Essays on Information Technologies, Social Networks and Individual Economic Outcomes

by

Guillaume B. Saint-Jacques
Master's in Economics, Paris School of Economics & École Normale Supérieure, 2009
Mater's in Management Science Research, MIT Sloan, 2015

SUBMITTED TO THE SLOAN SCHOOL OF MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
February 2018

Signature redacted

Author .....
Guillaume Saint-Jacques
MIT Sloan School of Management
Jan 8th, 2018

Signature redacted

Certified by .....
Erik Brynjolfsson
Schussel Family Professor of Management Science
Thesis Supervisor

Signature redacted

Accepted by .....
Catherine Tucker
Professor of Marketing
Chair, Ph.D. Program, MIT Sloan School of Management

Essays on Information Technologies, Social Networks and Economic Outcomes

by

Guillaume B. Saint-Jacques

Submitted To The Sloan School Of Management on January 12th, 2018

in Partial Fulfillment Of The Requirements For The Degree Of
Doctor Of Philosophy In Management Science

ABSTRACT

This four-part thesis focuses on the effect of Information Technologies and individual economic outcomes. The first two papers investigate the relationship between technology and individual pay. The last two focus on connectivity through social and communication networks and labor market outcomes.

The first paper offers a simple model of how technology may be reshaping the distribution of individual income in the US between 1960 and 2008. First, is shows fractal patterns of the income distribution, which indicate the presence of an increasingly unequal power law distribution at the top. Then, using IRS individual tax data, it shows two main trends: first, more and more individuals seem to draw their income from a Pareto distribution rather than a Lognormal distribution (typical of the "industrial economy"). Second, the tail index of the Pareto distribution seems to be getting lower, indicating increasing inequality at the top.

The second paper investigates the relationship between Information Technologies and CEO Pay. It shows, both at the industry and at the firm level, that IT intensity seems to increase CEO Pay. It shows support for three distinct mechanisms: first, IT makes firms bigger. Second, it increases their effective size (the size effectively affected by CEO decisions). Third, it increases mobility of CEOs, possibly because managing IT-intensive companies requires relatively more general skills.

The third paper is the first to match, at the individual level, complete Call Detail Records data with individual income for over 100,000 individuals. This allows to describe the associations between individual income and patterns of individual's social networks. We find that wealthier individuals have higher degree, much higher network diversity, higher local centrality, and more social engagement, but lower density and reciprocity in their individual networks.

The last paper attempts a causal study of the relationship between social tie strength and individual labor market outcomes (measured as job mobility) using LinkedIn's People-You-May-Know randomizations. It shows an inverted U-shape relationship between tie strength and job transmissions, as well as a globally negative relationship between clustering coefficient and labor market mobility, suggesting that even individually, strong ties are not always more useful than weak ties.

Thesis Supervisor: Erik Brynjolfsson

Title: Schussel Family Professor of Management Science, MIT Sloan

Acknowledgements

MIT was an incredible place to do research, and I owe a great debt of gratitude to many members of the MIT community.

First, none of this research would have been possible without Erik Brynjolfsson's help, advice, and encouragement. His immense intelligence and taste in research questions, combined with such modesty and humor were an inspiration every day. His office was the place of many stimulating conversations, and I can't thank him enough for his support throughout these years. Working with him has been a pleasure and a privilege, and has profoundly shaped the way I think about the world.

I would like to thank Sinan Aral for his advice, and for to teaching me the art of formulating a research question. To Edoardo Airoldi, thank you for your advice, for your feedback, and for your friendship.

I am also indebted to Thomas Piketty, who got me interested in research in the first place. During my master's program, Thomas was an exceptionally attentive advisor. Working with him on the *Revolution Fiscale* project gave me a first-hand look at how research can have an impact on the world.

I am also grateful my coauthors: Nick Fazzari, Eaman Jahany, Andrew Lo, Sandy Pentland, Jean Pouget-Abadie, Neil Thompson and Ya Xu.

I would also like my friends, whose friendship I enjoyed over these years. Special thanks go to Daniel Rock, who was my brother in arms throughout the PhD process, as well as an exceptional friend. I would also like to thank Maxine Van Doren, Seth Benzell Fernanda Bravo, Audren Cloitre, Maxime Cohen, HyungJune Kang, Arvind Karunakaran, Laura Rock and Leon Valdes, whose friendship and support will not soon be forgotten!

Finally, I can't express how grateful I am for the support of my parents and my sister, as well as my wife's. To my mother: your unwavering strength and determination and the high standards you always set for yourself have always been an inspiration. To my wife Elizabeth: every day you teach me to enjoy life a little more, every day is a new adventure. Finally, to Alfred Bittendiebel, my late grandfather, who is my model in everything I do.

Thank you all!

Guillaume

Table of Contents

Information Technology and the Rise of the Power Law Economy

Introduction

In 1970, a person in the top 1% of the wage distribution earned between seven and eight times as much as someone in the bottom 90%. By 2008, this ratio had roughly tripled to about 22. While US has experienced a tremendous increase in income inequality over the past forty years, research on this phenomenon is at a critical crossroads. On one hand, the facts documenting the increase in US inequality are clear. On the other hand, there is not a consensus on the driving forces that caused it. Establishing the role of technology, if any, and specifically the ways that information technologies can affect the income distribution, is a key task for IS researchers.

Most of the literature on the effects of information technologies and inequality has focused on skill-biased technical change, documenting and explaining a broad increase in the relative wages of skilled workers, such as college graduates, over unskilled workers. A less well studied phenomenon, which we focus on in this paper, is the potential of technology to create superstar effects. Digital technologies can amplify the ideas, talents or luck of a small handful of innovators, vastly increasing their income as they reach a broader market. In contrast to earlier production and distribution technologies, goods and services that are digitized can be replicated at nearly zero cost, with perfect fidelity and reach global audiences almost instantaneously. This fundamentally changes the way that value is distributed and allocated, and it would be surprising if it didn't have significant effects on the income distribution.

In this paper, we argue that the shape of the income distribution is changing in specific way: using nearly 50 years of tax data from the United States Internal Revenue Service (IRS) and a new modelling approach, we show that a bigger share of individual incomes are drawn from a power law, or Pareto distribution, as opposed to the long-established log-normal distribution that historically governed incomes. We argue that the increased prevalence of Power Law distributions is consistent with the effects of the diffusion of Information Technology, because digitization and networks facilitate winner-take-most markets. We present a simple theoretical model of income distribution and of the role of income technology, and estimate it using the IRS public use tax files. We find that more and more individuals seem to participate in winner-take-most markets, and that, within these markets, competitiveness has been steadily increasing.

The remainder of this paper is organized as follows: Section 2 highlights some of the relevant literature. Section 3 introduces simple models of the key economics of information technology and its potential to produce highly skewed income distributions, and specifically power laws. Our models imply four hypotheses on the effects of IT the distribution of income and ways to empirically test them. Section 4 presents our basic tax data and shows evidence of fractal effects in income, which are consistent with the presence of power laws. Section 4 presents our empirical strategy and assesses our hypotheses more formally. Section 5 concludes with a summary and some implications.

Relevant Literature

Recent Trends in Inequality

This paper draws on three main spans of literature. First, it builds on recent research on income inequality and the role of technology. Second, it builds on the more specific span of the economics literature dedicated to understanding functional forms of the income distribution. Finally, it draws on recent research in statistics and actuarial science for tools to estimate specific mixture models of statistical distributions.

The recent increase in income inequality has been widely documented. Some of the most recent research started in France (Piketty 2001) and in the US (Piketty and Saez 2003), and was then extended to a large number of countries (Atkinson and Piketty 2010). The starting point of this literature is the Kuznets curve: the idea that inequality follows an inverted U-shape over time, in the sense that the industrial revolution brought about an increase in inequality which gradually reversed over time (Kuznets 1955). After forty more years of data, Piketty argues that the Kuznets curve has now turned up again: after decreasing around WW2, inequality has been increasing again in the United States since the 1980s. This phenomenon has accelerated in recent years: while the 1990s expansion led to a 10% increase of top 1% incomes and a 2.4% of bottom incomes, the more recent 2001-2007 expansion led to an 11% increase at the top and an increase under 1% at the bottom (Saez 2013). In other words, 75% of recent income growth went to the top 1%.

Why should anyone care about the income distribution? There are several types of arguments present in the literature: First, even if increased inequality is linked with a higher mean income, large groups of the population can be left significantly worse off. This has

been the case for the median family between 1990 and 2008 (Bernstein 2010). Second, increased inequality is associated with decreased mobility (Corak 2013). This finding was coined by Krueger (2013) as the "Great Gatsby curve", and may create concern that inequality can be self-perpetuating. Recent evidence shows that household mobility as become very low (Debacker et al. 2013). Third, increased economic inequality may lead to increased political inequality, leading to a vicious circle (Acemoglu and Robinson 2012). Finally, a great deal of happiness seems to be dependent of relative levels of income and wealth rather than absolute levels, and a dislike of inequality may directly enter into the utility function for many people.

Theories of Inequality: Institutions vs Markets

There is no general consensus regarding the driving forces behind inequality. A first class of arguments emphasizes the impact of institutions and historical events. Among the proponents of institutional explanations, one can find arguments relating to recent drops in tax rates (Piketty and Saez 2003) favoring the rich, or to various types of rent-seeking (Bivens and Mishel 2013). It is also argued that past reductions in inequality were mostly accidental rather than the result of market forces. In this literature, emphasis is put on the effects of the two World Wars and on the great depression as the main driver of the large drop in inequality that was observed in the first half of the 20th century: a lot of wealth was destroyed, and wealthy individuals were disproportionately affected.

