \(M_{cp}\). Their methodology consists of iteratively calculating the average value of the previous-iteration properties of a node’s neighbors and is defined as the set of observables:
Initially, \( k_{c,0} \) and \( k_{p,0} \) represent the number of products with significant prominence out of country \( c \), and the number of countries that export product \( p \) with significant prominence, respectively. \( n > 0 \) indicates the number of iterations of refinement that the metric has gone through. Therefore, \( k_{c,n} \) converges to the country complexity index \( CCI \), see fig.2-2, and \( k_{p,n} \) converges to the product complexity index \( PCI \) for large \( n \) values.

These network representations of the data have shown great success in capturing information about a country’s level of “knowhow and knowledge” [59], demonstrated by predicting new exports of countries and correlating to GDP growth, even when controlling for various covariates [62]. It is important to note that instead of defining economic growth as the difference of logs between two years [62], we use a more accepted definition of it as the annualized growth rate:

\[
\text{Annualized GDP Growth}(t, t' = t + \Delta t) = \left[\left(\frac{GDP_{t'}}{GDP_t}\right)^{1/\Delta t} - 1\right] \times 100
\]

The regressions’ coefficient values and their significance estimates using this metric for economic growth are somewhat different than the ones reported in Hidalgo et al.’s original paper. This is to be expected given the change to the way economic growth is calculated. However, their hypothesis and final conclusions still hold.

Even though the previously discussed method has shown tremendous results, some in the literature have claimed that there are theoretical issues with it [103], ranging from assimilation of the theory [41], to instability of the method’s iteration values (see first plot in fig.2-2) [42]. To tackle these issues, a new method was developed in [103] by altering the equations and normalizing in each step of the reflections method:
This new version of the method (referred to as “fitness”) is very similar to the previous method (i.e. “complexity”). This can be seen from the work of Albeaik et al., which demonstrates that both methods have the same raw equation form with differing parameters (i.e. \((\alpha, \beta, \gamma, \delta, \epsilon, \theta)\) which can only take the values of \((1, 0, -1)\) [4]:

\[
\begin{align*}
\tilde{k}_{c,n} &= \frac{1}{\sum_p M_{cp} k_{p,n-1}} \\
\tilde{k}_{p,n} &= \frac{1}{\sum_c M_{cp} k_{c,n-1}} \\
k_{c,n} &= \tilde{k}_{c,n}/ <\tilde{k}_{c,n}> \\
k_{p,n} &= \tilde{k}_{p,n}/ <\tilde{k}_{p,n}>
\end{align*}
\]

\[
k_{c,n+1} = \frac{(\sum_p M_{cp} k_{p,n})^\gamma}{(\sum_p M_{cp})^\epsilon}
\]

\[
k_{p,n+1} = \frac{(\sum_c M_{cp} k_{c,n})^\delta}{(\sum_c M_{cp})^\theta}
\]

where \((\alpha, \beta, \gamma, \delta, \epsilon, \theta)\) are parameters that produce a family of metrics \((3^6 = 729)\), with \((1, 1, 1, 1, 1, 1)\) producing complexity and \((1, -1, 1, -1, 0, 0)\) producing fitness.
Figure 2-2: Complexity and Fitness methods’ convergence plots. Plot (A) shows all iterations of the complexity method, while plot (B) only shows the even iterations. Plot (C) however, is the z-score for each even iteration of complexity, making it more closely resemble fitness. Finally, plot (D) shows all iterations of the fitness method.

First, it should be noted that there are significant impacts (as can be seen in the last plot of fig.2-2) for normalizing in each iteration and taking the reciprocal of the sum of the reciprocals, which is done in the fitness approach. It is clear that the newer method of fitness is stable even when viewing the raw evaluations (in contrast to complexity’s raw iterations, fig.2-2). Second, fitness converges to a steady state,
but it does not converge to the same value as complexity does (with the variance becoming smaller and smaller in magnitude with more iterations).

2.3 Quantitative Comparison

First, in order to quantitatively compare the two methods, we can run the same regressions previously done [62] after modifying how GDP growth was calculated, and we can then compare the regression results of the two methods. For example, regressing the two methods over the GDP per capita values for different years and comparing resulting Adjusted $R^2$ produces an interesting result (see fig.2-3). Namely, Figure 2-3 shows the effects of the instability in the complexity method and how well it fits the dependent variable in different iterations. The results also show that complexity produces a better fit than fitness does, even though fitness is more stable and reaches a steady state faster. It should be noted that the superiority of one method over the other in these $R^2$ results is not indicative of much, since there are other covariates that would explain away much of the correlation power.
Figure 2-3: Adjusted $R^2$ values (y-axis) from regressing GDP for various years over the successive iterations (x-axis) of the two different metrics (complexity and fitness).

However, the hypothesis presented in the original literature hints towards a causal link. Hence, we further examine the relationship between GDP and the two different methods. To this end, we use GDP growth as our dependent variable, as the hypothesis presented in the literature indicates such a link. In addition, we add a few control variables to ensure that no meaningful and reasonable confounders that we should control for are left unused. The first of these control variables is the original GDP value at the initial point of time, to control for the size of the economy. The second is population size, as this could explain economic growth. The last control variable is the total enrollment in tertiary education, expressed as a percentage of the total...
population, as education could also explain economic growth. As for the treatment, we use a binarized value for the two metrics (larger than the average or less than the average):

$$ \text{Annualized GDP Growth}(t, t') \sim \text{treat} + \text{GDP}(t) + \text{Pop}(t) + \text{Edu}(t) + C \quad (2.6) $$

where $\text{GDP}$ represents the log of a country’s PPP GDP per capita, $\text{Pop}$ is the log of the country’s population, and $\text{Edu}$ is the percentage of enrollment in that country’s tertiary education. Moreover, $t$ and $t'$ they represent the year under study, with $t' = t + \Delta t$, and treat is the treatment variable (i.e. binarized complexity or fitness). Converting the treatment into a binary variable has some significant implications for what relates back to the raw metric, but one can assume that if the binary version shows a causal relationship, then that should generalize to the raw variable. Hence, the only result we hope to obtain this experiment is whether or not either or both treatments are indicative of economic growth; we do not assume that we can gain a good measurement of the magnitude of that relationship. Thresholding the treatment metric significantly hinders our ability to test its causality links, but if we still find a causal link after thresholding, that evidence will transfer to the continuous version as well. Otherwise (in cases where no statistical evidence is found), nothing can be stated about any causal link with regard to the continuous version. We do this for both $\text{treat} = \text{binary(Complexity)}$, and $\text{treat} = \text{binary(Fitness)}$. However, before running our regression, we use matching to approximate (simulate) a controlled experiment and reduce the imbalance in the covariates as much as possible. The method we use for matching is the coarsened exact matching (CEM) method for the cases of two and all three covariates (see fig.2-4).
After matching and controlling for covariates, we see that both complexity and fitness mostly have a large, statistically significant effect when regressing over economic growth (5 years and 20 years, but not 10 years). This does not mean that there is an issue with the 10-year growth rates, but that the problem is the lack of data points after matching for that time period (see fig.2-5 and fig.2-6). Figure 2-4 demonstrates a causal relationship between both complexity and fitness and GDP growth when controlling for two covariates.

Although coarsened exact matching is powerful in enhancing our ability to control for confounders, it omitted a significant amount of the limited data we had (right figure in fig.2-6). Fortunately, we still see that there is a significant sign for a causal link, and therefore conclude that complexity (and in a similar result, its variation fitness) both capture a dimension that has a causal impact on economic growth for countries (independent of the population and economic size), with no method having
Figure 2-5: The coarsened exact matching scatter plot visual for the two-variate matching, to demonstrate which data points were kept (matched) vs. dropped (unmatched). We see a significant number of remaining data points for statistical analysis.

Figure 2-6: This figure demonstrates the data points kept by the coarsened exact matching method, with the left figure being identical to fig. 2-5, and the right plot representing the three-variable variant of the matching. Note the limited number of remaining (matched) data points in the three-variable matching.
a significant advantage over the other.

2.4 Conclusion

These preliminary analyses indicate that further investigation might prove fruitful, such as incorporating causal inference analysis to study what relationships these metrics have with other economic factors (unemployment, income inequality, resistance to shocks). In addition, it could be interesting to conduct a deeper analysis of the methodologies applied and the parameter choices that were made, and whether other directions might be more appropriate. However, the results generally indicate that there is good reason to adopt these methods for our purposes in this thesis of studying the labor market dynamics of individuals/cities and their skills.
Chapter 3

The Importance of Social Skills in Labor Markets

This chapter discusses the role of social skills in today’s labor markets on both a micro and macro scale (e.g. individuals and cities). Specifically, the chapter introduces a novel approach of constructing a complementarity network for all of the skills in the labor market. Thus allowing us to unpack the polarization of workplace skills, which has been deemed to be the cause for the hallowing of the middle class, and then we investigate the value of social skills for occupations’ salaries and labor markets’ (i.e. cities’) economic well-being.¹

3.1 Introduction

Economic inequality is on the rise, making it one of the central challenges facing policy-makers today [67, 5, 39]. In recent decades, the middle class has shrunk in the vast majority of U.S. metropolitan areas [34]. For example, consider that absolute income mobility—the fraction of children who earn more than their parents—has fallen dramatically in the U.S. from 90% for children born in 1940 to 50% for children born in 1980 [36]. Some have declared that the diminishing opportunity for prosperity and

¹This work is to be published in “Ahmad Alabdulkareem, Morgan Frank, Lijun Sun, Bedoor Alshebli, Cesar Hidalgo, and Iyad Rahwan. Unpacking the polarization of workplace skills. Science Advances, 2018.”
success marks the fading of the “American dream” [93, 65], an ideal that is intimately associated with the U.S.’s national identity and ethos. This highlights the growing need to characterize low- and high-wage occupations, and to identify the constraints on career mobility between the two.

In contemporary political debate, one of the main culprits behind economic inequality is the lack of “good jobs”. Both nationally and in a majority of U.S. metropolitan areas [34], economists have identified occupational polarization: an increasing proportion of high- and low-wage employment, accompanied by a relative decrease in employment share in middle-wage occupations [13, 14, 1]. The result is a “hollowing” of the middle class. Mechanisms driving this trend include the off-shoring of work [48], which has triggered recent shifts in international trade policy. Another mechanism is the automation of routine work, which has sparked major concerns about the impact of automation on the future of work [75, 12, 24].

However, while mechanisms like off-shoring and automation ultimately impact people’s jobs, they do not typically operate at the level of occupations. Rather, they alter the demand for specific workplace skills, tasks, knowledge, and abilities (hereafter, “skills”). Thus, if individual workers, or even entire cities, are unable to adapt their own skills appropriately, their ability to compete in the national and global labor market may be diminished.

Despite the important role of skills in occupational polarization, existing studies have explained the hollowing of the middle class in terms of annual wages [11] and broad, subjectively defined occupational categories, such as “cognitive” versus “physical”, or "routine" versus “non-routine” [13]. For example, suppose we use wage as a proxy for skill (that is, high-wage occupations are considered high-skilled occupations, etc). Then, if we find that growth in employment in middle-wage occupations is slower than growth in low-wage and high-wage occupations, we may conclude that demand for high-skills and low-skills are driving economic inequality. However, this coarse-grained distinction may miss important relationships between skills that impact how workers adapt. This motivates the set of questions we wish to explore in this chapter:
Q. Can we recover occupational polarization, at the finer-grained level of underlying skills, using an objective (unsupervised) data-driven clustering? How many distinct clusters, if any, does this skill structure contain? And does the skill structure exhibit smooth or abrupt transition between skill clusters?

To answer these questions, we apply data-driven methods to map skill complementarity as a network. We find that workers leverage skill complementarity between their existing skills to make career changes [31]. Similarly, cities leverage complementarity between industries to optimize productivity and increase their competitiveness in a global economy [91, 92, 87, 86]. We use techniques from network science to identify distinct clusters of skills, and we find that the structure of skill complementarity explains many stylized observations about occupational polarization and the hollowing of the middle class.

### 3.2 Data Sets

The first data source for this chapter is the Occupational Information Network (O*NET): a program by the U.S. Department of Labor which annually produces a publicly available database detailing the importance of 161 workplace skills, knowledge, and abilities for the completion of each of the 672 occupations recognized under the Standard Occupational Classification (SOC) System, see figure 3-1.A.

This allows us to understand not only the decomposition of any given occupation, but also its relationship to all other occupations. The O*NET database is updated regularly, allowing for annual snapshots of the relationships between occupations and skills through continual survey of workers from each occupation. We use annual O*NET data from the years 2010 through 2015. We denote the importance of skill $s \in S$ to occupation $j \in J$ using $I(j, s) \in [0, 1]$, where $I(j, s) = 1$ indicates that $s$ is essential to $j$, while $I(j, s) = 0$ indicates that workers of occupation $j$ does not need to possess or perform $s$. 
Figure 3-1: (A) Heat-map visualization of the O*NET data matrix, which represents the relationships between occupations in the U.S. labor market and 161 skills \( I_{j,s} \). Every row is an occupation, while every column is a skill. (B) Heat-map visualization of the Bureau of Labor Statistics data matrix used, representing the number of employees for different occupations within the various cities. Every row is a city, while every column is an occupation.

The second data source is the Bureau of Labor Statistics (BLS), which annually produces publicly available data detailing the distribution of SOC occupations in each U.S. metropolitan statistical area (MSA) (see fig.3-1.B). MSAs represent an entire urban system, including areas with large proportions of commuters employed in the city proper. We use the terms MSA and “city” interchangeably. Along with the numbers of workers in each occupation, BLS provides additional details about the annual salary of each occupation in each city.

An important feature that should be noted, and that will be useful in future chapters, is of the nature of job markets’ evolution (whether for cities and jobs or for jobs and skills). This is demonstrated when we visualize the cross-sectional view of various densities of the different components (i.e. jobs or skills) within all of the hosts (i.e. cities or jobs), respectively. By sorting the hosts from most diversified to least...
diversified in the y-axis and arranging the components from the most ubiquitous to the least common, we obtain figures 3-1.A and 3-1.B. Notice the important characteristic of the heat-map structure being triangular in both cases. This is indicative of many features. We explain these features below using cities and jobs, but the same holds for jobs and skills.

- Ubiquitous jobs (first columns) are in almost all cities (all rows).
- Rare jobs (last columns) are mostly located in diversified cities (first rows).
- Non-diversified cities' job markets (last rows) are composed mostly of ubiquitous occupations (first columns).
- Not only do diversified cities have rare occupations, but they also maintain a strong presence in almost all occupation categories, indicating an additive nature to job market evolutions.

Therefor, the triangular structure hints towards a pattern of growth: cities grow in a path that follows the cities that are more diversified than them but similar in their composition (cross-sectional analysis can provide a longitudinal understanding and prediction). This leads us to further investigate the possibility that patterns of job market growth could be evaluated and predicted using only cross-sectional data sets and without any longitudinal inputs (i.e. predicting future evolutions using a single year's data set). This will be investigated in the chapter following this one.

3.3 The SkillScape: A Skill Complementarity Map

Typically, occupations are the unit of interest in labor dynamics. However, in other situations, occupations are broken down even further because the labor requirements that define an occupation are reflected in the skills possessed by workers of that occupation (see fig.3-2).

These skill requirements represent key features that uniquely identify occupations, hence we seek a data-driven methodology that maximizes the information about each
Figure 3-2: An occupation is identified through the skills of workers of that occupation. The bipartite network connecting occupations to required skills is a result of an underlying tripartite network containing workers as a conduit between occupations and skills. Relationships between skills are determined from their co-occurring importance across occupations.

occupation while minimizing the potential bias that can accompany investigations through ad-hoc skill aggregations. However, raw O*NET data do not control for ubiquitous skills, such as “Identifying Objects” and “Communicating with Supervisors and Peers” (see SI Appendix, fig.8-1). Therefore, we focus on skills that are over-expressed in an occupation by calculating the revealed comparative advantage (RCA) \[63, 62, 59\] of each skill in an occupation according to

\[
rca(j, s) = \frac{I(j, s)}{\sum_{s' \in S} I(j, s')} = \frac{\sum_{j' \in J} I(j', s)/\sum_{j' \in J, s' \in S} I(j', s')}{\sum_{j' \in J} I(j', s')/\sum_{j' \in J, s' \in S} I(j', s')}. \tag{3.1}
\]

where “j” represents an occupation within the set of all occupations “J”, and “s” represents some skill in the set of all skills “S”. \(I(j, s)\) is the importance of skill sto occupation jwithin the O*NET data set (see fig.3-1.A). Occupations are distinguishable from each other according to their “effective use” of skills. We denote effective use of skills using
\[ e(j, s) = \begin{cases} 1 & \text{if } \frac{I(j,s)}{\sum_{s' \in S} I(j,s')} > \frac{\sum_{s' \in S} I(j,s')}{\sum_{s' \in S, j \in J} I(j,s')} \\ 0 & \text{otherwise} \end{cases} \] (3.2)

Therefore, \( e_{j,s} \) represents a binarized version of \( I_{j,s} \). If the fractional importance of skill \( s \) to occupation \( j \) is higher than the fractional importance of skill \( s \) to all occupations, then occupation \( j \) does indeed require skill \( s \), otherwise no link is considered. For example, if the fractional importance of hand dexterity to all occupations is 1%, any occupation that has a fractional importance for hand dexterity higher than 1% is considered as needing it (or is linked to it, e.g. truck drivers are linked to hand dexterity).

Now that we have assigned skills to occupations, we have created an unweighted bipartite network of occupations and skills. We will now use the occupations that skills inhabit as a proxy for connecting skills together, to better understand their relationships. To relate pairs of skills based on their coexistence in different occupations or lack thereof, we use conditional probability. That is, given that some random occupation \( j \) requires some skill \( s (e(j, s) = 1) \), what is the probability that the same occupation also requires another skill \( s' (e(j, s') = 1) \):

\[ P(e(j, s') = 1 | e(j, s) = 1) = \frac{\sum_{j' \in J} e(j', s') \ast e(j', s)}{\sum_{j' \in J} e(j', s)} \] (3.3)

However, the conditional probability in equation 3.3 is vulnerable to outliers. For example, if a rare skill “rare s” is connected only to one occupation \( j \), then the conditional probability of all other skills \( s \) that are also connected to occupation \( j \) will equal one: \( P(e(j, s) = 1 | e(j, \text{rare } s) = 1) = 1 \) for all \( s \) connected to \( j \). To adjust for outliers and noise, in addition to smoothing out the network to make it an undirected network, we use the minimum of both directions for the conditional probabilities:

\[ \theta(s, s') = \min \left\{ P(e(j, s) = 1 | e(j, s') = 1), P(e(j, s') = 1 | e(j, s) = 1) \right\} \] (3.4)

Applying equation 3.4 ensures that even if there is a rare skill, it will only be
strongly connected to other rare skills that inhabit the same occupations, which in that case is an indicator of an actual strong relationship and not the result of one of the skills being an outlier. In other words, the “Skill Complementarity” [87, 31] (denoted by $\theta$) represents the pairwise skill complementarity by taking the minimum of the marginal probabilities of two skills each being effectively used by the same occupation:

$$\theta(s, s') = \frac{\sum_{j \in J} e(j, s) \cdot e(j, s')}{\max\left(\sum_{j \in J} e(j, s), \sum_{j \in J} e(j, s')\right)}$$ (3.5)

The distribution of complementarity values is provided in Figure 3-3. This novel approach identifies skill pairs that co-occur across occupations and represent key occupational features. Co-occurrence captures how two skills support each other, either by boosting the productivity of a worker who possesses both skills, or by making it easy to acquire both skills simultaneously. These proximities represent the relationships between every pair of skills, representing what we call the “SkillScape” network. Figure 3-3 shows the distribution of all the proximities in the SkillScape.
Figure 3-3: The distribution of all co-occurrence proximities between skills. Insets represent counterpart networks from other related works of literature for comparison. Unlike previous applications of co-occurrence networks (insets), the SkillScape contains a bimodal distribution of pairwise skill complementarity.

Notice how in comparison to other similar works in the literature (Producer, Product, and Occupational Spaces, etc.), which are unimodal distributions (e.g. a Weibull distribution [58, 59, 63, 99]), our case demonstrates a bimodal distribution. This segregation of proximities is indicative of a strong substructure (sub network represented by the right mode), followed by the weaker remaining part of the network (left mode). If we were to visualize the SkillScape in its entirety, we would simply obtain a fully connected (161x160) network with varying weights (figure 3-3). However, we identify skill types using Louvain community detection [27], which greedily identifies node communities by comparing the density of connections within a community to the density of connections between communities. The method requires no assumptions about the number of communities to be found. Furthermore, this community detection method has been widely used in a variety of fields, including neuroscience [96, 101], transportation research [20], social science [23], business/management research [45],
Figure 3-4: We identify two polars of skills by applying Louvain community detection to the complete SkillScape network (i.e., no minimum $\theta$). We notice that this almost perfectly captures the right mode of the bimodal distribution. That is, the links that are within one of the two clusters (intra-polar edges) mostly belong to the strong mode of the distribution, while the weak links within the weak mode almost exclusively belong to edges that are between clusters (i.e., inter-polar edges). The visual on the right only contains $\theta > .6$ edges. The complete list of O*NET skills in each cluster is presented in the SI Appendix.

climatology [49], and cybersecurity [40]. For visualization purposes, we filter out the weak part of the network, represented by the left hump (mode) of the weight distribution, but keep any edges that are part of the maximum spanning tree (MST) to yield an insightful result. We obtain the skeleton structure of the network colored by the communities identified by Louvain community detection, which demonstrates the very strong polarization in the network (see fig.3-4).

If we color the skills based on their O*NET categorization, we obtain the resulting filtered network of skill complementarity, which we call the “SkillScape” visual (see fig.3-5). This network has $\approx 2000$ edges from the right hump of the weight distribution (with weights $\geq .6$), plus 160 edges from the MST that will maintain the network as a single connected component. Node sizes in the figure represent PageRank values of the nodes in the complete network (not the visualized segment, which has $\approx 20\%$ of the edges), while node colors indicate the category to which a skill belongs.

While we will delve into them more deeply in later sections, the following are important features of this network that should be noted. One is the fact that the left side almost exclusively consists of high-level social and cognitive skills (e.g., prob-
Figure 3-5: A filtered visual for the SkillScape. The SkillScape thresholded according to a minimum skill similarity (i.e. $\theta > 0.6$) visibly reveals two communities of complementary skills and respects expert-derived O*NET categories (colors). Node sizes reflect the PageRank values, while color indicates the O*NET categorization of the skill.

Problem solving, communication, and scientific knowledge), while physical skills are on the right (e.g. manual labor, dexterity, strength), in addition to finer segmentation within each of the two main clusters. Another observation is the PageRank values for the different skill nodes: they are larger for social, intellectual, and communication skill types (the highest PageRanks belong to “Communicating with Persons Outside Organization”, “Social Perceptiveness”, and “Critical Thinking” skills), and lower for the physical and hard labor skill types (e.g. “Dynamic Flexibility”, “Explosive Strength”, “Building and Construction”). Finally, the aggregate structure in the skill network should correspond to meaningful labor dynamics. For example, node communities in the skill network represent clusters of complementary skills that define important
Intra-Category Edges (Within Category)

Inter-Category Edges (Across Category)

Figure 3-6: The SkillScape network respects experts' skill categorization. For each O*NET skill category, we measure the distribution of $\theta$'s for pairs of skills within a category (blue) and compare it to the distribution of $\theta$'s for each edge connecting a skill within the category to a skill outside of the category (red). The complementarity for skills within a category is significantly stronger according to the KS statistic (title) than the complementarity for skills across categories.

types of labor. This can be demonstrated if we study the connectivity of similar types of skills, see figure 3-6.

3.4 The SkillScape and Occupations

The divide between traditionally "technical" and "non-technical" skills largely supports previous findings characterizing U.S. occupational polarization. For example, let $SC$ denote the set of social and cognitive skills according to the community detection algorithm (blue nodes in fig.3-4 or fig.3-10). We measure the SocioCognitive skill fraction of a job $j$ by calculating the fraction of effectively used skills (e.g. see
(A) Nuclear Technicians

(B) Nuclear Medicine Technologist

(C) Nuclear Power Reactor Operators

(D) Nuclear Engineers

Figure 3-7: Four different occupations, showing how they drastically differ in SkillScapes. Nodes of skills that an occupation requires \( e(j, s) = 1 \) are maintained (colored in black), while skills that the occupation does not require are ignored (colored in white). It is clear that a job such as (A) nuclear technician requires mostly sensory-physical skills, in contrast to (D) nuclear engineer, for example, which requires mostly SocioCognitive skills.

Jobs with higher \( \text{SocioCognitive}_j \) tend to yield higher annual wages (see fig.3-8, Pearson correlation \( \rho = 0.46, p_{val} < 10^{-37} \)). This result demonstrates the direct link between the skill polarization we have identified and occupational polarization, which is characterized by growing employment share for high- and low-wage occupations [11].
Figure 3-8: Occupations requiring more SocioCognitive skills tend to make higher annual salaries.

To test how much information is captured with SkillScape, we regress the mean annual occupation salaries for all occupations within the U.S. labor market against other data variables such as education and total skill magnitude. The results are shown in table 3.1.
Table 3.1: We consider the SocioCognitive skill fraction ($SocioCognitive_j$) and the total embeddedness ($\sum_{s \in S} I(j, s)$) of each occupation in addition to educational variables in linear regression models. Each educational variable represents total employment in that occupation whose highest educational degree is a high school diploma, bachelor’s degree, etc. All variables were standardized before regression. Models for occupational wages are improved and less susceptible to over-fitting when accounting for the SocioCognitive skill fraction of occupations, as can also be seen in fig.3-9.

Figure 3-9: Performed out-of-sample testing for each model in table 3.1 through 1,000 trails of randomly selecting 75% of the occupations as a training data set and measuring the root-mean squared error of the resulting model applied to the remaining 25% of occupations. We represent the resulting model performance as box plots. Medians are represented by a red line, while the mean error is represented by the triangles.
Model 1 demonstrates the superior performance of SocioCognitive\textsubscript{j} \((R^2 = 0.15)\) compared to the educational variables in model 2 \((R^2 = 0.126)\), where we consider variables for each occupation’s total employment whose highest educational attainment is a high school diploma, a bachelor’s degree, etc. In addition, we consider the total skill content required by each occupation (i.e. \(\sum_s \text{onet}(j, s)\) which is equivalent to \(\sum_{s \in S} I(j, s)\)) in Model 4 \((R^2 = 0.301)\). The regression results in table 3.1, especially models 1, 4, and 5, demonstrate that the SkillScape’s SocioCognitive variable captures information that is orthogonal to what is captured by the raw data feature \(\sum_s I_{j,s}\), demonstrated by the \(R^2\) value of model 5 equaling the sum of models 1 and 4’s \(R^2\) results. In addition, we provide out-of-sample testing to demonstrate the robustness of our models’ performance. We find that the SocioCognitive models (1,5) produce the best results, which also have the least variance in model performance and do not suffer from over fitting.