A second class of arguments emphasizes markets and technology as drivers of inequality. These "market" arguments can roughly be decomposed into two main categories: *skill-biased technical change* and *superstars*. Skill biased technical change is the idea

that new technologies emerge that have much higher complementarities with skilled labor than unskilled labor, therefore increasing wages of skilled individuals and depressing the wages of others. SBTC is an important part of the labor economics literature focusing on the college premium. For a comprehensive review of this literature, see Katz and Autor (1999), Katz (2000), and Acemoglu and Autor (2012). It is worth noting that this literature typically focuses on the bottom 99%, i.e. tends to exclude top incomes from the analysis. Generally speaking, the bulk of the income distribution seems to follow a log-normal distribution. This seems to be the case over time, and in a large number of countries (Lopez and Serven 2006).

On the other hand, the literature on superstars is concerned with a much smaller number of individuals: rather than investigating the difference between skilled and unskilled labor (two very large groups), it focuses on a very small of fraction of the population which receives very high incomes. While small in numbers, this group can command a large (and growing) share of aggregate income. The literature on superstars emphasizes the effects of technology through economies of scale that allows the most talented individuals to replicate their talent across larger and larger markets (Rosen 1981). The very top of the income distribution seems to follow a Pareto distribution rather than a lognormal one. Pareto distributions seem to be present in a very large array of phenomena: City Size, frequency of words in languages, casualties in wars, cotton prices, movie profits, publications and citations by researchers, social networks, and many more. As analyzed in the literature on the "Long Tail", sales of books and other products online can be very well described by a simple Pareto distribution (Brynjolfsson, Hu and Smith,

2003). For a list of documented phenomena following a Pareto distribution, see Andriani and McKelvey (2009).[1]

Theory: Distributional Effects of Digitization

The dominant explanation of income inequality has historically been skill-biased technical change: new technologies are used that require higher-skilled labor, thereby increasing wages of high-skill workers and depressing wages of lower-skilled workers. Autor et al. (2008) show an increase in both inter-group (according to age, education and gender) and intra-group (residual) inequality. They document polarization, i.e. middle income workers being most negatively affected relative to those at the high and low end. They offer the explanation that computerization substitutes for routine information processing work more than non-routine cognitive work (at the high end) or non-routine manual work (at the low end). They also emphasize the importance of residual inequality: even within an industry and a job category, inequality is increasing. In an earlier paper, Autor et al. (1998) show the influence of computerization on inequality through skill upgrading: the college premium increases faster in computer-intensive industries. Bresnahan et al. (2002) also show that greater use of information technology within a

[1] It is worth mentioning that more complex statistical distributions have also been used to describe incomes: Lognormal and Pareto distribution have recently been put together in the double Pareto Lognormal distribution, which is shown to fit income data very well (Hajargasht and Griffiths 2013). However, it is interpreted in terms of proportional random shocks applied to an exponentially growing population rather than in terms of individual productivity or markets (Reed and Jorgensen 2004).

firm leads to the employment of higher educated workers and greater investments in training.

However, these explanations focus on the "other 99%" (Autor 2014), and do not address the top 1%, which is a main cause of the increase in aggregate inequality. The chart below shows the evolution of the ratio of reported income of the top 1% of taxpayers to the bottom 90% of taxpayers. The increase since the 1980s is rather dramatic: where the average wage in the top 1% was roughly 10 times higher than the average wage in the bottom 90%, it is now over 20 times higher. This is evidence that understanding inequality likely requires paying special attention to the top 1% of wage earners.

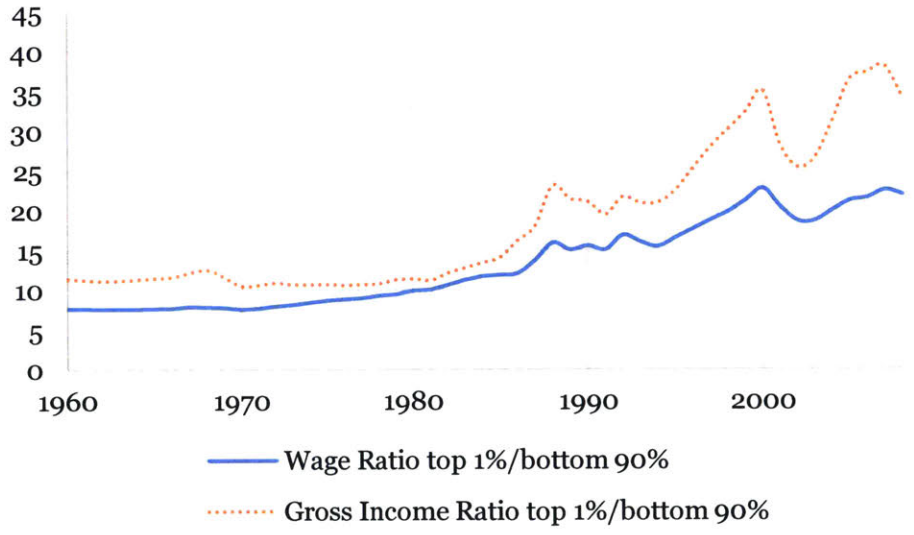


Figure 1: Ratio of average income reported by the top 1% to the bottom 90%

Recent research has argued that some people are earning a disproportionate share of income because they have become much more productive. This seems to be the case for managers (Kaplan and Rauh 2013) and CEOs (Gabaix and Landier 2008). The link between increases in managerial productivity and information technology can be established using the concept of effective size (Kim and Brynjolfsson 2009). As companies become more IT-intensive, managers can control larger and larger spans of activity. A good manager can create a lot of value, and a bad manager can destroy a lot of it.

Investigating the increase of top 1% incomes likely requires focusing on the properties of the digital revolution. The digital economy seems to have two main characteristics: first it is *digital,* allowing for production that is nearly *free, perfect,* and *instantaneous.* Digital goods can be replicated at almost zero marginal costs: For example, streaming music to two customers costs virtually the same as streaming to one customer only. Digital goods can be replicated perfectly, as the act of copying data does not degrade it. Further, goods can be delivered nearly instantaneously to anyone connected to the global internet. This alone can lead to large economies of scale and generate winner-take-most economies.

Second, the digital economy is fundamentally *networked.* For example, when choosing a new instant messaging app to use, or a new smartphone game to download, users may be influenced by the number of other people already using these services. The more people use the platform, the more likely they will learn of it and, in many cases, the more valuable it becomes to any individual user. As a consequence, individuals tend to preferentially attach to services that already have a large user base. Or, they may primarily choose their new apps from the top-ranked apps from their favorite app store. Fleder

and Hosanagar (2009) have shown that such recommender systems can create biases towards already popular products, creating a rich-get-richer effect.

How the Industrial Economy Generates Log-normal Income Distributions

The main theory of the income-generating process that leads to a lognormal distribution posits that individuals receive a series of random shocks to their income (Gibrat 1931; Mincer 1958). Individual characteristics seem to often be roughly normally distributed. This is the case for height, IQ, gripping strength, and many more. This motivates the question: "How can one reconcile the normal distribution of abilities with a sharply skewed distribution of incomes"? (Pigou 1932)

Lognormal distributions are typically generated as the product of a large number of random variables. For example, assume that the productivity of a car mechanic is the product of a number of her individual abilities, such as attention to detail, information gathering ability, skill in operating vehicles, decision making skill, ability to establish and maintain relationships, strength, fine motor skills, and so on. The higher the number of relevant characteristics, the more then distribution of productivity in the population will resemble a lognormal. This follows straightforwardly from the central limit theorem in log space:

If $X_1, \dots X_n$ are i.i.d. individual abilities with mean 0 and variance $\sigma^2 < \infty$, and $Y = \Pi_{i=1..n} X_i$, then

$\log(Y) = \sum_{i=1..N} \log(X_i)$, and by the C.L.T, $\log(Y) \xrightarrow{d} (0, n\sigma^2)$.

Two main implications follow: first, if a car mechanic becomes 25% stronger, his productivity will increase by 25%. But if he becomes both 25% stronger and has a 25% increase

in fine motor skills, one can expect his productivity to increase by over 56%. In other words, there are complementarities between individual abilities, so that the overall productivity effect of having high abilities is higher than the sum of marginal productivity effects of these abilities taken individually. If compensation is proportional to marginal productivity, then this process will create a log-normal distribution of income.

Second, as the relevant number of individual characteristics increases, the observed log variance of productivity should increase as well: if new abilities, such as various abilities linked to interacting with computers, become required, the observed $n\sigma^2$ should increase. Conversely, if this process slows, $n\sigma^2$ should stagnate. This increase in $n\sigma^2$ can be thought of as roughly capturing "industrial" technical change.

How the Digital Economy Generates Power Laws

Previous theories of Pareto distributions in labor income often rely on matching individuals with economic fundamentals that are already known to be Pareto distributed. For example, Gabaix and Landier (2008) match managers with firms, the size of which empirically follows a Pareto distribution. This can generate Pareto distributions for managerial compensation as well. A key takeaway from this literature is that infinitesimal differences in individual ability may result in very large differences in pay. This often results from matching processes between individuals and jobs that have higher and higher stakes, creating winner-take-most job markets.

We argue that digitization and networks can generate a Pareto distribution of income more directly. Let us go back to the example of competing messaging apps. Let us assume there are K apps, which all start out with N_0 users each. At each time t, L new users

arrive and face the choice of which app to use. If there are network effects (i.e. if an app is more valuable when it has more users), then each new user will tend to choose apps that already boast a large user base. For simplicity, let us assume that a new user's probability of choosing app *i* is simply app *i*'s share of total users.

$$P(\text{"new user chooses } i\text{"}) = \frac{N_i}{\sum_{j=1,\ldots K} N_j}$$

At *t=0,* all apps have equal probability of gaining a new users. At *t=1,* the app that ended up gaining the arriving users now has higher probability of gaining future users than other apps. This self-reinforcing mechanism, known as the preferential attachment model, leads to a power law, as was shown in the context of social networks (Barabási and Albert 1999). A simple simulation can illustrate this point.