3.5 Skill Polarization Using a Bottom-Up Approach

Previous investigations of occupational polarization have examined the role of workplace skills. However, in this study, we employ an entirely data-driven bottom-up approach to the problem. Unlike previous studies, our methods for measuring skill complementarity make no assumptions about the meaningful aggregations of skills (i.e. skill clusters represented by network communities) that may result from the assessments of actual workers through the O*NET data set. For example, existing studies have explained the hollowing of the middle class [5] in terms of annual wages [11] and broad, subjectively defined occupational categories such as “routine” versus “non-routine” [13]. It has also been shown that some decades are marked by a relative increase in the share of employment in high-wage and low-wage jobs, at the expense of workers in middle-wage jobs. While these results identify the outcome of labor polarization, they do not relate this polarization to the underlying topology of skills. These limitations led researchers to call for new high-resolution models that more accurately account for raw workplace tasks and skills [1].
On aggregate, our cluster analysis reveals that the skill network is highly polarized into a SocioCognitive and a sensory-physical cluster of skills (see fig.3-10). This polarization is not an artifact of the methods we employ, and is significantly different from comparisons to a null model (see fig.3-3). Unlike previous studies that fall short of a comprehensive understanding of the entire space of workplace skills, our data-driven methods grant us the highest resolution possible by capturing raw workplace skills.

One might wonder whether our approach to skill polarization captures factors beyond those well known in the literature. For example, Autor et al. presented an ad-hoc distinction between jobs based on their reliance on routine versus non-routine
skills [15, 16], and they subsequently used this distinction to highlight occupational polarization [1]. Does our approach to skill polarization add further predictive power? Furthermore, education level is a key factor in determining wages [66, 11]. Educational institutions act as a social “sorting machine” [66] when students begin their careers. Indeed, if we correlate O*NET skill requirements and the average degree requirement for each occupation, we find that skills in the SocioCognitive cluster indicate higher education requirements across occupations. Conversely, occupations with more lenient degree requirements tend to rely on sensory-physical skills.

So, to re-iterate the question: are differences in wages explained by the existing ad-hoc distinction between routine vs. non-routine skills, and the level of education? Or does the polarized structure of the skill network we have identified play an independent role? We investigate these questions by comparing different regression models in Figure 3-11.

In Model 1, we consider the relative importance of routine labor by combining O*NET data with the routine O*NET variables defined in [15], that is:

\[
\text{Routine Labor} = \sum_{s \in R} I(j, s) / \sum_{s \in S} I(j, s)
\] (3.7)

where \( R \) are routine O*NET variables, \( R^2 = 0.12 \). Model 2 demonstrates the superior performance of SocioCognitive \( j \) \( R^2 = 0.15 \). Models 4, 5, and 6 demonstrate total skill content, and SocioCognitive skill fraction outperforms models using the variable for routine labor (Model 6 has \( R^2 = 0.46 \)). Finally, Model 8 demonstrates the improved performance from including the variable for routine labor and total skill content \( R^2 = 0.42 \), but maximum performance is achieved by including SocioCognitive \( j \) as well (Model 9 has \( R^2 = 0.49 \)). We provide out-of-sample testing to demonstrate the robustness of our models’ performance. We find that the inclusion of skill-related variables in Models 8 and 9 reduces the variance in model performance.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
<th>Model 6</th>
<th>Model 7</th>
<th>Model 8</th>
<th>Model 9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent Variable:</strong></td>
<td>Annual Wage of Occupations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive</td>
<td>0.357***</td>
<td>0.352***</td>
<td>0.614***</td>
<td>0.572***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.036)</td>
<td>(0.078)</td>
<td>(0.062)</td>
<td>(0.063)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routine Labor</td>
<td>-0.352***</td>
<td>-0.288***</td>
<td>0.203***</td>
<td>-0.268***</td>
<td>0.233***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.037)</td>
<td>(0.078)</td>
<td>(0.062)</td>
<td>(0.063)</td>
<td>(0.032)</td>
<td>(0.031)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sum_{j \in S} I(j,s)$</td>
<td>0.548***</td>
<td>0.513***</td>
<td>0.574***</td>
<td>0.488***</td>
<td>0.550***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.033)</td>
<td>(0.031)</td>
<td>(0.031)</td>
<td>(0.031)</td>
<td>(0.033)</td>
<td>(0.031)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No GED</td>
<td>-0.094*</td>
<td>-0.066*</td>
<td>-0.077*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.045)</td>
<td>(0.037)</td>
<td>(0.035)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H.S. Diploma</td>
<td>-0.170**</td>
<td>-0.087*</td>
<td>-0.077*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.047)</td>
<td>(0.038)</td>
<td>(0.035)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bachelor’s Degree</td>
<td>0.097*</td>
<td>0.062*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.042)</td>
<td>(0.036)</td>
<td>(0.035)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Master’s Degree</td>
<td>0.064</td>
<td>-0.040</td>
<td>-0.057*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.041)</td>
<td>(0.036)</td>
<td>(0.035)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doctoral Degree</td>
<td>0.240***</td>
<td>0.191***</td>
<td>0.151***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.037)</td>
<td>(0.031)</td>
<td>(0.029)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| $R^2$                     | 0.124         | 0.150         | 0.301         | 0.151         | 0.582         | 0.460         | 0.126         | 0.424         | 0.490         |
| adj. $R^2$                | 0.122         | 0.149         | 0.299         | 0.148         | 0.380         | 0.458         | 0.119         | 0.418         | 0.483         |

Figure 3-11: Reliance on SocioCognitive skills predicts increased annual wages. As a baseline, we consider the relative importance of routine labor using routine O*NET variables from [15]. In addition to cognitive skill fraction ($SocioCognitive_j$), we calculate the total skill content ($\sum_{j \in S} I(j,s)$) of each occupation. Each educational variable represents the total employment in that occupation whose highest educational degree is a high school diploma, bachelor’s degree, etc. All variables were standardized before regression. Standard errors are reported in parentheses and asterisks indicate the statistical significance of coefficient approximations. We perform out-of-sample testing for each model through 1,000 trails of randomly selecting 75% of the occupations as training data and measuring the root-mean squared error of the resulting model applied to the remaining 25% of occupations. We represent the resulting model performance as box plots. Median error is represented by a red line, while the mean error is represented by the triangles.

In summary, we find that the SocioCognitive skill fraction ($SocioCognitive_j$) explains the annual wages of occupations better than models using routine labor or educational variables alone.
3.6 Moving into Labor Markets

Labor markets are one of the main drivers of human society, from fostering innovation, and dictating unemployment, to having a profound impact on the economy. Understanding the ecosystem of occupations and its possible changes is key when planning for the future growth of a society. As per Adam Smith’s concept of division of labor, knowledge and knowhow are being “divisioned”, manifested in the segregation of capabilities, information, and tasks that are required by different occupations. Individuals who hold those occupations are therefore carriers for different segments of knowledge and knowhow. Many studies have attempted to capture those pieces of knowledge and knowhow through proxies; whether through the products that societies create (Hidalgo et al. 2007), research publications they publish (Guevara et al. 2016), Technologies they produce (Alstott et al. 2015), the occupations they hold (Shutters et al. 2015), or even the countries they inhabit (Bahar et al. 2012) [63, 58, 7, 99, 18].

The main approach of this thesis is not to use proxies (products, publications, industries, or even technologies) that capture the outputs of those carriers of knowledge and knowhow. Instead, the aim is to peek into the actual components of knowledge and knowhow and how they interdependently interact with each other, in addition to their interactions with their hosting ecosystem. This is done in this section using O*NET, which provides us with the different requirements, skills, and tasks that can be thought of as pieces of knowledge and knowhow (i.e. “skills”) required for every occupation in the U.S. labor market.

First, we start by analyzing the labor market using only the City/Job matrix (see fig.3-1.B) and process the data to produce a binary matching of jobs and cities. If we connect jobs together to better understand their relationships using the cities that they inhabit as a proxy, we obtain a similar network as the one presented by Shutters et al. [99] (see fig.3-12.A). If we analyze the edge distribution for this network in its entirety (not the minimal spanning tree), we see a single mode distribution (inset in figure 3-3). Figure 3-12.A shows the minimum spanning tree of that network, with node sizes reflecting the number of employees for that job within the U.S., while node...
colors indicate the industry cluster to which that job belongs, for each of the 600 occupations in the U.S. To demonstrate some examples of the relationships within the job market network, fig.3-12.B provides a closer examination of a subpart of the graph and a few examples of jobs that strongly correlate and "coexist" together.

Figure 3-12: The labor market network connecting jobs based on their coexistence in cities, similar to the one presented by Shutters et al. Panel (A) The job market network, with node colors representing the industry cluster to which an occupation belongs, where red = Production Occupations; blue = Education, Training, and Library Occupations; green = Healthcare Practitioners and Technical Occupations; cyan = Life, Physical, and Social Science Occupations; purple = Office and Administrative Support Occupations; and gray = all other clusters. (B) provides a closer examination of a subpart of the graph and a few examples of jobs that strongly "coexist" together.
However, even in the work of Shutters et al. [99], the network was highly limited in its ability to demonstrate insightful information without subjective categorizations of occupations. Furthermore, using our additional data set of occupations and skills, we have a new dimension to infuse into our analysis.

3.7 The SkillScape and Cities

In this section, we need to know the explicit relationship between cities and skills. We achieve this by using the previously mentioned O*NET data set in combination with cities’ job composition (BLS) as a proxy. This produces a link between cities and the prevalent skills contained in their job markets, to capture how much each urban workforce relies on each skill (i.e. a city/skill matrix, see fig.3-13.C).

Figure 3-13: Heat-map visualizations of the data. (A) represents the number of employees for different occupations within the various cities. (B) represents the relationships between occupations in the U.S. labor market and 161 skills “I(j, s)”. (C) shows that by multiplying the BLS city/occupation data set with the occupation/skill data set from O*NET, we obtain a matrix that relates all cities to every skill “I(c, s)”.

Denoting the number of workers in city c with occupation j using \( bls(c, j) \), we combine the two data sets according to

\[
I(c, s) = \sum_{j \in J} bls(c, j) \cdot I(j, s), \tag{3.8}
\]
where $I(c, s)$ denotes city $c$'s reliance on workplace skill $s$ (a city's equivalent to occupations' $I(j, s)$).

As with the raw O*NET data, certain jobs and certain skills are ubiquitous across many cities. We apply the location quotient (LQ), which is the geographical equivalent to RCA, to $I(c, s)$ to identify which skills are effectively used in each city according to

$$LQ(c, s) = \frac{\frac{I(c, s)}{\sum_{s \in S} I(c, s)} \sum_{c \in C} \frac{I(c, s)}{\sum_{c \in C, s \in S} I(c, s)}}{\sum_{s \in S}}.$$

(3.9)

Similar to occupations, $LQ(c, s) > 1$ indicates the effective use of $s$ in $c$; that is, if $LQ(c, s) > 1$ then $e(c, s) = 1$, otherwise $e(c, s) = 0$. Using the relationships between cities and skills (i.e. $e(c, s) = 1$), we obtain figure 3-14. We can see how different cities with varying skill compositions can easily be distinguished by looking at their SkillScapes, providing a high level insight into a city's labor market (in this binary “black and white” visual). In addition, we can visually see an evolutionary path that cities follow in growing their labor markets. Smaller, simpler job markets are in one (less complex sensory-physical) area of the network, while the larger, more competitive job markets are in the other (more desired SocioCognitive) area of the network, with a few cities in transition.

Figure 3-14: The SkillScapes of different U.S. cities, showing an evolutionary path through different SkillScapes. Nodes of skills in which a city has significant presence, i.e. $e(c, s) = 1$, are maintained (colored in black), while other skills are ignored (colored in white).

### 3.8 Social Skills and Cities' Economic Well-Being

Occupations are often considered to be part of an urban workforce [105, 50, 72] because cities represent humanity’s hubs for innovation [90, 26, 52] and economic
growth [56, 94]. Does skill polarization explain economic well-being across cities? Inspired by work detailing the occupations and labor complementarity of cities [99, 68, 17], we use employment distributions for U.S. MSAs compiled by the BLS. By considering \( I(c, s) \) in place of \( I(j, s) \) in equation (3.6), we can compute the same SocioCognitive skill fraction (denoted \( SocioCognitive_c \)) for entire cities. Similar to occupations, cities that employ larger workforces tend to require more SocioCognitive skills (\( \rho = 0.39, p_{val} < 10^{-13} \)), and hence have higher median household incomes (\( \rho = 0.38, p_{val} < 10^{-13} \)) according to the 2014 American Community Survey (see Fig.3-15).

Figure 3-15: Larger cities (i.e. in population and number of workers in the labor force) increasingly rely on SocioCognitive skills, leading to economic well-being (e.g. higher median household income and GDP). Example cities are projected onto the SkillScape using black nodes for effectively used skills.

Together, these results suggest that inequality between cities may be driven by processes that operate at the level of skill supply, and the ability of cities to effectively
exploit skill complementarity within the SocioCognitive niche.

3.9 Conclusion

Occupational polarization has been studied in the literature using broad, subjective occupation categories (i.e. "routine" or "non-routine") that fail to capture the dynamics of workplace skills between low- and high-wage occupations. Instead of subjective occupation categories determined entirely by annual wages, we proposed a purely data-driven methodology to map the space of workplace skills based on skill complementarity. The resulting network of skills is polarized in a way that respects stylized facts about occupational polarization; in particular, skill communities distinguish between occupations of different annual wages, thus demonstrating the direct connection between skill polarization and the "hollowing of the middle class".