We start with K=5 apps, having one user each. At each time, three new users arrive and preferentially choose one of the apps based on its current market share:

	Market Shares					Log(Users) / Rank plot at t=1M for 100 apps
t	App	App	App	App	App	
0	20.0%	20.0%	20.0%	20.0%	20.0%	
10	11.4%	11.4%	2.9%	71.4%	2.9%	
100	7.2%	4.3%	0.3%	87.9%	0.3%	
1,000	6.8%	2.4%	0.1%	90.6%	0.0%	
10,000	6.4%	2.8%	0.2%	90.5%	0.1%	
100,00	6.3%	2.8%	0.2%	90.7%	0.1%	

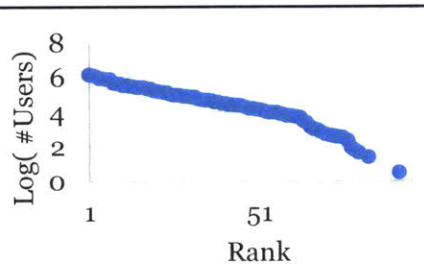


Figure 2A-2B: A simulation of power laws through preferential attachment

A common way of recognizing a power law is to regress the logarithm of the level of the variable on the rank of the observation. Figure 2B illustrates that power laws look like straight lines when represented this way. As can be seen from figure 2A, such mechanisms can converge to a very unequal distribution of users (and therefore of income), which then persists over time. This is a fundamentally different dynamic than what is observed in the industrial economy, and can have a large effect on income distribution[2].

A commonly used power law is the Pareto distribution, with survival function:

$$\Pr(X > x) = (\frac{c}{x})^{\alpha} \; \forall x \in (c, +\infty)$$

When α decreases, the distribution gets a fatter and fatter tail. When α drops below 2, the variance is infinite. When α drops below one, the mean becomes infinite as well.

A Simple Mixture Model of Income Distribution

Considering that software investments have increased over 50% between 2000 and 2012 we now turn to a model of an increasingly digitized economy.

---

[2] The large concentration in information goods market has been documented in (Jones and Mendelson 2011)

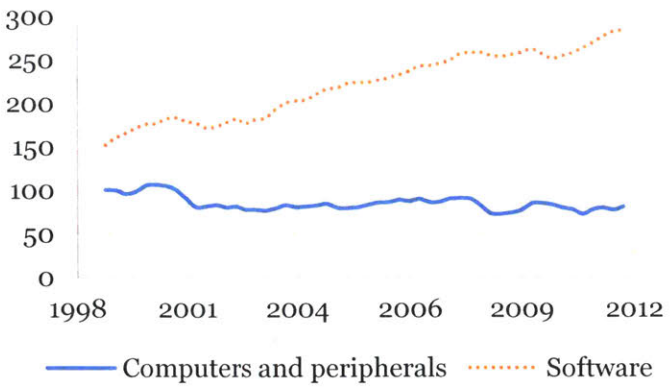


Figure 3: Software and Computer investment in billions of dollars, 1999-2012.

If, as Marc Andreessen put it, "software is eating the world" (Andreessen 2011), then our model predicts that a power law will govern an increasing share of the income distribution. The distributional consequences can be large. For instance, more and more people will be below average income, and a smaller number of individuals will have a larger share of the overall income (Brynjolfsson et al. 2014).

Let us consider what happens when individuals optimally chose which sector of the economy they want to partake in, based on their expected income in each sector. An example could be an accountant who programs mobile apps on evenings, or a Hollywood waiter who is considering becoming an actor.

We posit that each person has two potential incomes drawn from two distributions:

$$X_L \sim LogNormal(\mu, \sigma)$$

$$X_P \sim Pareto(c, \alpha)$$

Where X_L is the income the individual would have if she took a job in the traditional economy (i.e. if she became an accountant or waiter), and X_P is the income she would have if she tried her luck in the digital economy (as an actor or app programmer). For simplicity, assume that each individual perfectly observes income offers from each option, and then chooses

$$Y = max(Y_P, Y_L).$$

Throughout the rest of this paper, we will refer to this model as the "Max model". Note that there are other ways to mix a Pareto distribution and a Log-Normal one, such as using a spliced model, where a lognormal distribution is fitted at the bottom and a Pareto at the top under a set of restrictions that guarantee that the resulting distribution is continuous and differentiable. Such approaches introduce scaling factors which complicate interpretation and do not map into a simple model of income generating processes, which is why we do not discuss them in the body of the paper. However, as shown in the appendix, they generate findings that are broadly consistent with the ones we derive using the Max model.

Our theory implies four hypotheses. Using the Max model, we can formally state each of them, and we can use our data to test them:

> *Hypothesis 1: If superstar effects are an important part of the economy, a model of income distributions incorporating both log-normal and Pareto distributions should fit the data significantly better than a traditional log-normal alone.*

If the lognormal variance does capture some dimension of SBTC, then we could expect it to increase over time, but much more sharply in early years than in recent years:

> *Hypothesis 2A: Industrial technical change should lead the variance of the underlying lognormal distribution to significantly increase over time.*
> *Hypothesis 2B: If industrial technical change is slowing, the observed variance of the log-normal part of the income distribution should be stabilizing over time.*

Similarly, if "software is eating the world", we should see more and more people joining the digital economy. If the "networked" aspect of the digital economy is accentuating, we should expect skewness to increase there.

> *Hypothesis 3: If more and more people are joining the digital economy, the number of individuals whose income seems to be drawn from a Power law should be increasing over time.*

> *Hypothesis 4: If the digital economy is becoming more digital (closer to free, perfect, and instant) and more networked, incomes within that economy should become increasingly skewed, resulting in a lower Pareto parameter, α.*

Data and results

Relevant Data Sources

Studying income inequality can usually be achieved using one of two different data sources:

- Survey data, which may contain a wealth of demographic variables, information on sectors of activity and occupation. The main limitation of survey data lies in their poor sampling properties at the top: top incomes represent a very small fraction of the population, and are likely to be missed by survey sampling. However, these individuals account for a very significant fraction of total income. This is why survey data is generally not used to compute income shares. Furthermore, the study of the global shape of the income distribution requires fine-grained data, and would therefore not be achievable on survey data.
- Tax data. Because filing taxes is mandatory above a certain income threshold, federal income tax files are very helpful in studying income shares and the income distribution. However, they are surrounded by a number of strict confidentiality policies, so that demographic information is poor.

For the purposes of this paper, we use the second kind of source, tax data. We do this in particular because we are interested in studying power laws, and because so much of the action in power laws occurs at the very top of the distribution where there are very few observations, tax data provides more suitable information.

We use the IRS Tax Model Files, which are samples of US Federal Individual Income Tax returns between 1960 and 2008 (except for 1961, 1963 and 1965). The data has been

anonymized and blurred out of concern for taxpayer's privacy. In particular, the data has very little demographic information, such as job or industry. For example, even state of residence is not available to researchers for all taxpayers whose income exceeds $200,000. However, its very good sampling properties at the top make it the best dataset to estimate income distributions.

In order to establish the validity of our approach, we first share a few summary statistics about the sampling design and argue that it can help us estimate a parametric model of income distribution much better than survey data could. Then, using the "wages" variable contained in the data, we provide some summary statistics about inequality, including top income shares. The following table illustrates the quality of sampling for year 1960:

q0-q50		**q50-q80**		**q80-q90**		**q90-q95**	
n	w/n	n	w/n	n	w/n	n	w/n
45,000	676	17,000	1111	6,000	1035	6,000	514
q95-q99		**q99-q99.9**		**q99.9+**			
n	w/n	n	w/n	n	w/n		
12,000	210	7,000	81	10,000	6		

*Note: **n** represents the number of observations in each quantile, and **w/n** the average weight per observation in each quantile. A lower value for **w/n** at the top of the income distribution means better sampling at the top.*

Table 1: Sampling properties of IRS public use tax data

As can be seen from table 1, the sampling design of our data seems to lend itself well to a study of top incomes: relative to the rest of the population, the top incomes are significantly oversampled. Where, on average, a thousand individuals whose labor income falls between the 50th and 80th percentile are grouped into one observation, only 6 individuals from the top 0.1% are grouped into one observation. This level of precision allows us to work on the shape of the income distribution as a whole, rather than conditional averages.

A first look at income densities

Let us now have a first look at the shape of the distribution of income. Using Gaussian kernel smoothing as in (DiNardo et al. 1996) allows us to graph densities of log incomes for year 1960 and 2008 as follows.

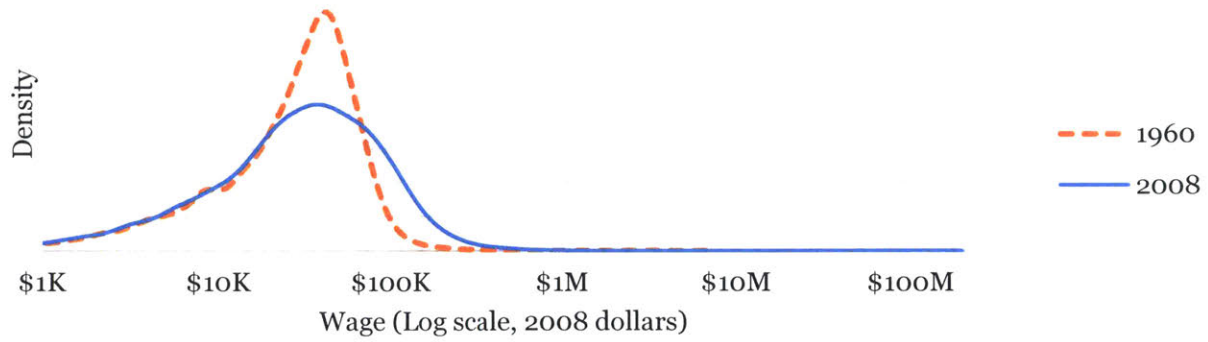


Figure 4: Kernel-Smoothed densities of wage distributions in 1960 and 2008.

One can notice that the bulk of both the 1960 and 2008 distributions look roughly log-normal. However, they show a very long right tail. Furthermore, the tail appears to have very significantly extended between 1960 and 2008. Yet this representation may still

understate the importance of changes in the tails, because they are hard to see. Accordingly, we offer an illustration of fractal effects in income shares.

Fractal effects in income shares

Perhaps one of the most compelling ways of illustrating trends in income inequality is to compute "income shares." Income shares tables illustrate the share of all of a specific kind of income that is earned by a given percentile of earners. Here and throughout the rest of the paper, we restrict our analysis to wages (as reported to the IRS). This allows us to exclude capital income, which may be subject to different dynamics and interfere with our focus on individual productivity.

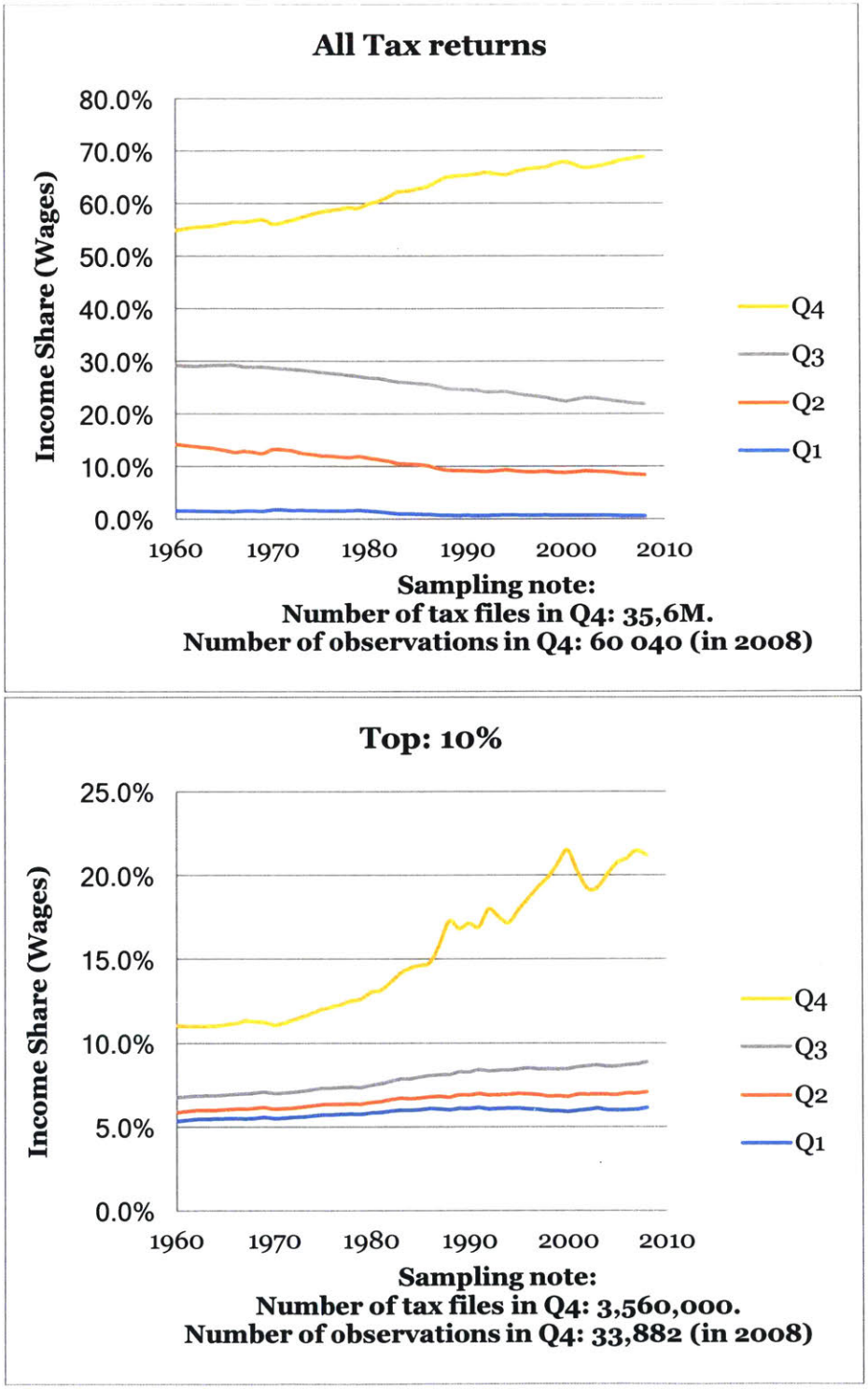
In order to give a sense of the widening of income inequality and in order to underline the influence of power laws, we offer the following visualization:

- First, we sum all the reported wages across all individuals and compute the share of these wages earned by the top 100%, 10%, 1%, 0.1%, 0.01% and 0.001% of the distribution.
- We then break these into quarters and show the evolution of their income shares since 1960.

For example, figure 5B is a decomposition of income shares of the top 10%: Q1 represents the bottom quarter of that group (i.e. individuals between *p90* and *p92.5*), and Q4 represents the top quarter (*p97.5* and above). The other figures represent the same composition, but of different portions of the population (respectively: all of it, and the top 1%, 0.1%, and so on).

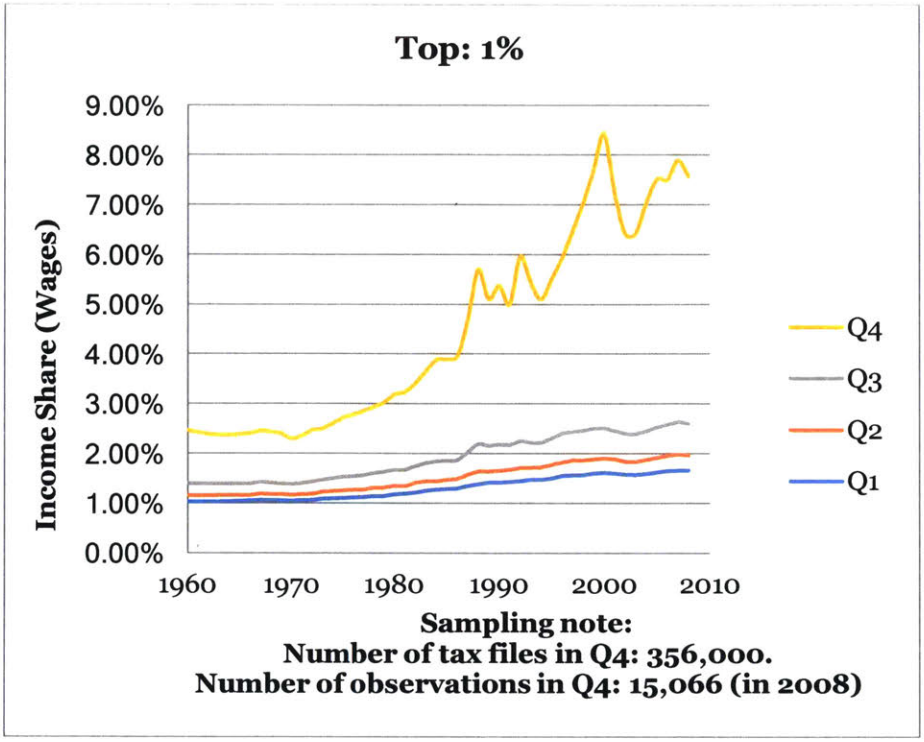
This visualization has two main advantages. First, it shows that income increases are concentrated at the very top of the income distribution. More specifically, it seems that a very significant share of income inequality can be traced back to the top 0.1%. Figure 5C shows that the income share of individuals located in the top quarter of the top 1% (i.e. *p99.75* and above) went from about 2.5% in 1960 to almost 8% in 2008. About half of that 5.5 point increase can be traced back to the top quarter of the top 0.1% (i.e. *p99.975* and above): their income share went from .5% to over 3%.

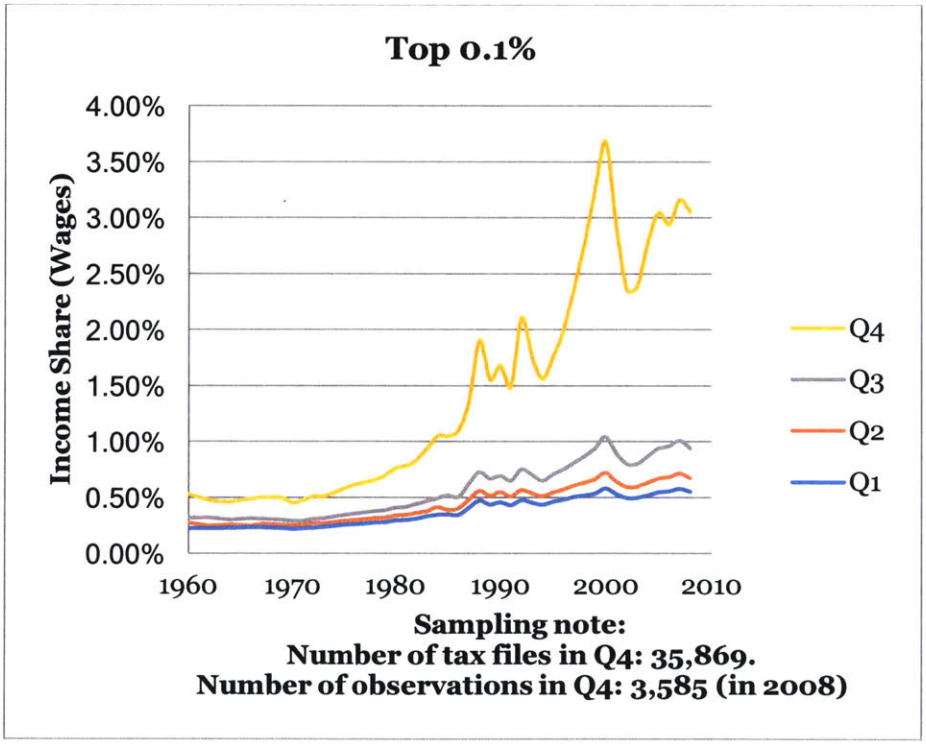
Second, this visualization shows a "fractal effect". No matter how far we zoom in, the global picture is apparently the same. As you go up the income distribution, we see that the 1% have their own 1% and so on, with a similar (if somewhat noisier) pattern of income distribution. This remarkable scale independence is a feature of power laws (also called scale-free distributions).