While applications of similar methods (complementarity networks) to other systems reveal unimodal edge-weight distributions [63, 99, 58], the SkillScape contains a bimodal distribution of skill complementarity, forming a polarized network of two skill communities according to Louvain community detection [27]. One community is comprised of SocioCognitive skills and one of sensory-physical skills (see SI for a list of O*NET skills comprising each cluster). Projecting occupations onto the SkillScape according to the skills they effectively use revealed that higher income occupations rely more strongly on SocioCognitive skills. Furthermore, we demonstrated that SocioCognitive skills indicate a city’s economic well-being.

While our methods provide more texture to changing labor demands, they have some limitations. Firstly, while the O*NET database facilitates the improved resolution of our model, the taxonomy of O*NET skills may not capture the real-time dynamics of skill categories. For example, consider that a job listing for a software developer in the 1990s may only require "programming" skills, while modern listings might require specific types of programming skills, including proficiency in Hadoop, Java, or Python, for example. The O*NET database may miss this change in skill specificity until the taxonomy of skill categories is explicitly updated. External data
sources, such as LinkedIn, provide user-defined skills that may allow the future study of skill category dynamics, though they suffer from being non-representative. The second limitation is that while our methods provide a data-driven view of the structure underlying U.S. labor, they do not account for general market equilibrium dynamics that accompany changing skill demands. For example, how would the advent of new technology that performs a specific workplace skill change the skill network? And how does the relative cost of capital equipment play into decisions to retrain workers or purchase software or hardware? Answering these types of questions requires knowledge of other mechanisms, such as demand elasticity or capital availability, in addition to knowledge about the skill’s location in the skill network. Nevertheless, we hope our framework inspires further investigation into how skill structure dynamics interact with economic equilibrium dynamics studied in traditional models.

In conclusion, the SkillScape provides a structure of workplace skill complementarity by which the different cities’ and occupations’ skill composition can be viewed and evaluated. These preliminary analyses indicate that further investigation might prove fruitful, such as incorporating causal inference analysis to study the relationships between cities’ skill composition evolutions and economic factors (economic growth, unemployment, and income inequality). In addition, it could be interesting to conduct a deeper study of the applied methodologies and parameter choices that were made, and whether other directions might be more appropriate.
Chapter 4

Social Skills for Labor Mobility and Combating Automation

This chapter demonstrates how the impacts of automation could lead to various labor dynamics (e.g. labor mobility), and the role of social skills in combating those negative impacts. Specifically, the chapter examines the value of social skills in worker job transitions, occupation redefinition, and urban workforce adaptation to inform strategies for mitigating the negative effects of automation and off-shoring on employment. Ending with the result that larger cities are hosts to skills that allow for more labor mobility, thus not only being more desired, but also more robust towards the impacts of automation.

4.1 Introduction

Having mapped the structure of skills in the previous chapter, an obvious question to ask is: How can we use this structure (e.g. for labor mobility)? Studies have identified the aggregate effects of skill complementarity on labor dynamics, such as the redefinition of skills comprising each occupation [24]. We unpack the role of skill complementarity in labor dynamics by exploring the following questions:

Q1. Can the skill structure help us predict changes in the skill requirements of a given job, i.e. how the job composition itself changes over time, for example
due to automation?

Q2. Can the skill structure help us predict changes in the skills of individual workers as they transition from one job to another?

Q3. Can the skill structure help us predict changes in the latent skills of different urban labor markets (cities)? That is, given the skills used effectively in a given city at time $t$, can the network structure help us predict new skills in which that same city will become competitive in at time $t + 1$?

Having shown in the previous chapter that skill polarization exists and affects some key dynamics, the final natural question to ask is:

Q4. Is worker mobility between sensory-physical and SocioCognitive occupations constrained by the polarized structure of skills?

Our analysis suggests that the answer is “yes”. We provide three types of evidence: i) workers tend to transition between occupations relying on the same skill set, ii) workers are unable to switch away from occupations relying equally on SocioCognitive and sensory-physical labor, and iii) this constraining effect is reflected in national employment statistics.

4.2 Data Sets

The U.S. Census Bureau and BLS produce a monthly current population survey (CPS) through a continuous survey process that produces representative samples of the U.S. population. Providing high-resolution labor statistics is one of the primary goals of the CPS, and, in particular, the CPS records changes in occupations of survey participants over a 1.5-year period for which those participants are active contributors to the survey. For our purpose, we are interested only in participants who reported one occupation when they were first surveyed in 2014 and then reported working in a different occupation when they were surveyed one year later in 2015. There are several methods for joining different time periods of the CPS data [79], so we employ
strict merging criteria including participant ID, gender, sex, state of residency, and age to verify the validity of our occupation transitions. The result is a dataset of 5,400 occupation transitions for individual U.S. workers from 2014 to 2015. In addition, we use the automation probabilities published by Frey and Osborne [54] to superimpose automation probabilities onto our SkillScape, and compare those probabilities with both $SocioCognitive_j$ and $SocioCognitive_c$.

4.3 Motivation

There is a profound change sweeping through our times. The fourth industrial revolution is upon us and will result in increased automation of many tasks and even some complete occupations in the near future. Frey and Osborne [54] have published a detailed analysis of the likelihood of automation for many professions in the labor market (~700 occupations). In the following, we use their results of the automation probability for each skill in the SkillScape (see fig.4-1), which allows us to compare their automation probabilities with our SocioCognitive fraction metric for both occupations and cities (see fig.4-2).
Figure 4-1: Frey and Osborne's [54] automation probabilities superimposed onto our SkillScape.

Figure 4-2: (A) Relationship between the predicted impact of automation on cities and the cities' SocioCognitive$_c$ metric. (B) Relationship between the predicted probability of automation for occupations and their SocioCognitive$_f$ fraction.

### 4.4 Methods

These methods are used to try to determine whether or not skill complementarity (i.e. $\theta$) correspond to "nearby" skills in practice. We capture this using a measure for the network "proximity" between each pair of skills based on the network topology
and an empirical measure for skill acquisition. Let $E_t^\lambda(j)$ represent the set of skills that job $j$ effectively uses at time $t$ according to some threshold $\lambda \geq 0$; that is:

$$E_t^\lambda(j) = \{s \in S \mid rca_t(j, s) > \lambda\}. \quad (4.1)$$

A skill is "acquired" if it was not effectively used at time $t_1$ and becomes effectively used at $t_2$. Specifically, we denote occupation $j$'s set of acquired skills using:

$$Acquired_{t_1,t_2}^{\lambda_1,\lambda_2}(j) = \{s \in S \mid s \notin E_{t_1}^{\lambda_1}(j), s \in E_{t_2}^{\lambda_2}(j)\}. \quad (4.2)$$

According to this definition, two different thresholds, $\lambda_1$ and $\lambda_2$, are selected for time step $t_1$ and $t_2$, respectively. This allows us to vary the magnitude of skill change we are interested in; that is, $\lambda_2 - \lambda_1$ determines the severity of the skill change in order for a skill to be acquired for $\lambda_2 > \lambda_1$. Notice that if $\lambda_1 > \lambda_2$, then this would be skill loss instead of acquisition. For the analysis in the main text, we consider discrete choices of $\lambda$ according to each percentile of empirical RCA values (i.e. $\lambda_1, \lambda_2 = 0\%, 1\%, \ldots, 99\%, 100\%$ such that $\lambda_1 < \lambda_2$).

For a measure to be predictive of skill acquisition, skills with high scores (e.g. in O*NET) should have a higher probability of being acquired for each choice of $\lambda_1$ and $\lambda_2$. For example, if we consider raw O*NET values (i.e. $onet(j,s)$) as a proxy for skill acquisition, then skills that are not effectively used by an occupation (i.e. $s \notin E_{t_1}^{\lambda_1}(j)$) but have a high score (i.e. $onet(j,s) \rightarrow 1$) should have a higher probability of being acquired. We capture this by ordering pairs of occupations and skills by their O*NET value such that the skill is not effectively used by that occupation (i.e. $s \notin E_{t_1}^{\lambda_1}(j)$) and binning these pairs into 30 quantiles according to their associated O*NET value (i.e. $onet(j,s)$). For each pair, we calculate the probability that the skill is acquired in $t_2$ (i.e. $s \in Acquired_{t_1,t_2}^{\lambda_1,\lambda_2}(j)$) across all choices of $\lambda_1$ and $\lambda_2$. This produces several points for each quantile; we use the average and the 95% confidence interval for each quantile to simplify the data for visualization. This method is similar to previous studies' use of network topology to predict the regional acquisition of new industries [87]. In the main text, we visualize an interpolation through the averages
of each quantile. In addition to raw O*NET as a proxy for skill acquisition, we also consider RCA values and a measure of network skill proximity (described below). In the main text, we consider interpolation through the averages of each quantile. However, we also provide bar plots with the associated error bars in the SI Figure 8-7.

For each non-effectively used skill "s" (i.e. \( s \notin E_{t_1}^{\lambda_1}(j) \)), we say that it is near occupation \( j \) if that skill \( s \) has strong average complementarity with the effectively used skills of \( j \) (i.e. \( s' \in E_{t_1}^{\lambda_1} \)). We capture this by introducing a topological measure for proximity according to

\[
proximity(j, s) = \frac{\sum_{s' \in E_{t_1}^{\lambda_1}(j)} \theta(s, s')}{\sum_{s' \in S} \theta(s, s')}.
\]  

This proximity measure only utilizes information at \( t_1 \) to evaluate the status of all skills. Note that analogous calculations can determine SkillScape proximity from urban workforces by considering \( rca(c, s) \) instead of \( rca(j, s) \), and similarly for individual workers. SI Appendix Figure 8-14 provides an alternative analysis using receiver operating characteristic curves (ROC).

### 4.5 Labor Mobility

Skill acquisition through explicit education can be costly and time consuming. Therefore, workers more commonly transition between occupations based on the similarity of their skill set and the skill requirements of each occupation. The granular network topology of skill complementarity should capture this dynamic. In combination with the aggregate polarization of skills, we also expect that worker mobility between skill categories should be constrained. This hypothesis is not directly testable because we do not understand the precise mechanisms for worker adaptation, nor do we understand that mechanism's interplay with other market equilibrium dynamics [1, 24].

However, the hypothesis does have has three testable predictions. First, we pre-
dict that the topological proximity of skills on the network are informative about the
dynamics of skill-related trends, including the changing skill requirements of individu-ual workers, the dynamic skill requirements of occupations, and changes in the latent skill sets of urban labor markets. Second, we also predict that workers are more likely to transition to occupations relying on skills in the same skill cluster. Third, skill polarization represents a bottleneck in workers’ upward mobility towards high-wage occupations. This should lead to disproportionately high employment below a certain SocioCognitive threshold, rather than a smooth distribution of employment across the range of SocioCognitive values. In the remainder of this section, we demonstrate these predictions empirically.

We validate our first prediction in Figure 4-3 using a topological measure for skill proximity (i.e. proximity$((j, s))$, see fig.4-3.A for an example of proximity). Having a skill be near (i.e. have a high “proximity”) in terms of network topology should make that skill more likely to be obtained by a worker of that occupation, and more likely to become a skill requirement of that occupation. Analogously, nearby skills to a city’s local labor market should be more likely to be obtained by workers in that city. We empirically validate our proximity measure by demonstrating its relationship to the probability that a skill will be acquired (i.e. $s \in Acquired_{i_1, i_2}$) by a city (see fig.4-3.B), an occupation (see fig.4-3.C), or an individual worker (see fig.4-3.D). In each case, network proximity most strongly indicates newly acquired skills, thus demonstrating the highly granular relationship between the skill network topology and labor dynamics. We provide an alternative analysis in the SI Appendix (e.g. fig.8-14), as well as bar plots including 95% confidence intervals in the figure 8-7 of the SI.
Figure 4-3: Skill proximity predicts worker transitions between occupations, skill redefinition of occupations, and skill acquisition in cities. (A) An example demonstrating SkillScape proximity (i.e. proximity(j, s)) as a proxy for the connections between effectively used skills and other skills. (B) Skills with high proximity to the effectively used skills of an urban labor market in 2010 are more likely to be effectively used by that workforce in 2015. (C) Skills with high proximity to the effectively used skills of an occupation in 2010 are more likely to be effectively used by that occupation in 2015. (D) The effectively used skills in a worker’s occupation in 2015 are more likely to be effectively used in the worker’s next occupation in 2016. We provide bar plots including 95% confidence intervals for these probabilities in SI figure 8-7, and we consider an alternative area under the receiver operating characteristic curve (AUROC) analysis in SI figure 8-14.

For our second prediction, since occupational transitions represent local changes in workers’ skill requirements, the polarized network of skills should constrain mobility between low-wage sensory-physical occupations and high-wage SocioCognitive occupations. We capture this explicitly by binning occupation transitions into quantiles (each representing 780 transitions) according to the SocioCognitive skill fraction of the workers’ starting occupation (SocioCognitive_{ja}) and examining the average SocioCognitive change (i.e. \( \Delta \text{SocioCognitive} = \text{SocioCognitive}_{ja} - \text{SocioCognitive}_{ja} \)).
see fig.4-4.A) and average magnitude of this change (fig.4-4.B) for each bin. We consider workers selecting their new occupations at random as a null model for comparison. Workers transitioning from sensory-physical occupations tend to move to new occupations with a higher SocioCognitive skill fraction, but the magnitude of change is less than would be expected under random occupation selection (and vice versa for the other end of the spectrum). In contrast, workers transitioning from mid-quantile occupations, which represent starting occupations that effectively use SocioCognitive and sensory-physical skills evenly, exhibit larger magnitudes of change in SocioCognitive compared to the null model. In conclusion, workers of occupations relying strongly on one skill community tend to move towards other occupations within the same skill community, thus validating the second prediction.