Note: In 2008, the top 10% starts at an annual wage of $96,000$. The top 1% starts at $260,000. The top 0.1% starts at $860,000. The top 0.01% starts at $3.7M. Finally the top 0.001% starts at $15.4M.

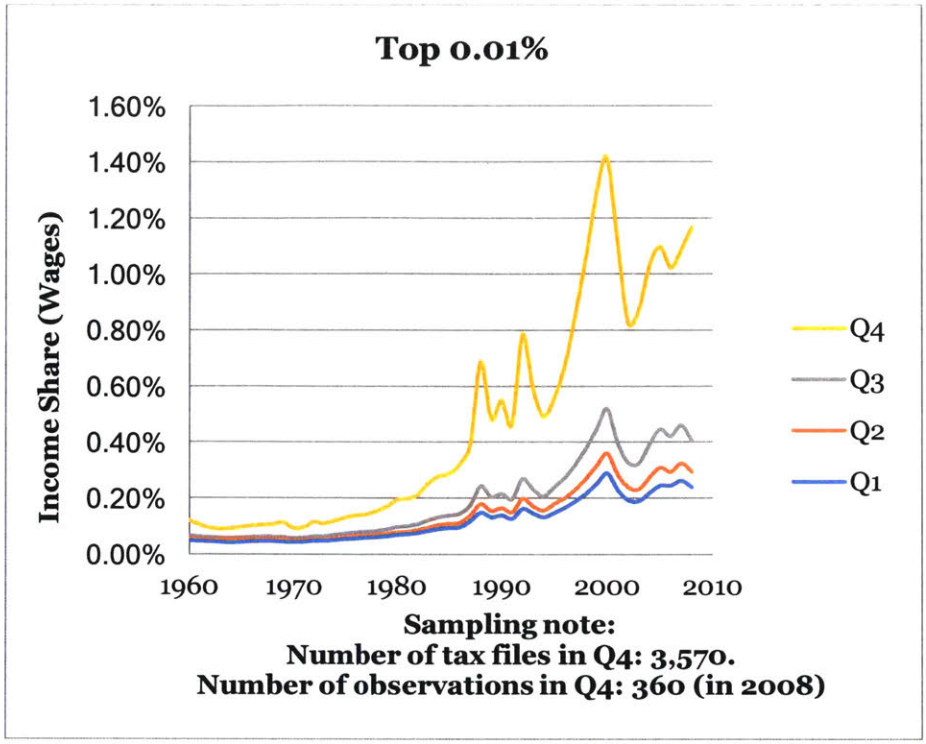
Figures 5A-5B: Income shares fractal patterns, 1960-2008

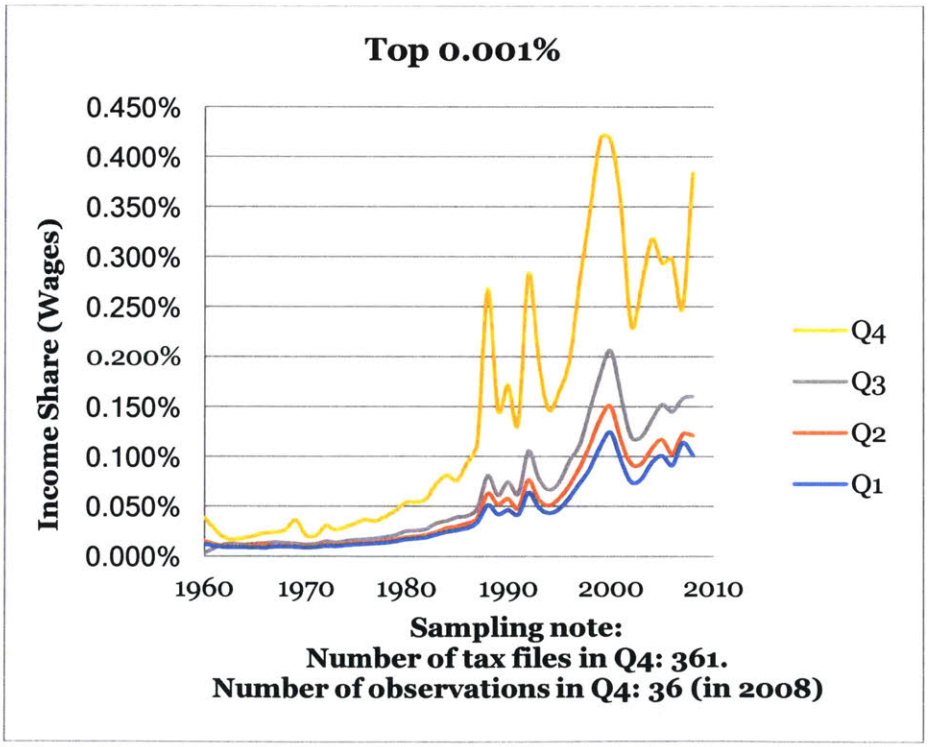




Note: In 2008, the top 10% starts at an annual wage of $96,000$. The top 1% starts at $260,000. The top 0.1% starts at $860,000. The top 0.01% starts at $3.7M. Finally the top 0.001% starts at $15.4M.

Figures 6C-5D: Income shares fractal patterns, 1960-2008





Note: In 2008, the top 10% starts at an annual wage of $96,000$. The top 1% starts at $260,000. The top 0.1% starts at $860,000. The top 0.01% starts at $3.7M. Finally the top 0.001% starts at $15.4M.

Figures 7E-5F: Income shares fractal patterns, 1960-2008

Estimating the Max model

Having documented the presence and significance of power laws in the U.S. income distribution, let us go back to our model where

$$Y = max(Y_P, Y_L).$$

The underlying potential incomes, Y_P and Y_L are fundamendally unobservable to us as econometricians. We only observe the mixture Y, and recover the original parameters (μ, σ, c, α) through numerical Maximum Likelihood Estimation and Goodness of fit estimation (using variations of the Anderson-Darling statistic). Once the parameters are recovered, we can simply obtain $Q = P(Y_P > X_L)$, the proportion of people selecting their Pareto draw over their lognormal draw.

Note that the tax data contains many individuals (roughly around 15%) who report an income of zero, as well as many individuals with a very low income. We therefore fit a censored lognormal distribution. For simplicity, we give the Pareto and the Lognormal distribution the same support, by fixing c at the 25th percentile and then truncating the lognormal distribution at the same value of c. While this model embodies a number of simplifications, our results do not seem qualitatively sensitive to different specifications (see appendix for the "spliced" model).

It is straightforward to show that the resulting probability density function of this model is:

$$\frac{e^{-\frac{(\mu-\log(x))^2}{2\sigma^2}}(1-(\frac{x}{c})^{-\alpha})}{\sqrt{2\pi}x\sqrt{\sigma^2}} + \frac{\left(\frac{x}{c}\right)^{-1-\alpha}\alpha(1-\frac{1}{2}\Psi(-\frac{\mu-\log(x)}{\sqrt{2}\sqrt{\sigma^2}}))}{c}$$

Similarly, the resulting cumulative distribution function is:

$$(1-(\frac{x}{c})^{-\alpha})(1-\frac{\Psi(-\frac{\mu-\mathrm{Log}(x)}{\sqrt{2}\sqrt{\sigma^2}})}{\Psi(-\frac{\mu-\mathrm{Log}(c)}{\sqrt{2}\sqrt{\sigma^2}})})$$

Note that by design we have $0 < c < x$. $\Psi(.)$ is the canonical complementary error function, i.e.

$$\Psi(x) = \frac{2}{\sqrt{\pi}}\int_{x}^{\infty} e^{-t^2}dt$$

Maximum likelihood Estimates (MLE) are obtained by numerical optimization (using the PDF above), as there is no closed form for them because of the need bottom-truncate the distributions. We fit the model using MLE as well as different types of goodness of fit statistics (using the CDF above), including the Anderson-Darling (AD) statistic as well as ADR and AD2R, which are variants of AD that emphasize fit in the right tail (Luceño 2006). These methods yield results that are broadly similar to the ones obtained with MLE, but somewhat more dramatic: they identify a lower Pareto α and a higher number of individuals in the digital economy. However, the trends they identify over time are the same as the ones obtained through MLE. For the sake of clarity, we focus on MLE results for the remainder of this paper. The below figures show MLE estimates of all underlying parameters of the Max model:

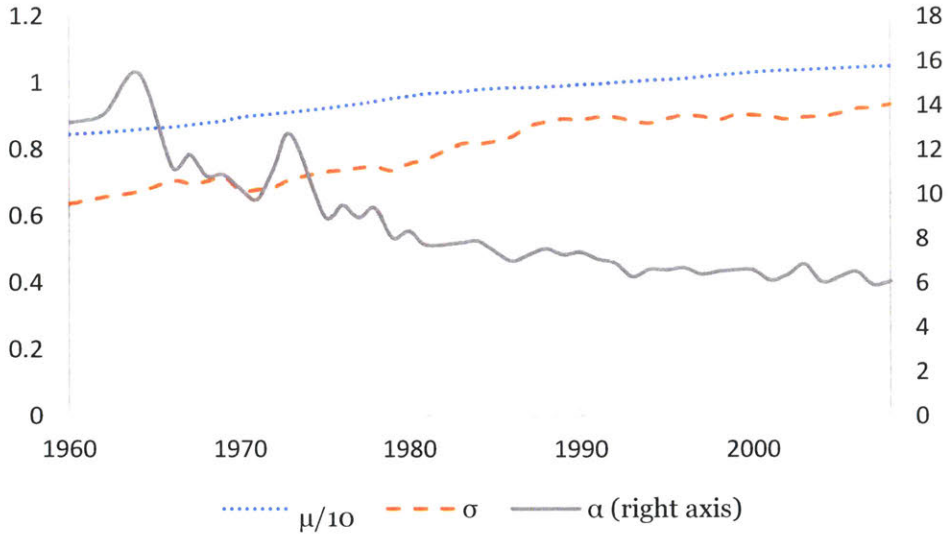


Figure 8: MLE estimates of the Lognormal/Pareto Mixture based on our model

The most interesting measure here is the estimated number of individuals having chosen to enter a winner-take-most sector. The following table describes a confidence interval for that measure, based on a 500 bootstrap sample:

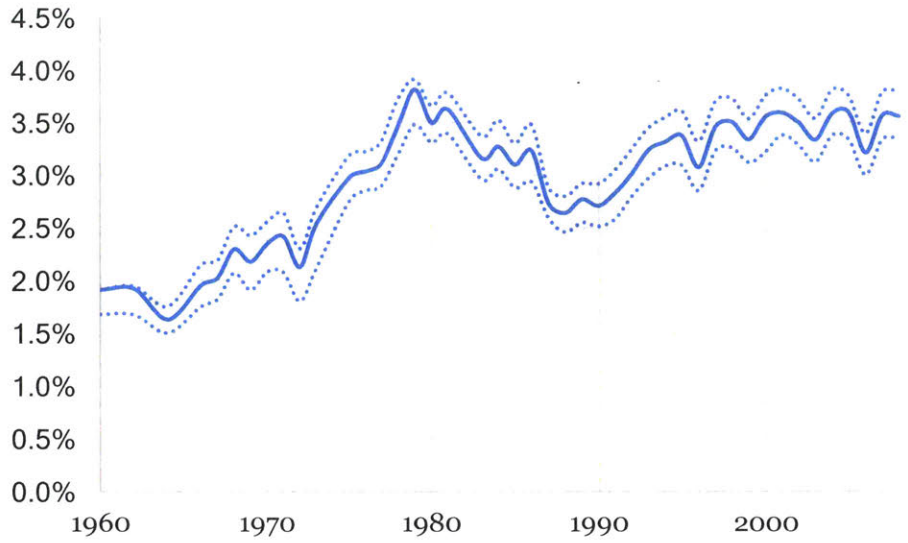


Figure 9: Percentage of individuals whose income is drawn from the Pareto distribution (Q), with 95% confidence interval.

The estimates of this model seem to recover the qualitative facts discussed earlier: more and more people are part of the Pareto distribution, and the Pareto distribution is becoming more and more competitive.

Therefore, we test our four hypotheses using bootstrap: we resample our data 500 times and estimate the model on each sample. This procedure takes about 5 hours using 36 processors. The distribution of the parameters we obtain in this way are then used to build confidence intervals and to test our hypotheses.

Hypothesis 1: The hypothesis that our mixture model fits income data significantly better is tested using differences in Bayesian Information Criteria between the lognormal model (null) and the Max model (alternative), reproduced in the table below. The standard deviations of parameter estimates (in parentheses) are obtained by bootstrap.

	Max(P,LN) model				Lognormal Only		
year	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\alpha}$	$\hat{Q}$	$\hat{\mu}$	$\hat{\sigma}$	ΔBIC
1960	8.42	0.63 (0.005)	13.28 (1.28)	1.79% (0.48%)	8.40	0.65	2028***
1962	8.48	0.65 (0.005)	13.74 (1.5)	1.8% (0.51%)	8.46	0.67	2285***
1964	8.57	0.67 (0.003)	15.29 (1)	1.62% (0.27%)	8.55	0.68	2583***
1966	8.64	0.7 (0.004)	11.15 (0.69)	1.92% (0.44%)	8.62	0.72	2190***
1967	8.70	0.69 (0.004)	11.75 (0.72)	2% (0.43%)	8.68	0.71	2451***
1968	8.76	0.7 (0.005)	10.77 (0.64)	2.25% (0.46%)	8.74	0.72	2247***
1969	8.82	0.71 (0.005)	10.82 (0.74)	2.14% (0.51%)	8.80	0.73	2368***
1970	8.93	0.67 (0.005)	10.1 (0.83)	2.32% (0.47%)	8.91	0.69	1735***
1971	8.99	0.67 (0.006)	9.77 (0.88)	2.34% (0.55%)	8.96	0.69	1610***
1972	9.04	0.68 (0.006)	11.17 (1.42)	2.01% (0.56%)	9.01	0.70	1553***
1973	9.08	0.71 (0.008)	12.74 (1.74)	2.37% (0.77%)	9.05	0.73	2445***
1975	9.20	0.73 (0.004)	8.9 (0.45)	2.95% (0.41%)	9.16	0.75	1773***
1976	9.27	0.73 (0.003)	9.43 (0.34)	3.03% (0.28%)	9.23	0.76	1915***
1977	9.33	0.74 (0.003)	8.85 (0.35)	3.1% (0.32%)	9.29	0.77	1796***
1978	9.41	0.74 (0.003)	9.27 (0.49)	3.46% (0.34%)	9.36	0.78	1881***
1979	9.50	0.73 (0.006)	8.01 (0.71)	3.68% (0.62%)	9.46	0.76	1305***
1980	9.57	0.75 (0.002)	8.25 (0.28)	3.5% (0.23%)	9.52	0.78	1470***
1981	9.64	0.77 (0.004)	7.63 (0.43)	3.58% (0.44%)	9.60	0.80	1364***
1983	9.71	0.81 (0.003)	7.74 (0.3)	3.15% (0.31%)	9.66	0.84	1449***
1984	9.76	0.81 (0.003)	7.84 (0.36)	3.27% (0.35%)	9.72	0.85	1464***
1985	9.80	0.82 (0.003)	7.35 (0.3)	3.1% (0.29%)	9.76	0.85	1218***
1986	9.83	0.84 (0.005)	6.92 (0.44)	3.22% (0.47%)	9.79	0.86	1212***
1987	9.83	0.87 (0.003)	7.22 (0.23)	2.73% (0.26%)	9.79	0.89	1337***
1988	9.86	0.88 (0.003)	7.48 (0.3)	2.65% (0.32%)	9.82	0.91	1402***
1989	9.89	0.89 (0.005)	7.27 (0.49)	2.72% (0.49%)	9.85	0.92	1415***
1990	9.93	0.89 (0.004)	7.34 (0.36)	2.71% (0.36%)	9.89	0.92	1345***
1991	9.95	0.9 (0.005)	7.04 (0.52)	2.8% (0.51%)	9.91	0.93	1272***
1992	9.99	0.9 (0.004)	6.83 (0.32)	3.02% (0.38%)	9.95	0.93	1237***
1993	10.03	0.88 (0.004)	6.24 (0.36)	3.23% (0.43%)	9.98	0.91	959***
1994	10.07	0.88 (0.003)	6.56 (0.31)	3.32% (0.32%)	10.02	0.91	1047***
1995	10.09	0.89 (0.004)	6.54 (0.35)	3.36% (0.42%)	10.04	0.93	1073***
1996	10.11	0.9 (0.004)	6.63 (0.29)	3.07% (0.37%)	10.07	0.93	1064***
1997	10.16	0.9 (0.003)	6.34 (0.26)	3.45% (0.35%)	10.11	0.93	986***
1998	10.21	0.89 (0.004)	6.51 (0.38)	3.47% (0.4%)	10.16	0.92	893***
1999	10.24	0.9 (0.004)	6.6 (0.34)	3.3% (0.38%)	10.19	0.94	905***
2000	10.28	0.9 (0.006)	6.59 (0.59)	3.46% (0.61%)	10.23	0.94	888***
2001	10.32	0.9 (0.003)	6.07 (0.23)	3.59% (0.3%)	10.27	0.93	764***
2002	10.34	0.89 (0.003)	6.28 (0.24)	3.5% (0.29%)	10.29	0.92	838***
2003	10.35	0.89 (0.003)	6.78 (0.28)	3.33% (0.3%)	10.30	0.93	982***
2004	10.38	0.9 (0.003)	6.01 (0.22)	3.58% (0.3%)	10.33	0.93	729***
2005	10.40	0.9 (0.005)	6.27 (0.46)	3.52% (0.52%)	10.35	0.94	791***
2006	10.42	0.92 (0.004)	6.5 (0.37)	3.17% (0.44%)	10.37	0.96	861***
2007	10.44	0.92 (0.003)	5.84 (0.24)	3.56% (0.29%)	10.39	0.96	663***
2008	10.46	0.93 (0.003)	6.03 (0.23)	3.55% (0.31%)	10.41	0.97	768***

Table 2: MLE estimates of the mixture model and Likelihood Ratio Statistic when compared to the simple lognormal model

Using a bootstrap approach, we determine the 0.1% critical value for the ΔBIC statistic for each year. We do not use likelihood ratio tests because of the presence of sample

weights in the data, which may lead us to reject the null too often. We obtain BIC critical values by simulating 10,000 datasets drawn from the null distribution, and computing the difference in the BIC of the two models. For year 2008, it is equal to 310. The null model is rejected for all years at the 0.1% level, leading us to conclude that the Max model fits the data significantly better than a simple lognormal. In other words, focusing on the industrial economy only would mean missing an important part of the dynamics of income.

Hypothesis 2A (σ increasing)**:** Increased σ (standard deviation of the lognormal distribution underlying the industrial economy) is a prediction of skill-biased technical change. Out of 500 bootstrap estimations, the intersection of the interval of estimated σ for 1960 and the interval of estimated σ for 2008 is empty. This gives us an estimated p-value under 0.002, and we can conclude with confidence that σ has significantly increased between 1960 and 2008.