For our third prediction, first note that the definition of skill complementarity [31] indicates increasing returns of combining skills within each skill community. Therefore, skill communities may be explained by the easy acquisition of related skills or by production efficiencies offered by workers who have complementary skills. However, this also means that workers relying on sensory-physical skills will face difficulty acquiring SocioCognitive occupations because they are unprepared to exploit large proportions of the SocioCognitive skills. Until they have a sufficient proportion of SocioCognitive skills, sensory-physical workers are bottle-necked by the polarized structure of skill complementarity. If this is true, then we expect disproportionately high employment in occupations under some threshold of SocioCognitive.
Figure 4-4: The polarized skill network constrains worker mobility. Binning by the \( \text{SocioCognitive}_j \) of the worker’s occupation in 2014 reveals the (A) expected SocioCognitive change and (B) the expected magnitude of SocioCognitive change when workers change occupations. Random occupation selection is considered as a null model (gray). Standard error bars are provided, but are small. Actual occupation transitions are provided as examples in (A). (C) The national distribution of employment by \( \text{SocioCognitive}_j \) with the distribution of individual occupations as an inset.

Indeed, binning national employment according to \( \text{SocioCognitive}_j \) yields a tri-modal distribution (see fig.4-4.C). The upper and lower modes of the distribution correspond to workers who are effectively exploiting the skill complementarity within each of their respective skill communities. The presence of a third mode in the middle suggests that skill polarization constrains workers from obtaining attractive SocioCognitive skills. This confirms the third prediction, yielding more evidence for
our hypothesis that the network of skill complementarity constrains labor mobility.

Finally, Figure 4-5 quantifies the average complementarity score of each skill as an approximation of that skill's network embeddedness. Considering our hypothesis and the strong relationship between skill proximity and skill acquisition, network embeddedness should correlate with increased labor mobility (individual skills are shown in Supplementary Materials Fig. S6).

![Skill Categories That Increase Mobility](image)

Figure 4-5: The connectivity and embeddedness for each skill category (by averaging the zscores for the PageRanks that each skill possesses in each skill category). The measure corresponds to worker mobility because skill proximity is indicative of skill acquisition. This highlights the importance of social skills for labor mobility. For a detailed result of all skills, see fig.8-2.

This demonstrates how SocioCognitive skills (social skills in particular) have the highest complementarity scores, which we showed is linked to labor mobility. Combined with figure 4-6, which shows how SocioCognitive skills also have superlinear growth rates with city size (i.e. population size), the results show how larger cities have a higher concentration of jobs that allow for better labor mobility.
Figure 4-6: The growth rate exponents for the two skill clusters (SocioCognitive vs. sensory-physical). This demonstrates that the occupations that have SocioCognitive skills grow superlinearly with city size and therefore, the presence of such skills in cities also grows superlinearly with city size.

4.6 Online Supplementary Materials

We also provide an online interactive tool to explore occupations and urban workforces on the SkillScape at skillscape.mit.edu.

4.7 Conclusion

After discussing the motivating issue of the impact of automation on large segments of the labor market, we demonstrated that the raw topology of the network corresponds to pathways along which labor dynamics can occur. Specifically, the “network proximity” between skills predicts i) skill adaptation in cities, ii) skill redefinition of occupations, and iii) the changing skill requirements of individual workers as they transition between occupations. Finally, combining these observations of skill polarization with the labor dynamics determined by the network topology highlights the importance of skills’ complementarity value, and indicates that skills that are more
"central" and connected are more valuable. Thus, we quantify the complementarity boost that each skill possesses, demonstrating that social skills have the best results by far, which is valuable considering our results. This high-resolution framework for understanding workplace skill requirements provides policy-makers with a new explanation for hindered career mobility, while also providing a tool to workers and urban planners trying to traverse the space of workplace skills.

However, some questions remain, such as the important question of why: not only why social skills are important for labor mobility, but also why they grow superlinearly with city size, as can be seen in figure 4-6. We attempt to answer this in the next chapter, where we address the question of social tie formation and the benefits that these ties bring to individuals, and to society as a whole. Regarding the limitations of the work, there are many. One important limitation is the assumption of causation; that is, given that we demonstrated that the skill structure is predictive of labor mobility, we assumed that skills with higher connectivity should therefore be sought after. However, this assumption is certainly a leap, and should encourage future research to study whether or not a causal link indeed exists. In addition, while our analysis identifies the specific skill requirements of low- and high-skill occupations that characterize occupational polarization, our work does not reveal whether occupational polarization is a result of skill polarization, or vice versa. In fact, many external factors, such as automation [75, 24] and off-shoring, likely contribute to both effects. Nevertheless, the SkillScape comprehensively explains the polarization of high- and low-skill occupations as a separation between workers with SocioCognitive skills and sensory-physical skills.
Chapter 5

Social Ties

In previous chapters, we presented numerous results demonstrating the importance of social skills in labor markets across a wide variety of circumstances. One of the most important results was describing how social skills are useful to individuals’ labor mobility (see fig.4-5), in both small and large labor markets. In addition, as cities become larger, their labor markets superlinearly gain social skills (see fig.4-6). This result is novel in the domain of labor markets, however, the general result of superlinear growth for social ties in cities is a well-known fact [90], not to mention the superlinear growth of many other socioeconomic indicators [25].

The formation and evolution of social ties is a fundamental but poorly understood problem in the field of social networks. There are multiple important theories, such as those by Sim et al. [100] and Pentland et al. [90], that focus on generative models that explain social ties’ superlinear growth, while others, e.g. that of Hausmann et al. [57], work on complexity to explain why cities have this observed feature. Several theories have intended to explain the mechanism of social ties’ formation, among which homophily and heterophily are two important accounts. Therefore, this chapter presents a computational model considering both homophily and heterophily to explain the dynamics of social networks, and hence explain some of the observations in previous chapters.
5.1 Introduction

Contradictory archaeological accounts of the change in cultural traits range between the change consistently increasing and the change being punctuated, with long periods of stasis interspersed by sudden gains or losses [69]. Nonetheless, all accounts agree that overall, change is positive, with an exponential accumulation of scientific knowledge [73].

Models of cultural evolution utilize a multitude of methods to explain the observed growth, such as social learning combined with the differential distribution of traits among social groups [69]. One important feature is the combinatorial process of merging different inputs to produce a novel output, which is one of the main sources of innovation, i.e. serendipity, incremental improvement (fig.5-1), and recombination [84, 61]. This is made clearer by studies that take an in-depth view of the “footprints” of innovation: patents [106].

One consistently critical component across all works in the literature is social ties as conduits to facilitate idea transmission. In fact, some have argued that all innovations are an emergent property of cultural learning critically through social networks, instead of being an output of a talented few. However, the formation and evolution of social networks is a fundamental but poorly understood problem in social network analysis. Large-scale data sets from widely adopted communication technologies, such as cellphone communications, have opened up new possibilities for studies of social networks. This chapter presents a computational model \(^1\) for social network formation which is rooted in social theory, in addition to leveraging large-scale data sets to test the effectiveness and prediction ability of the model.

---

\(^1\)This work was in collaboration with Yuan Yuan and is to be published in “Yuan Yuan, Ahmad Alabdulkareem, and Alex Sandy Pentland. Trade-Off Between Social Exchange and Coordination. 2018.”
5.2 Network Embedding Framework for Modeling Network Tie Formation

Due to the importance of social networks in societies, serving as conduits to exchange various benefits (e.g. innovations, information, and support [95, 19, 53]), they have received a great amount of attention and study. For instance, they have attracted the interest of scientists from a wide variety of fields, ranging from physicists to social scientists [88, 104, 64, 30], leading to important implications ranging from improving the social welfare to political participation [37, 28]. In addition, they are also essential in the study of macro scale phenomena, such as social polarization and contagion [51, 10, 82]. Even though the various approaches that are used for modeling social tie formation in networks are numerous, they all suffer from different drawbacks.
For example, various approaches ranging from game theory to agent based modeling, model the micro-level dynamics of individuals (deeply based on social and economic theories), however they are can only study and compare the macro-scale characteristics of the network. While on the other hand, network embedding approaches capture the micro-scale features of individuals, which produce testable predictions for the captured information about the modeled individuals’ in the network. However, they are orthogonal to any specific theory, at least not explicitly. They do however implicitly capture the tendency for homophily, by rightfully assuming that individuals connect with similar individuals, but those approaches are not easily extendable for the addition of other social theories.

The proposed model on the other hand approaches the problem of modeling social tie formation from a similar mindset to network embedding algorithms, in the sense that it tries to infer the features of individuals that would have lead them to being connected to every one of their neighbors in the network (while not connected to the others).

\[
\text{Social Tie}(i,j) = \begin{cases} 
1 & \text{if } u(v_i, v_j) > 0 \\
0 & \text{otherwise}
\end{cases}
\]  

(5.1)

Where Social Tie(i,j) refers to whether or not individual ‘i’ has a tie with individual ‘j’, with \( u(v_i, v_j) \) referring to the utility of that tie. The utility is calculated based on some functional form that compares the two individuals’ feature vectors (\( v_i \) and \( v_j \) respectively) based on some social theory (or theories). Notice that if the utility is not strictly positive, then there should not be a tie in that situation\(^2\).

However, the most critical difference of our model that sets it apart from other network embedding algorithms is that it approaches the problem of inferring the individuals’ feature vectors using an equation that is based on some social theories, and allows for future research to add and edit as needed. It does so in its utility function, for example, we picked two main drivers for social tie formation:

---

\(^2\)The reason for the ‘otherwise’ situation to include the case of \( u(v_i, v_j) = 0 \) is to disallow self ties.
The rational for these theories comes from various works in the literature which argue for social ties being largely driven by these two drivers. For example, the work of Alpert et al. [6] demonstrates the results of their empirical analysis and findings (see fig. 5-2) where they highlight the fact that the optimal case is a balance between heterophily and lack thereof (i.e. homophily).

Figure 5-2: Figure recreated from [6], where they studied communication effectiveness. This plot demonstrates the results of their empirical analysis and findings where they highlight the fact that optimal case is a balance between heterophily and lack thereof (i.e. homophily).

As for the specific functional form, there are many versions that we could choose for each component, however some assumptions are made to come up with an example functional form to test out our model. For example, by assuming that the benefits part of the utility function would only be concerned with positive gains in the differences of the two feature vectors, that would make the ReLU function a perfect candidate (i.e. the rectified linear unit, or ‘max(input, 0)’). As for the cost part (which is driven
by homophily drivers), it is a reasonable assumption that small difference across the many dimensions of a feature vector are not as important as one large difference in one dimension (e.g. small differences in the race, language, and ethnicity are more aligned with homophily than a large difference in one of them). Hence the square of the second norm would insure that a higher penalty is given to larger differences in single dimensions.

\[ u(v_i, v_j) = \sum_{k \leq K} b_k \text{ReLU}(v_{jk} - v_{ik}) - \left\| c \circ (v_j - v_i) \right\|_2^2 \]  

(5.3)

Where \( \circ \) represents the element-wise multiplication between vectors. While \( b \) and \( c \) are scaling parameters measuring the importance of dimensions for benefits and costs, respectively.

As for the execution of our model, we fit our parameters to any real life social network by optimizing the inferred individuals’ feature vectors \( (v'_i) \) such that they would form the exact same social network if they were simulated using ABM, with the utility function \( (u(v_i, v_j)) \) as their social tie forming mechanism (see fig.7-1).
5.3 Data Set

Even though we tested our approach on a wide variety of data sets, this example data set demonstrates the strengths and weaknesses of our model at the same time. The data set is a collaboration network of cast or crew members, named The Movie DataBase (TMDB), which is publicly available\(^3\). We extract the actors/actresses along with directors, and then establish links between the director and the cast based on co-occurrences in movies, meaning that “the director invited the actor/actress to the collaboration and the actor/actress agreed”. We extracted 3,493 movies from 2000 to 2016, and kept individuals with at least 5 movies within the period, resulting in 799 directors, 2,789 actors/actresses, and 33,808 director-cast pairs.

\(^3\)https://www.kaggle.com/tmdb/tmdb-movie-metadata
5.4 Results

Even though our objective was not to compete with the state-of-the-art in terms of pure prediction ability, since our goal was interpretability and being rooted in social theory with flexibility for future research to easily extend the approach with additional or different theories, we should still use the state-of-the-art as a measuring stick for our performance.

As we can see, fig. 5-4 demonstrates the results of using the feature vectors produced by the proposed model vs DeepWalk for predicting the occupation and gender of individuals in the TMDB network. Even though our model was tested on a variety of data sets, these results highlights both homophily and heterophily at the same time. Since our proposed model, which considers both drivers, outperforms DeepWalk in predicting occupations (reasonably assumed to be more driven by heterophily), while it underperforms in Gender prediction which is assumed to be more driven by homophily (which DeepWalk is purely focused on).

Figure 5-4: This figure demonstrates the results of using the feature vectors produced by the proposed model vs DeepWalk for predicting the occupation and gender of individuals in the TMDB network. Even though our model was tested on a variety of data sets, these results highlights both homophily and heterophily at the same time. Since our proposed model, which considers both drivers, outperforms DeepWalk in predicting occupations (reasonably assumed to be more driven by heterophily), while it underperforms in Gender prediction which is assumed to be more driven by homophily (which DeepWalk is purely focused on).
5.5 Conclusion

This chapter has presented some additional evidence for the value of social ties to explain some of the results of previous chapters. By proposing a computational model of social network formation that would balance both homophily and heterophily drivers. Distinct from network embedding algorithms, which mostly have no social theory or only implicitly assume homophily, we have demonstrated the value of an interpretable computational model which explicitly considers the importance of social exchange as an key factor for social network formation. The proposed model calculates the high dimensional feature vectors for each individual in a way that would maximize their chances of forming a tie based on the utility function. Thus, it is different from most models in complex systems which compare simulation results and empirical data sets from a macro-scale view. To demonstrate our models ability for capturing real information in the inferred vectors, we predict various real individual features (see figure 5-3 for a high-level overview). We have tested this approach on a wide variety of synthetic and real large scale data sets. As an example of the model deployed for prediction in a real life large scale data set, we attempted to capture the features of individuals in a social collaboration network data and then accurately predicting occupation types.

There are many elements that require future research to address, for example, there are various additional social theories that can be added or even replace the ones that are considered in our model. As an example, monophily [8] (which refers to the extreme preference of an individual for a specific feature that they do not possess) has been shown to also be a driver in some social tie formation. The interesting feature of this driver is that it creates similarities between the various friends of an individual but which are at the same time dissimilar from the ego-individual across those relationships. Another example limitation is the lack of meeting probabilities in the model, which refers to the fact that we do not consider the randomness of interaction probability, which is an important precondition for a tie to be formed. This is a factor that many works in the literature consider [81, 38, 98]. However,
even though we do not explicitly consider meeting probability in our model before the utility function is applied for tie formation, the model is implicitly able to encode those factors into the feature vectors of each individual (e.g. location). This is the case because it is a reasonable assumption to say that homophily is related to the meeting probabilities (i.e. similar people have higher meeting probabilities).
Chapter 6

Social Privacy

This chapter demonstrates the importance of social privacy, then presents some approaches to handle the release of social information in a privacy-preserving way. First, previous chapters have shown how social ties can be used to predict a wide set of socioeconomic indicators about an individual. We start this chapter by demonstrating that same result in a new setting (i.e. malicious activity on Twitter). An important feature of the previous chapters was the use of the information of some individuals in the network to predict their neighboring nodes' information (i.e. predicting one’s friends’ information using one’s own). Therefore, we study the important question of whether or not people care about affecting each other’s privacy, and whether they care about it more than their own. Lastly, we attempt to provide an improvement to a privacy-preserving algorithm that maintains its privacy guarantee while increasing its utility. Here, we make the distinction between privacy in the sense of providing a level of obfuscation, and utility, which is the useful information that can be retrieved from the distorted data set for purposes such as prediction or learning. Using a method called eigenbehaviors combined with differential privacy applied on real datasets, we demonstrate that it is possible to achieve increased levels of utility for machine learning algorithms while satisfying the same levels of privacy, thus highlighting the importance of making the distinction between the utility of the released data and its privacy. The objective is to highlight the importance of approaches that consider behaviors when developing privacy-preserving data release mechanisms.
6.1 Introduction

Digital information about users is undoubtedly the oil of the new economy. Due to the wide-spread adoption of digital and pervasive technologies, the collection of large-scale, longitudinal personal data – data generated by the people and about the people [97] – is now ubiquitous. Our mobile communication records, GPS location traces, credit card transactions, web-page searches, and online social networks, to name a few examples, can generate high-resolution data about who we know, where we are, what we like, and how much we spend. Consequently, personal data has the potential to transform our understanding of human society and is considered to be the oil of the online economy [97, 71]. While the collection, processing, and leveraging of users’ data are the fuel that powers many online services, they also raise severe concerns for users’ privacy.

However, when longitudinal data pertaining to a single individual are gathered, temporally reoccurring patterns emerge. Methods of extracting those behaviors from the data such as eigenvector decomposition produce what are hereafter referred to as eigenbehaviors [46]. Using the strong signaled eigenbehaviors you guarantee that the typical behavior of the entire data set is preserved for whatever analysis purpose you want. In addition, it guarantees that future data set added to the existing data will be “similar” without too large of a divergence. The reason being, that the noise is a random displacement from the broad dynamics of the entire data set rather than a random displacement from the zero vector. Thus behavioral privacy is a much more “well behaved” method for practical deployments.

6.2 Motivation

Today, users share large amounts of information about themselves in their online social networks. Besides the intended information, this sharing process often also “leaks” sensitive information about the users and, by proxy, their peers. This section investigates the effect of awareness about such leakage of information on user behav-
ior. In particular, taking inspiration from “second-hand smoke” campaigns, this study creates a “social awareness” campaign where users are reminded of the information they are leaking about themselves and their friends. The results indicate that the number of users disallowing access permissions doubles with the social awareness campaign as compared to a baseline method. The findings are useful for system designers considering privacy as a holistic social challenge rather than a purely technical issue.

The growing popularity of online social networks (OSNs), such as Facebook, Twitter, and LinkedIn, has made them integral parts of contemporary online activities. Although OSNs are widely used and represent a rich source of information, much of their data is also sensitive and personal (e.g., demographic, interests, etc.) [35]. OSN users usually disclose personal information in order to participate in social communities or in return for services [70]. However, disclosing personal information in this case can be a double-edged sword. For example, such exposure might make the user vulnerable to personalized attacks such as stalking, identity theft, reputation slander, personalized spamming, and phishing. While most OSN services offer various levels of privacy protection (e.g., allowing only an authorized list of other OSN users, applications, third parties, etc.), users’ information may extend beyond the defined bounds, which in a privacy context is referred to as information leakage [21].

Using Twitter data that was collected over two months, fig.6-1 demonstrates that the strongest indicator of anomalous activity is social ties (in this case, normalcy vs. spammer). This is yet additional evidence (in addition to the results we presented in previous chapters) that social ties leak information about ourselves and our friends.

Information leakage is the phenomena where explicit information provided to a third party can be used to derive implicit and previously hidden information about an entity (e.g. as in previous chapters, we leak information about our friends). Much of the literature suggests that some reactive measures be taken to minimize the effects of information leakage. Such measures include suggesting some friends to “unfriend” to minimize the amount of leakage [60]. While later sections will provide alternative solutions, in this motivational section, we would like to suggest one approach to addressing this issue in a preemptive way rather than in a reactive way. Since the
Figure 6-1: The strength of different indicators for the prediction of anomalous behavior. Results demonstrate that social ties are better predictors than “tweet” or profile contents.

user is the entity in charge of such decision, we want to test whether informing the user before sharing personal information can minimize this effect preemptively. In addition, we examine the extent to which peer pressure influences users’ behavior. The experiment goal is to investigate how different users behave when they know that they are leaking sensitive information about themselves or their peers and how this affects their decisions. The study is conducted on an online social network platform with over 200 participants. The results show that users are more responsive to the peer pressure variable, which suggests that if users consider their peers when making their online privacy decisions, they will probably leak less information and increase their privacy level.

The main question that this study aims to answer is whether knowing the implication of sharing pieces of information may make users change their behavior. Specifically, in this section, we try to answer the following questions:

I. If the user is presented with a visualized numerical quantification of the amount of information leaked, does this affect his/her behavior? II. What are the differences in
Figure 6-2: The three messages shown to the three different groups, Control, Self Information Leakage, and Social Information Leakage Group, respectively.

user behavior when informed about leaking information about themselves, as opposed to leaking information about their peers? This is especially interesting because: a) it grants the user a sense of “agency”, and b) it brings out the effect of direct and indirect peer pressure on user behavior. Previous results of smoking campaigns (e.g. second-hand smoke affects our loved ones) and healthy behavior adoption [29] have suggested an impact of social peer pressure on user behavior.

To answer these questions, the experiment is designed such that users are randomly assigned to one of three groups. Control Group: This group is presented with only a text-based message shown as a typical terms and conditions page. This group acts as the baseline control. Self Information Leakage Group: This group is presented with both a text message stating that using the app will result in leakage of some of the user’s own sensitive information, along with a visual spider graph to emphasize the before/after effect of leaking the user’s information. Social Information Leakage Group: This group is presented with a text message that states that using the app will result in leakage of sensitive information about the user’s friends, and similar to the previous group, a spider graph is shown to visualize the amount of leakage. However, this spider graph now states that the user’s friends’ information will be leaked, alongside profile pictures of the user’s friends that are pulled and displayed as part of the social message to add social pressure and test the peer effect on the user’s behavior. All users in all groups have the same two options: either to accept and proceed to the app page, or to decline and exit the app.

Due to its popularity, Facebook is the chosen platform in this experiment. Each
user who participates in the Facebook app experiment is assigned to one of the three user groups presented earlier. For each participant, there are two measured variables: (1) the user’s action when presented with the information leakage alert message; and (2) how much time it takes the user to respond to the message.

First, we present the decision analysis results, focusing on the differences between each group in terms of the users’ response to the self information leakage (SIL) message. As can be seen in figure 6-3, the results show that in Group 1, 91% of the users agree to the terms and conditions page and thus respond with Yes to grant the app access to their data, while 9% respond with No. Similarly, in Group 2, 90% of the users respond with Yes knowing that the app will leak some of their personal information, and only 10% respond with No. However, in Group 3, when presented with their friends’ pictures, 19% of the users deny the app access and respond with No. The results show that the third group behaves differently than the other two groups. This indicates that peer pressure has a strong effect on users’ decisions. Figure 6-3 shows the percentages of each decision per user group.

Secondly, the decision-time analysis focuses on studying the differences between user groups in terms of the time spent making a decision regarding the SIL message. Again, the results show that users in Group 3 behave differently than users in the other two groups. As can be seen in figure 6-3, the median total time elapsed to make a decision is 11.0 seconds for users in Group 3 (either yes or no), 8.0 seconds for users in Group 2, and 6.3 seconds for users in Group 1.

We can see that in both kinds of analysis, Group 3 behaves differently than the other two groups. In addition, Group 1 and Group 2 are fairly similar in time taken to make their decisions. Thus, the quantification and visualization of information leaked does not seem to have any major effect on users’ decisions and behavior, which answers the first question. Moreover, when users are informed that they are leaking sensitive information about themselves, they are more willing to proceed with the app, as opposed to when their friends or peers are affected. This behavior is consistent in both the decisions they take and in their taking more time to think when they are aware that their decision will affect their friends.
Thus, knowing that one is compromising one’s friends’ privacy affects one’s decision-making process. The analysis shows that peer pressure influences users’ behavior. To the best of our knowledge, this is the first attempt to study peer pressure and its effects on users’ behavior with regard to privacy. We hope the presented results will bring useful insights for policy and application design of future social applications with respect to privacy.

### 6.3 Background

There are numerous types of privacy-preserving mechanisms beyond the simple removal of explicit identifying information. When the aim is to release data, non-interactive privacy mechanisms are appropriate; otherwise, interactive mechanisms might fit (Q&A types of frameworks). Non-interactive privacy methods (the case in this thesis) maintain privacy by either removing some parts of the data (anonimiza-
tion), or distorting some or all of the data (adding noise). Through generalization and suppression, at first glance anonymization might seem like a highly promising approach. However, although there are numerous ways of anonymizing personal data, all models face various critical limitations [44]. For instance, the $k$-anonymity model makes every record in a given table indistinguishable from at least $k - 1$ other records, thereby making it impossible to identify an individual in that table [102]. Variations and alternatives of $k$-anonymity include $\ell$-diversity [76, 77], which prevents attacks that use a lack of diversity in the sensitive data, and $t$-closeness [33, 74], which aims to maintain the distribution of sensitive data. The reader is referred to the surveys [2, 55] for further details.

Even though all of these models aim to minimize privacy risks while keeping data utility as high as possible, they often mix privacy and utility definitions, where both focus on the raw data points. In addition, all of these approaches were designed with relatively low-resolution data in mind, and they are not easily extendable to high-dimensional data, such as GPS location or accelerometer readings. Therefore, in recent years, there have been several technological and legal movements towards a new practical way of protecting individuals' privacy and regaining control through personal data stores (PDS) [22, 83, 44]. Such distributed architectures allow individuals to collect, store, and safely share personal data under their control. For example, OpenPDS and SafeAnswers are new ways of dynamically protecting personal data. OpenPDS is a “personal data management framework that allows individuals to collect, store, and give fine-grained access to their metadata to third parties” [44]. SafeAnswers is a practical way of protecting the privacy of personal data at the individual level, by providing the ability to answer services’ questions by directly calculating it against the raw personal data, instead of trying to anonymize it. Only the answers to the questions (that are significantly lower-dimensional) are shared back with the services, and therefore are unlikely to be re-identifiable or contain sensitive pieces of information.

The other categories of privacy mechanisms tackle the problem by adding noise, which can be done in a multitude of ways. The approach of choice in this chapter
is differential privacy, a state-of-the-art privacy approach that uses a parameter $\epsilon$ to specify the amount of controlled random noise that is added. The advantage of differential privacy is that it guarantees a level "$\epsilon$" of privacy, regardless of what an adversary has as side information (which he/she hopes to use to make released data as the last breadcrumb required to breach privacy). While most other approaches (including anonymization) make explicit or implicit assumptions about the adversary's side information (information an attacker attained elsewhere which he/she uses to de-anonymize/attack the data), these assumptions and/or restrictions cannot be guaranteed in most if not all cases. In contrast, differential privacy assumes nothing and mathematically guarantees that an adversary will not gain any "bonus breadcrumbs" due to having any side information. Differential privacy is more of a mathematical constraint on a privacy algorithm than one specific algorithm, as stated by Cynthia Dwork et al. [43].