Hypothesis 2B (regime change in σ)**:** Figure 6 above illustrates the evolution of the log-normal σ over time. Visually, the data estimates seem to exhibit a clear break in trends in the 1980s, with the slope of the evolution of σ declining sharply. More formally, we can test for a structural beak by regressing our estimated σ (from our bootstrap sample of 500 estimates of σ per year) on time and performing a Chow test. Testing for a structural break in year 1988 yields a very large F-statistic of 1464.38, so we can reject the null that there is no trend change in sigma at the 0.1% level.

Hypothesis 3 (q increasing)**:** Q represents the number of individuals who select into the digital economy. Out of 500 bootstrap estimations, the intersection of the interval of estimated q for 1960 and the interval of estimated q for 2008 is empty. This gives us an

estimated p-value under 0.002, and we can conclude with confidence that q has significantly increased between 1960 and 2008. Alternatively, we can simply look at the 95% confidence intervals in Figure 7, and see that that the top of the confidence interval for 1960 is significantly below the bottom of the confidence interval for 2008.

Hypothesis 4 (α decreasing)**:** α measures how fat the tail of the Pareto distribution is. A lower α means fatter tails and more skewed rewards. Out of 500 bootstrap estimations, the intersection of the interval of estimated alphas for 1960 and the interval of estimated alphas for 2008 is empty. This gives us an estimated p-value of under 0.002, and we can conclude with confidence that α has significantly decreased between 1960 and 2008.

Conclusion

Our results suggest that as the economy becomes more and more digitized and networked, the income distribution is changing dramatically. While the idea that increased productivity may benefit everyone in the very long term is certainly plausible, in this paper we show that this need not be the case in the short or medium term. The key to understanding inequality seems to lie in understanding the underlying market mechanisms yielding highly skewed income distributions.

We find that (1) the distribution of income is better approximated by a model where individuals choose optimally between draws from a Pareto distribution (digital economy) and Log-normal distribution (industrial economy) than by a Log-normal distribution alone. Using this model, we find that the evolution of the shape of the income distribution is consistent with (2) a leveling off of the variance of the underlying lognormal

distribution in the industrial economy, which is consistent with the hypothesis that skill-biased technical change is no longer accelerating as it did earlier in the period; (3) an increased number of individuals selecting into the digital economy which generates a power law distribution of income, and (4) a digital economy delivering increasingly skewed rewards: the winners are winning bigger than ever.

More and more people moving into the digital economy does not necessarily mean more and more people being better off in absolute terms because of digitization. Indeed, our model accommodates both people moving to the digital economy because of increasing opportunity there and people moving because of decreasing opportunity in the industrial economy. Furthermore, we show that the digital economy is producing increasingly unequal rewards.

If the trends we document in this paper continue, the income distribution will become increasingly skewed toward not just the top 1%, but the top 1% of the top 1%. Future research is needed in order to more causally distinguish between different types of technologies and their effects on inequality. Furthermore, the effects of digitization on worldwide inequality patterns seems like a worthwhile extension of our research.

Appendix

As an alternative to our Max model we also use a "spliced" model of the Pareto/Log-Normal mixture. Specifically, we model the income distribution with the following PDF:

$$f(x) = \begin{cases} \beta \dfrac{f_1(x|\mu,\sigma)}{F_1(\theta|\mu,\sigma)} \forall\, x \leq \theta \\ (1-\beta)\dfrac{f_2(x|c,\alpha)}{1-F_2(\theta|c,\alpha)}, \forall\, x > 0 \end{cases}$$

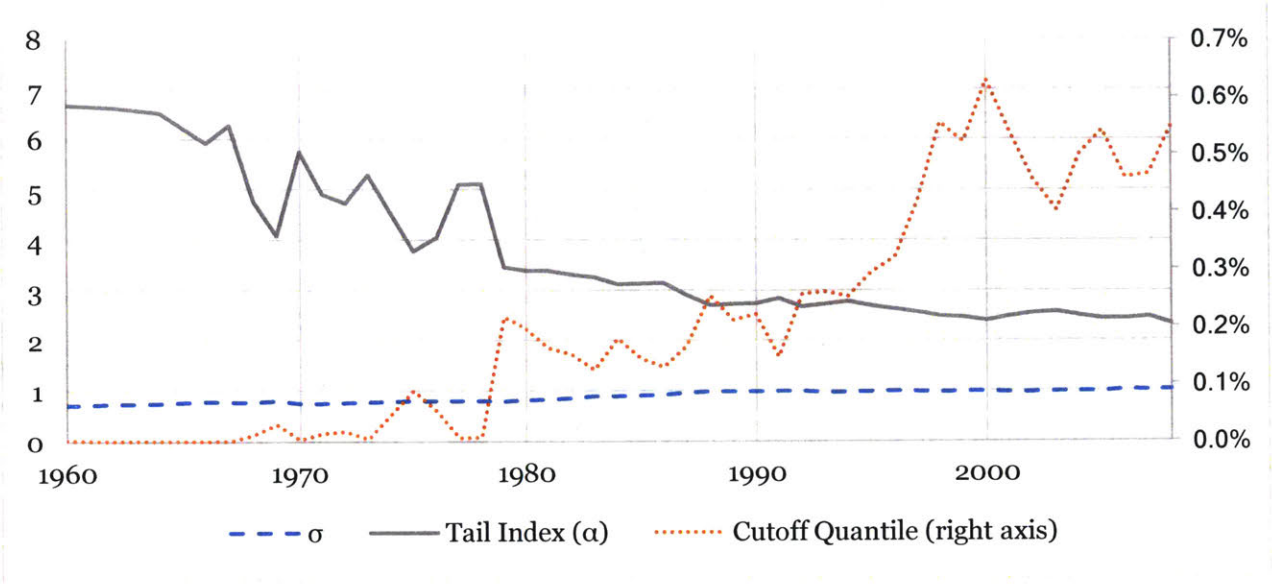
Were $f_1(x|\mu,\sigma)$ is the probability distribution function of a lognormal distribution with parameters μ and σ, and $f_2(x|c,\alpha)$is the p.d.f of a simple, type I Pareto distribution with cutoff c and shape α. β is a weighting coefficient. Note that this distribution is not in general continuous or differentiable, which may be undesirable for modelling purposes, as the empirical income distribution function does not seem to exhibit any major discontinuity. However, ensuring continuity is possible by placing restrictions on μ and β (Nadarajah and Bakar 2012). This requires imposing the following:

$$\mu = \ln(\theta) + \sigma^2 + \frac{\theta\sigma^2 f_2'(\theta|c,\alpha)}{f_2(\theta|c,\alpha)}$$

$$\phi = \frac{f_1(\theta|\mu,\alpha)(1-F_2(\theta))}{f_2(\theta)F_1(\theta)}$$

(with notation change $\beta = \frac{1}{1+\phi}$).

This setup gives us a continuous and differentiable p.d.f. Note that the likelihood function is not in general continuous at θ, so that MLE is consistent but not necessarily efficient (Bee 2012). MLE estimation (fixing c at the 25th percentile) gives the following results:



The move in the number of individuals is in the Pareto economy is lower. This is expected, since this model equates "being in the Pareto" and "being at the top", whereas the Max model allows for individuals with low income to self-select into the digital economy (because their industrial income would be lower still).

References

Acemoglu, D., and Autor, D. 2012. "What Does Human Capital Do? A Review of Goldin and Katz's *The Race between Education and Technology*," *Journal of Economic Literature* (50:2), pp. 426–463 (doi: 10.1257/jel.50.2.426).

Acemoglu, D., and Robinson, J. 2012. "Why Nations Fail: The Origins of Power," *Prosperity, and Poverty, Crown Business, New York*.

Andreessen, M. 2011. "Why Software Is Eating the World," *The Wall Street Journal* (available at http://www.wsj.com/articles/SB10001424053111903480904576512250915629460).

Andriani, P., and McKelvey, B. 2009. "Perspective—From Gaussian to Paretian Thinking: Causes and Implications of Power Laws in Organizations," *Organization Science* (20:6), INFORMS, pp. 1053–1071 (doi: 10.1287/orsc.1090.0481).

Atkinson, A. B., and Piketty, T. 2010. *Top incomes: A global perspective*, Oxford University Press.

Autor, D. H. 2014. "Skills, education, and the rise of earnings inequality among the 'other 99 percent,'" *Science (New York, N.Y.)* (344:6186), pp. 843–51 (doi: 10.1126/science.1251868).

Autor, D. H., Katz, L. F., and Kearney, M. S. 2008. "Trends in U.S. Wage Inequality: Revising the Revisionists," *Review of Economics and Statistics* (90:2), pp. 300–323 (doi: 10.1162/rest.90.2.300).

Autor, D. H., Katz, L. F., and Krueger, A. B. 1998. "Computing Inequality: Have Computers Changed the Labor Market?," *Quarterly Journal of Economics* (113:4), pp. 1169–1213 (doi: 10.1162/003355398555874).

Barabási, A.-L., and Albert, R. 1999. "Emergence of scaling in random networks," *Science* (286:5439), American Association for the Advancement of Science, pp. 509–512.

Bee, M. 2012. "Statistical analysis of the Lognormal-Pareto distribution using Probability Weighted Moments and Maximum Likelihood," *Department of Economics Working Papers*, Department of Economics, University of Trento, Italia (available at http://ideas.repec.org/p/trn/utwpde/1208.html).

Bernstein, J. 2010. "Three Questions About Consumer Spending and the Middle Class," *Bureau of Labor Statistics* (available at http://www.bls.gov/cex/duf2010bernstein1.pdf; retrieved May 3, 2015).

Bivens, J., and Mishel, L. 2013. "The Pay of Corporate Executives and Financial Professionals as Evidence of Rents in Top 1 Percent Incomes," *The Journal of Economic Perspectives* (27:3), pp. 57–78 (doi: 10.2307/41955545).

Bresnahan, T. F., Brynjolfsson, E., and Hitt, L. M. 2002. "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence," *The Quarterly Journal of Economics* (117:1), Oxford University Press, pp. 339–376 (doi: 10.2307/2696490).