6.4 The Friends and Family Data Set

To test our method, we utilize the Friends and Family ("FunF") dataset [3] (see figure 6-4) pertaining to over 100 individuals, which contains immensely rich and dense information over a period of a year-long study. The FunF community includes a more diverse subject pool than previous social computing observation studies [47, 78]. FunF is essentially a passive sensing software which participants ran on their mobile phones, continuously collecting information (e.g. location, device proximity using Bluetooth, calling information, and dozens of other metrics). In addition, participants also completed surveys at regular monthly (e.g. relationships, interactions, and Big-Five personality test) or daily intervals (e.g. happiness, stress, and sleep). For each individual, we chose different portions of his/her data on which to test our methods (e.g. Bluetooth data of proximity to other individuals or number of phone calls in each hour), with two target attributes (happiness level and stress level on any particular day based on that day's data). These classes were calculated by detecting whether a participant’s daily survey response regarding his/her happiness and stress was above
Participants

<table>
<thead>
<tr>
<th>Participant Degree</th>
<th>40</th>
<th>20</th>
<th>3</th>
</tr>
</thead>
</table>

Clustering Algorithm

<table>
<thead>
<tr>
<th>Friendship Type</th>
<th>Unilateral</th>
<th>Reciprocal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendship Score</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

Figure 6-4: Snapshot of the social network of the Friends and Family data set. We can see many features, such as stronger and more bidirectional ties within cliques than across the clusters that were identified through the clustering algorithm (which is stronger than random, and visually apparent).

or below average, thus creating the two classes of happy/stressed or not, respectively.

6.5 Materials and Methods

Differential Privacy for Personal Data Stores

In this section, we show how the well-known differential privacy measure can be modified to fit the case of personal data stores. We start by describing the traditional differential privacy measure:

Definition: A mechanism $M$ satisfies $\epsilon$-differential privacy if for every pair of databases $X, X'$, and for every subset $S \subseteq \text{Range}(M)$, and every query $q \in Q$:

$$\frac{Pr[M(X, q) \in S]}{Pr[M(X', q) \in S]} \leq e^{\epsilon|x \oplus x'|} \quad (6.1)$$

However, this definition is designed for centralized settings, where a database contains data about a large number of individuals and answers are aggregate across
Figure 6-5: Probability distribution for 10,000 two-dimensional points from three different one-dimensional laplacian noise generations at different locations ($\epsilon = .3$ (blue), $\epsilon = .4$ (green), & $\epsilon = .5$ (red)). The middle inset plot is a scatter plot of the generated noise points, while the left plot represents many probability density functions for the noise points in two dimensions, and the right plot is a two-dimensional heat-map for the noise points.

these individuals. Unfortunately, this makes the definition poorly suitable for protecting the privacy of individuals in a PDS architecture, where one data store holds data about a single individual. Inspired by the traditional differential privacy, *geoindistinquishability* was suggested in the literature as a mechanism for adding noise (e.g. see figure 6-5) to the location of a single individual, while maintaining specific privacy guarantees and without compromising the quality of the application results too much [9]:

**Definition**: A mechanism $M$ satisfies $\epsilon$-differential privacy if for every pair of data points $x, x'$, and for every point $s \in \text{Range}(M)$:

$$\frac{\Pr[M(x) = s]}{\Pr[M(x') = s]} \leq e^{\epsilon \cdot d(x, x')}$$

(6.2)

While this definition represents a great improvement for the use of differential privacy in PDS environments, it is still highly limited as it was designed for location data only. To address this issue, the line of work in Andres et al. [9] sets up the framework and creates a two-dimensional version of differential privacy that operates on single personal data stores of geo-location. In PDSs, however, the data elements consist of many dimensions (e.g. location, number of events “a”, number of events “b”, 

101
time), and therefore the distortion method has to be able to adapt to the appropriate number of dimensions. Fortunately, this definition can easily be extended to higher-dimensional data. To extend the distortion to higher dimensions while maintaining the validity of the approach, we use N-Dimensional Quasipolar Coordinates[89]:

$$\begin{align*}
  x_1 &= r \times \Pi_{k=1}^{n-1} \sin(\theta_k) \\
  x_i &= r \times \cos(\theta_{i-1}) \times \Pi_{k=1}^{n-1} \sin(\theta_k) \\
  x_n &= r \times \cos(\theta_{n-1})
\end{align*}$$

where \( r \) is the amount of noise distortion that is to be projected into an \( n \) dimensional space (generated by a laplacian distribution, as most differential privacy algorithm implementations do [80]). While \( \theta_1, \theta_2, ..., \theta_{n-2} \) are angular coordinates ranging over [0, \( \pi \)) radians, \( \theta_{n-1} \) ranges over [0, 2\( \pi \)). This creates a set of \( x_1, x_2, ..., x_n \) variables that are the Cartesian coordinates of the generated distortion after the transformation into euclidean space.

**Behavioral Privacy**

When longitudinal data pertaining to a single individual are gathered, temporally reoccurring patterns emerge. Methods of extracting those behaviors from the data such as eigenvector decomposition produce what are hereafter referred to as eigenbehaviors [46]. These eigenbehaviors capture different amounts of information about the data, and using them as a bases of reconstruction demonstrates how much information they hold.

We suggest a new method that includes the behavioral aspect in the differential privacy process, henceforth called "behavioral privacy", see fig.6-6. The behavioral privacy method excludes eigenbehaviors that contain very low amounts of information and maintains only the top \( k \) eigenbehaviors (the choice of \( k \) changes from person to person depending on the eigenvalue associated with the eigenbehaviors, which represents the amount of information they contain).
6.6 Results

We apply the state-of-the-art differential privacy distortion method and behavioral privacy method on the data sets, producing two different versions of distorted data. When executing both methods, we choose the same $\epsilon$ values for both methods' differential privacy component. We do this to guarantee the same level of privacy. To evaluate the contained information remaining intact in the methods' distorted outputs (as a proxy for utility), we cross-validate (10-fold) different machine learning algorithms on the distorted data sets and record the area under the receiver operating characteristic (AUROC) curve (see fig.6-7).
Figure 6-7: Approach for comparing utilities of the state-of-the-art vs. the proposed behavioral privacy method.

Each execution applies six different machine learning algorithms attempting to classify the data elements based on the target variables. For each of the hundreds of individuals in the data sets, we test the effects of applying the state-of-the-art approach vs. the behavioral privacy method for each permutation (individual id, Bluetooth/calling data, stress/happiness levels, and which machine learning algorithm (ML) will consume the outputted data), producing a single output, fig 6-8.

Since both methods apply differential privacy, they produce different outputs for varying inputted $\epsilon$ values shown on the x-axis (the variable controlling the amount of distortion/privacy level). Due to differential privacy being stochastic, many executions for each $\epsilon$ value are run for both methods (50 times in the present results) and the results are aggregated as means and their standard error intervals (darker shade...
Figure 6-8: The state-of-the-art method (red) vs. the behavioral privacy variant (blue), using different privacy parameters (x-axis) and then applying a linear SVM on the resulting distorted data sets. We show the resulting utility measures (AUROC values) on the y-axis. Inset plots represent the kernel density estimator (KDE) and cumulative distribution function (CDF) for the y-axis when aggregating over all of the x-axis.
representing 1*standard error and lighter shade representing 2*standard errors). For each output figure, and for each x-axis tick, it is not sufficient to see which method has the superior mean. Instead, a Welch’s t-test is preformed to check whether there is significant evidence to reject the null hypothesis of equal means and therefore provide more evidence for the higher mean pertaining to a statistically superior method within that specific instance. Fig 6-8 demonstrates how in the case of individual FA-17, the behavioral privacy method greatly outperforms the standard state of the art (raw data + noise) for low amounts of distortion with statistical significance. However, for most if not all individuals (including FA-17), choosing a very high level of privacy (low ε) makes the two methods score a similar AUROC of 0.5. It is useful to note that when 1/ε → ∞, the output of both methods almost becomes completely useless (an AUROC of 0.5, which is random guessing). Fig 6-8 also demonstrates similar results on a different permutation of user SP-41 and features (Bluetooth data predicting stress levels).

A final note before the results are discussed concerns an unfair advantage that the behavioral privacy method has over the state-of-the-art in this experiment: by definition, it is infused with behavioral mining techniques (dimensionality reduction), which would give it a boost in the ML results. Therefore, the results of both the state-of-the-art method and the behavioral mining method are treated by dimensionality reduction (keeping the top “K” eigenbehaviors) in their output matrices.

Now that we have shown how experiments are carried out on each individual separately with the different possible permutations of data types, predictors, and ML approaches, we aggregate the outputs to generalize the results. As we saw, for each individual we calculated the different p-values of comparing the different methods with varying ε values. Now, we treat each p-value as either statistically significant (blue or red depending on whether behavioral privacy or the state-of-the-art is superior, respectively) or non-statistically significant (green) and tally up the results in fig 6-9. Note that the non-statistically significant bars are not shown, but can be calculated by taking 1-(blue + red).

The results in fig.6-9 demonstrate that in reasonable amounts of noise (1/ε ≤ 1),
behavioral privacy is superior to the state-of-the-art method, with statistical significance in around 5% to 20% of the datasets. However, in increased amounts of noise, both methods appear to be almost exactly the same (note that the percentage of non-statistically significant results are not shown, but are the complement to the sum of the shown bars). This indicates that if the objective of releasing such data sets was
their use in similar ML algorithms, the behavioral privacy approach would be more desirable since it would produce better results (i.e. higher utility, with the same level of privacy).

### 6.7 Conclusion

This section demonstrated some new motivational reasons for the importance of social privacy, and also endeavored to suggest new ways of thinking about the problem. That is, it focused on decoupling the objectives of privacy from utility, and showed positive results of such decoupling using an example case study of behavioral privacy. The objective was not to suggest the adoption of this method (since utility in this case was solely focused on machine learning), but to demonstrate the value of pursuing approaches that decouple utility and privacy, achieving higher levels for the former while maintaining the same level for the latter.
Chapter 7

Conclusion

7.1 Conclusion

In this thesis, using data-driven tools from computational science and machine learning, the first question we aimed to answer is what role do social skills play in today’s labor markets on both a micro and macro scale (e.g. individuals and cities). We addressed this question by constructing a network of skill complementarity and using it to demonstrate the value of social skills in various results (see Figure 7-1), such as their relationship to economic well-being for both individuals and cities. Second, given the current important topic of automation, we investigated the question of its negative effects and possible methods to combat them. Using the structure developed in the previous chapter, we predicted various labor dynamics, ending with the important role that social skills play in combating those effects, specifically in relation to career mobility. The goal is to inform strategies to mitigate the negative effects of automation and off-shoring on employment; therefore, we also provided an online interactive tool for exploring skills of occupations and urban workforces on the SkillScape.
The Importance of the Social Dimension

Figure 7-1: A summary for the flow of the thesis, starting from the initial chapters that transformed matrices of individuals’ job features to construct the skill network, which demonstrated the importance of the social dimension in the labor force (for both economic well-being and labor mobility). Then, using a city’s social network, later chapters employed a mechanistic model to capture the value of social ties for social exchange by reconstructing the individuals’ features embedded in the social network.

The third question involved what theoretical model can explain the demonstrated importance of the social dimension in cities, and what are its consequences. To answer this, we developed a computational model of social network formation and dynamics, which we validated on small and large empirical datasets while providing some of the properties that were extracted using our model (see Figure 7-1). Finally, we highlighted some possible solutions to combat the issue of the demonstrated vulnerabilities for invading individuals’ privacy.

7.2 Going Full Circle

After we demonstrated the value of our proposed model, which balances the cost/benefit of new social ties, we can take a step back and link this approach to the work in the previous chapters of this thesis. Taking inspiration from the various speeds of differ-
ent network spreading mechanisms, we can hypothesize on the optimal cost-benefit trade-off of acquiring new skills (for individuals, occupations, or cities). As an illustration, Figure 7-2 demonstrates how cities with various sizes are at different stages of evolution for their skill portfolio. One can then ask the question, what skills should a city (or even an individual, for that matter) in one of those earlier stages acquire?

Figure 7-2: Cities with various sizes are at different stages of evolution for their skill portfolio. That is, smaller cities, on the left, mostly possess Sensory-Physical skills, while larger cities, on the right, possess mostly SocioCognitive skills.

For example, if currently acquired skills possess an extremely high automation probability, what would be the next skill to aim at acquiring? While many criteria can be used to tackle this question, we focus on the metrics that can be extracted from the network we used in early chapters. That is, the benefit would be to acquire skills that are more desired (i.e. well-connected skills that are more SocioCognitive), with the limiting cost being the new skill’s complementarity with the previously possessed skills (see Figure 7-3).

Figure 7-3: (A) An example current status of acquired skills in the SkillScape. This could be an individual, occupation, or city. (B) demonstrates the weighing of costs and benefits for various skill acquisitions.
Thus, in the Figure 7-3 example, the optimal choice of balancing the benefit and cost of acquisition would be the skill “S2” (provided that equal importance was assigned to both metrics). The justification for the importance of combining these metrics is that we have demonstrated how high-proximity skills empirically possess the highest probability of being acquired, thus indicating that they are usually a very viable option. At the same time, we have calculated which skills are more connected and central (e.g. social skills), which would open up opportunities for better and more numerous future skills. In addition, some analysis could be conducted regarding the salary value for each skill, which would be a crucial metric. However, given that the resolution of the skills in the current analysis is very low, this would overload the already extended approach with too many assumptions.
Chapter 8

Appendix; Supporting Information

Figure 8-1: Transforming raw O*NET data with RCA. The first plot is the raw occupation-skill matrix, $I(j, s)$, the middle plot is the RCA occupation-skill matrix, $rca(j, s)$, and the final plot is the thresholded RCA job-skill matrix, $e(j, s)$, for 2014. Here, $e(j, s) = 1$ if and only if $rca(j, s) > 1$. Occupations (y-axis) are ordered by the sum of threshold RCA skill values, and skills (x-axis) are ordered by the correlation of their thresholded RCA values across occupations to the occupational sums.
Cluster | List of skill labels according their community cluster
--- | ---

Figure 8-2: Full list of skills for figure 4-5 in the main text. PageRank scores for every individual Skill (node in the network). That is, the connectivity & embeddedness of each skill. Color represents skill category.
Figure 8-3: Continuing PageRank of the SkillScape Skills in figure 8-2.
Figure 8-4: Instead of PageRanks in figure 4-5 (which is an n-th order calculation), this is the first order calculation for comparison. This figure demonstrates complementarity scores for every skill category. That is, the average Z-score of each categories’ node strengths (sum of it’s edges, or “complementarity weights” $\theta$). Color represents skill category. We can see Social Skills still achieve the highest score.
Figure 8-5: Complementarity scores for every individual skill (node in the network). That is, the Z-score of each node’s strength (sum of it’s edges, or “complementarity weights” θ). Color represents skill category.
Figure 8-6: Continuing Node Strengths of the SkillScape Skills.
Figure 8-7: Instead of the interpolated plots in the main text (figure 4-3), here we provide bar plots with the associated error bars.
8.1 Predicting Economic Well-Being with SocioCognitive Skills

<table>
<thead>
<tr>
<th>Label</th>
<th>Industry Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupation 1</td>
<td>Management, Business, &amp; Financial Occupations</td>
</tr>
<tr>
<td>Occupation 2</td>
<td>Computer, Engineering, &amp; Science Occupations</td>
</tr>
<tr>
<td>Occupation 3</td>
<td>Education, Legal, Community Service, &amp; Arts Occupations</td>
</tr>
<tr>
<td>Occupation 4</td>
<td>Healthcare Practitioners and Technical Occupations</td>
</tr>
<tr>
<td>Occupation 5</td>
<td>Service Occupations</td>
</tr>
<tr>
<td>Occupation 6</td>
<td>Sales &amp; Office Occupations</td>
</tr>
<tr>
<td>Occupation 7</td>
<td>Natural Resources, Construction, &amp; Maintenance Occupations</td>
</tr>
<tr>
<td>Occupation 8</td>
<td>Production, Transportation, &amp; Material Moving Occupations</td>
</tr>
</tbody>
</table>

Table 8.1: Descriptions of each occupation type indicator variable used in regression models. For each occupation, the indicator variable is 1 if and only if the occupation SOC code belongs to that occupation category. Each occupation belongs to exactly one occupation category.

8.1.1 Predicting Annual Wages of Occupations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SocioCognitive_j$</td>
<td>0.387***</td>
<td>0.403***</td>
<td></td>
<td>0.372***</td>
<td></td>
</tr>
<tr>
<td>Occupation 1</td>
<td>0.490***</td>
<td>0.050</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation 2</td>
<td>0.663***</td>
<td>0.544***</td>
<td>0.970***</td>
<td>0.648***</td>
<td></td>
</tr>
<tr>
<td>Occupation 3</td>
<td>0.112</td>
<td>-0.203*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation 4</td>
<td>1.320***</td>
<td>1.251***</td>
<td>1.627***</td>
<td>1.351***</td>
<td></td>
</tr>
<tr>
<td>Occupation 5</td>
<td>-0.785***</td>
<td>-0.674***</td>
<td>-0.478***</td>
<td>-0.588***</td>
<td></td>
</tr>
<tr>
<td>Occupation 6</td>
<td>-0.575***</td>
<td>-0.837***</td>
<td>-0.268*</td>
<td>-0.722***</td>
<td></td>
</tr>
<tr>
<td>Occupation 7</td>
<td>-0.483***</td>
<td>0.036</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation 8</td>
<td>-0.582***</td>
<td>-0.108</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.000</td>
<td>0.160***</td>
<td>0.059*</td>
<td>-0.147***</td>
<td>-0.027</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.150</td>
<td>0.386</td>
<td>0.429</td>
<td>0.312</td>
<td>0.424</td>
</tr>
<tr>
<td>adj. $R^2$</td>
<td>0.149</td>
<td>0.380</td>
<td>0.421</td>
<td>0.308</td>
<td>0.419</td>
</tr>
</tbody>
</table>

Table 8.2: Linear regression using standardized $SocioCognitive_j$ for each occupation and occupation type indicator variables.

$p_{val} < 0.1^*, p_{val} < 0.01^{**}, p_{val} < 0.001^{***}$
Figure 8-8: Out of sample testing of model performance from Table 8.2. For each model, 1,000 trials are run where 75% of the data is randomly selected as training data and the remaining 25% of data is used as validation. The distribution root-mean-square errors for each model is reported. Medians are represented by a red line, while the mean error is represented by the green square.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.000</td>
<td>-0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>SocioCognitive$_j$</td>
<td>0.387***</td>
<td>0.355***</td>
<td></td>
</tr>
<tr>
<td>No B.D. Employment</td>
<td>-0.264***</td>
<td>-0.234***</td>
<td></td>
</tr>
<tr>
<td>B.D. Employment</td>
<td>0.216***</td>
<td>0.094*</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.150</td>
<td>0.090</td>
<td>0.203</td>
</tr>
<tr>
<td>adj. $R^2$</td>
<td>0.149</td>
<td>0.087</td>
<td>0.199</td>
</tr>
</tbody>
</table>

Table 8.3: Linear regression using SocioCognitive$_j$ and employment in each occupation with a bachelor’s degree (denoted B.D. Employment) and without a bachelor’s degree (denoted No B.D. Employment). Each variable has been standardized. Employment by level of education for each occupation is taken from onet data.
Figure 8-9: Out of sample testing of model performance from Table 8.3. For each model, 1,000 trials are run where 75% of the data is randomly selected as training data and the remaining 25% of data is used as validation. The distribution root-mean-square errors for each model is reported. Medians are represented by a red line, while the mean error is represented by the green square.
### 8.1.2 Predicting Median Household Income of Cities

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.000</td>
<td>0.000</td>
<td>−0.000</td>
<td>0.000</td>
<td>−0.000</td>
</tr>
<tr>
<td>$SocioCognitive_c$</td>
<td>0.216***</td>
<td>−0.162</td>
<td>0.338***</td>
<td>0.407***</td>
<td>−0.147*</td>
</tr>
<tr>
<td>Occupation 1</td>
<td>0.354***</td>
<td>0.402***</td>
<td>0.407***</td>
<td>0.287***</td>
<td>0.407***</td>
</tr>
<tr>
<td>Occupation 2</td>
<td>0.229**</td>
<td>0.247**</td>
<td>0.233***</td>
<td>0.287***</td>
<td>0.287***</td>
</tr>
<tr>
<td>Occupation 3</td>
<td>−0.076</td>
<td>−0.040</td>
<td>0.338***</td>
<td>0.407***</td>
<td>0.287***</td>
</tr>
<tr>
<td>Occupation 4</td>
<td>−0.084</td>
<td>−0.066</td>
<td>0.338***</td>
<td>0.407***</td>
<td>0.287***</td>
</tr>
<tr>
<td>Occupation 5</td>
<td>−0.005</td>
<td>−0.013</td>
<td>0.338***</td>
<td>0.407***</td>
<td>0.287***</td>
</tr>
<tr>
<td>Occupation 6</td>
<td>−0.248***</td>
<td>−0.229**</td>
<td>−0.240***</td>
<td>−0.181**</td>
<td>−0.181**</td>
</tr>
<tr>
<td>Occupation 7</td>
<td>0.118</td>
<td>0.053</td>
<td>0.338***</td>
<td>0.407***</td>
<td>0.287***</td>
</tr>
<tr>
<td>Occupation 8</td>
<td>−0.089</td>
<td>−0.165</td>
<td>0.338***</td>
<td>0.407***</td>
<td>0.287***</td>
</tr>
</tbody>
</table>

$R^2$ | 0.046 | 0.456 | 0.458 | 0.424 | 0.435 |

$adj. R^2$ | 0.043 | 0.439 | 0.440 | 0.417 | 0.426 |

$p_{val} < 0.1^*, p_{val} < 0.01**$, $p_{val} < 0.001***$

Table 8.4: Linear regression using standardized $SocioCognitive_c$ for each city and employment in that city of each occupation type. All variables have been standardized.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.000</td>
<td>0.000</td>
<td>−0.000</td>
<td>0.000</td>
<td>−0.000</td>
</tr>
<tr>
<td>$SocioCognitive_c$</td>
<td>0.216***</td>
<td>−0.278***</td>
<td>−0.266***</td>
<td>0.436***</td>
<td>0.318***</td>
</tr>
<tr>
<td>No GED</td>
<td>−0.048</td>
<td>0.007</td>
<td>0.436***</td>
<td>0.318***</td>
<td>0.426***</td>
</tr>
<tr>
<td>H.S. Diploma</td>
<td>−0.340***</td>
<td>−0.402***</td>
<td>−0.351***</td>
<td>−0.423***</td>
<td>−0.423***</td>
</tr>
<tr>
<td>Associate’s Degree</td>
<td>−0.258***</td>
<td>−0.175**</td>
<td>−0.257***</td>
<td>−0.183**</td>
<td>−0.183**</td>
</tr>
<tr>
<td>Bachelor's Degree</td>
<td>0.317***</td>
<td>0.436***</td>
<td>0.318***</td>
<td>0.426***</td>
<td>0.426***</td>
</tr>
<tr>
<td>Master's Degree</td>
<td>−0.061</td>
<td>−0.001</td>
<td>0.317***</td>
<td>0.436***</td>
<td>0.318***</td>
</tr>
<tr>
<td>Doctoral Degree</td>
<td>0.047</td>
<td>0.056</td>
<td>0.317***</td>
<td>0.436***</td>
<td>0.318***</td>
</tr>
</tbody>
</table>

$R^2$ | 0.046 | 0.351 | 0.381 | 0.348 | 0.378 |

$adj. R^2$ | 0.043 | 0.337 | 0.364 | 0.341 | 0.369 |

$p_{val} < 0.1^*, p_{val} < 0.01**$, $p_{val} < 0.001***$

Table 8.5: Linear regression using $SocioCognitive_c$ and education variables. Education variables represent the employment in each city by highest educational degree attainment. All variables have been standardized.
Figure 8-10: Out of sample testing of model performance from Table 8.4. For each model, 1,000 trials are run where 75% of the data is randomly selected as training data and the remaining 25% of data is used as validation. The distribution root-mean-square errors for each model is reported. Medians are represented by a red line, while the mean error is represented by the green square.
Figure 8-11: Out of sample testing of model performance from Table 8.5. For each model, 1,000 trials are run where 75% of the data is randomly selected as training data and the remaining 25% of data is used as validation. The distribution root-mean-square errors for each model is reported. Medians are represented by a red line, while the mean error is represented by the green square.
8.2 SkillScape’s Predictive Power

Many publications in the literature [32][63] show how network features can predict the dynamics of the studied objects (individuals, occupations and cities in our case). They demonstrate this by reporting AUROC results, with network features as predictors as demonstrated in figure 8-12.

One important note however is that when picking the starting and ending thresholds for $\lambda$ (i.e. $\lambda_1$ and $\lambda_2$ in $Acquired_{t_1,t_2}^{\lambda_1,\lambda_2}$) not all occupations or cities would be included in the predictions. In other words, picking a large difference between the $\lambda$s would mean a very large jump which is a rarer occurrence, and thus would produce a smaller sample to study the AUROC results for such an instance. Figure 8-13 shows how this rarity is highlighted in the case of cities (since large jumps would require a huge shift in the labor force).
Figure 8-13: The top figures represent the fraction of instances (for cities or occupations) that have a change (i.e. $Acquired_{t1,t2}$) with the x and y axis representing the respective $\lambda_1$ and $\lambda_2$ values. While the dotted boxes in each of the top figures represent the zoomed area that the AUROC values will be studied in the bottom figures. The bottom figures represent comparison of AUROC distribution for all of the various cities of occupations. That is, for each $\lambda_1$ and $\lambda_2$ value, there are many instances of cities or occupations that have such a change, and we study the AUROC results for predicting such jumps using the different indicators (raw onet, RCA, or SkillScape's proximity metric).
AUROC for Predicting which Cities' Skills will jump from $LQ_{2005} < .9$ to $LQ_{2015} > 1$

Figure 8-14: The varying averages of AUROC achieved by combining the different variables with varying degrees (two network based, and one raw data based), creating this Dirichlet triangle. SkillScape is a network features, while $I$ & $LQ$ are none network features.

One interesting approach in the literature [32], was merging different metrics (network features) with raw data features, the result is a much better predictive model than any of the results for any individual feature. Even though, their intention was to show that network feature are better predictors than other non-network metrics (which still holds in our result), it would be important to mention this finding which we show in figure 8-14 that a combination of all 3 features gets an improved Area Under the ROC, but not by much, and still is slanted towards the network feature.

Not only does figure 8-14 show that combination of variables still lean more favorably towards the network based metric, but it also shows how the decline to low AUROC happens quite drastically when nearing the pure raw data based metrics.
8.3 How Educational Requirements Relate to Skill Requirements for Occupations

Figure 8-15: The skill requirements of an occupation indicate the education required. In each panel, we plot the SkillScape network thresholding edges with $\theta > 0.6$. Nodes (or skills) are colored according to the Pearson correlation between $onet(j, s)$ and the proportion of workers of each occupation with a given degree (title).
Figure 8-16: The correlation between social status and degree distribution in the Andorra dataset.

8.4 Human capital vs. social capital

In order to study some of the characteristics of the inferred feature vectors, we plot the heat maps of degree distribution in TMDB dataset (under $K = 8$). Since this is a bipartite undirected graph, we show the results of cast and directors respectively. As shown in Figure 8-17, human capital is strongly positively correlated with degree ($\rho = 0.57, p < 0.00$) for directors, while it is weakly correlated with degree ($\rho = 0.05, p < 0.00$) for cast members, as shown in Figure 8-17.

Figure 8-17: The correlation between human capital and social capital in TMDB dataset.
Bibliography


[67] Rakesh Kochhar, Richard Fry, and Molly Rohal. The American middle class is losing ground, 2015.