Brynjolfsson, E., Hu, Y. (Jeffrey), and Smith, M. D. 2003. "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science* (49:11), INFORMS, pp. 1580–1596 (doi: 10.1287/mnsc.49.11.1580.20580).

Brynjolfsson, E., McAffee, A., and Spence, M. 2014. "New World Order: Labor, Capital, and Ideas in the Power Law Economy," *Foreign Affairs* (doi: 10.3318/ISIA.2004.15.1.3).

Corak, M. 2013. "Income Inequality, Equality of Opportunity, and Intergenerational Mobility," *The Journal of Economic Perspectives* (27:3), American Economic Association, pp. 79–102 (doi: 10.2307/41955546).

Debacker, J., Heim, B., Panousi, V., Ramnath, S., and Vidangos, I. 2013. "Rising Inequality: Transitory or Persistent? New Evidence from a Panel of U.S. Tax Returns," *Brookings Papers on Economic Activity*, Brookings Institution Press, pp. 67–122 (doi: 10.2307/23594863).

DiNardo, J., Fortin, N. M., and Lemieux, T. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica* (64:5), The Econometric Society, pp. 1001–1044 (doi: 10.2307/2171954).

Fleder, D., and Hosanagar, K. 2009. "Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity," *Management Science* (55:5), INFORMS, pp. 697–712 (doi: 10.1287/mnsc.1080.0974).

Gabaix, X., and Landier, A. 2008. "Why Has CEO Pay Increased so Much?," *The Quarterly Journal of Economics* (123:1), Oxford University Press, pp. 49–100 (doi: 10.2307/25098894).

Gibrat, R. 1931. *Les inégalités économiques*, Recueil Sirey.

Hajargasht, G., and Griffiths, W. E. 2013. "Pareto--lognormal distributions: Inequality, poverty, and estimation from grouped income data," *Economic Modelling* (33), Elsevier, pp. 593–604.

Jones, R., and Mendelson, H. 2011. "Information Goods vs. Industrial Goods: Cost Structure and Competition," *Management Science*, INFORMS (available at http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1100.1262).

Kaplan, S. N., and Rauh, J. 2013. "It's the Market: The Broad-Based Rise in the Return to Top Talent," *Journal of Economic Perspectives* (27:3), pp. 35–56 (doi: 10.1257/jep.27.3.35).

Katz, L., and Autor, D. 1999. "Changes in the Wage Structure and Earnings Inequality," *Handbook of Labor Economics* (3), pp. 1463–1555 (doi: 10.1016/S1573-4463(99)03007-2).

Katz, L. F. 2000. "Technological change, computerization, and the wage structure," in *Understanding the Digital Economy: Data, Tools, and Research*E. Brynjolfsson and B. Kahin (eds.), MIT Press, Cambridge MA.

Kim, H., and Brynjolfsson, E. 2009. "CEO Compensation and Information Technology," in *ICIS 2009 Proceedings* (available at http://aisel.aisnet.org/icis2009/38).

Krueger, A. B. 2013. "Land of Hope and Dreams: Rock and Roll, Economics and Rebuilding the Middle Class," *Speech at the White House*.

Kuznets, S. 1955. "Economic growth and income inequality," *The American Economic Review* (45:1), American Economic Association, pp. 1–28 (doi: 10.2307/1811581).

Lopez, H., and Serven, L. 2006. "A normal relationship? Poverty, growth, and inequality," *Poverty, Growth, and Inequality (January 2006). World Bank Policy Research Working Paper* (3814).

Luceño, A. 2006. "Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators," *Computational Statistics & Data Analysis* (51:2), pp. 904–917 (doi: 10.1016/j.csda.2005.09.011).

Mincer, J. 1958. "Investment in Human Capital and Personal Income Distribution," *Journal of Political Economy* (66:4), The University of Chicago Press, pp. 281–302 CR – Copyright © 1958 The University (doi: 10.2307/1827422).

Nadarajah, S., and Bakar, S. A. A. 2012. "New composite models for the Danish fire insurance data," *Scandinavian Actuarial Journal* (2014:2), Taylor & Francis, pp. 180–187 (doi: 10.1080/03461238.2012.695748).

Pigou, A. C. 1932. *The economics of welfare*, Macmillan.

Piketty, T. 2001. *Les hauts revenus en France au 20ème siècle*.

Piketty, T., and Saez, E. 2003. "Income Inequality in the United States, 1913--1998," *The Quarterly Journal of Economics* (118:1), pp. 1–39 (doi: 10.1162/00335530360535135).

Reed, W. J., and Jorgensen, M. 2004. "The double Pareto-lognormal distribution -- a new parametric model for size distributions," *Communications in Statistics-Theory and Methods* (33:8), Taylor & Francis, pp. 1733–1753.

Rosen, S. 1981. "The Economics of Superstars," *The American Economic Review* (71:5), American Economic Association, pp. 845–858 (doi: 10.2307/1803469).

Saez, E. 2013. "Striking it Richer: The Evolution of Top Incomes in the United States," *University of California at Berkley working paper* (available at http://eml.berkeley.edu/~saez/saez-UStopincomes-2012.pdf).

THIS PAGE INTENTIONALLY LEFT BLANK

CEO Pay and Information Technology

We find that information technology (IT) intensity predicts the compensation of CEOs and other top executives, the dispersion of their pay, and their mobility. We explore three possible explanations:

1. IT may primarily affect CEO and executive pay by changing firm size through "winner-take-all" effects.
2. IT may increase the "effective size" of the CEO's firm by making performance more sensitive to the decisions of top executives, even with firm size held constant.
3. IT may thereby broaden the market for top executives by increasing the generality of skills required to be an effective top executive.

We examine panel data from 3413 publicly traded firms over 23 years, controlling for other types of capital, number of employees, market capitalization, industry turbulence, firm or industry fixed effects, and other factors and find the strongest evidence for the second and third hypotheses. In particular, industry IT intensity can help predict not only the level of CEO pay, but also the dispersion of CEO pay, and the mobility of executives.

"The [IT] dashboard is the CEO's killer app, making the gritty details of a business that are often buried deep within a large organization accessible at a glance to senior executives. ... Managers can see key changes in their businesses almost instantaneously -- and take quick, corrective action."

– Ante (2006)

1 Introduction

This paper examines the relationship between information technology (IT) and top executives' pay. A substantial rise in top executives' pay in the 1990s has been well documented (Bebchuk & Fried, 2005; Bebchuk & Grinstein, 2005; Frydman & Jenter, 2010; Frydman & Saks, 2007; Hall & Liebman, 1998; Hall & Murphy, 2003). For instance, the ratio of CEO pay to average worker pay increased from 60:1 in 1990 to 380:1 in 2000. Less publicized is that fact that the levels of top CEO pay in publicly traded companies have been relatively flat since 2002, while the median CEO pay has been still increasing (**Figure 11**). Since early 2000s, the ratio of CEO pay to median work pay has fluctuated between 180 and 350 (Mishel & Davis, 2014).

What explains this pattern? While increases in the size of firms may explain a part of the increase (Figure 2), another part of the story may be the changes in IT intensity over the same period. Corporations can be thought of as information processors. Hence, large declines in the costs of digital information processing are likely to affect monitoring and control capabilities. In particular, the increases in the quality and quantity of IT have changed the types of skills needed to be an effective C-level executive and have radically increased the ability of top executives to keep informed about activities throughout their organizations and to respond more quickly and precisely with instructions. This may have affected the level and dispersion of their pay, as well as their mobility. Our study provides a first look at the potential role of IT in top executive compensation and mobility.

2 Related Research

Our study is related to three streams of literature: one is the effect of IT on centralization and decentralization of decision-making, the second is the effect of IT on income inequality, and the third is the rise of CEO compensation.

2.1 IT and Organization

A large stream of literature studies effects of IT on command, control, coordination, and organization of firms. In theory, IT could shift power either toward the center or away from it, leading to centralization or decentralization of a firm, respectively. In the former scenario, IT makes local knowledge directly available to top managers and enables managers to quickly route messages directly to distant subordinates, which enables management to be more centralized. In this case, one might expect higher CEO relative pay. On the other hand, in the latter scenario, IT makes local knowledge of one department available to employees (as well as to top managers) in other departments, and employees can coordinate tasks among themselves more easily without the need for CEO involvement. In this case, one might expect lower CEO relative pay.

Ultimately, the net effect is an empirical question. Previous studies have found evidence of both effects, albeit during earlier time periods in which the technology and institutions were different than those that prevail more recently (Attewell & Rule, 1984; Bresnahan, Brynjolfsson, & Hitt, 2002; Brynjolfsson & Mendelson, 1993; Leavitt & Whisler, 1958). Changes in firm size induced by IT have been also been predicted by Brynjolfsson, Malone, Gurbaxani, & Kambil (1994) and by Gurbaxani & Whang (1991), drawing on the economic theories of Coase (1937), Williamson (1973, 1981) and

Grossman & Hart (1986). In short, there are theoretical reasons to expect that IT will change equilibrium firm size or CEO compensation but no consensus on the direction of these changes.

2.2 IT and Wage Inequality

The second related stream of literature is the role of IT in inducing wage inequality in the whole economy (e.g., Autor, Katz & Krueger, 1998). This can lead to effects of IT on CEO pay in two ways. The first is closely related to those of Garicano & Rossi-Hansberg (Garicano, 2000; Garicano & Rossi-Hansberg, 2006). They argue that knowledge has hierarchy; some knowledge attached to lower-ranked employees is used to solve routine problems, while other knowledge attached to top managers is used to process more convoluted non-routine problems. As the cost of communication among agents decreases, the complicated problems that lower-ranked employees cannot solve can be more easily passed to their superiors. Once solved, the solution can be disseminated more easily with the lowered cost of communication. This can lead to the dependency of the problem-solving on a few "superstars" (Rosen, 1981) and thus a higher wage for the superstars. This explanation is consistent with our view that IT increases the effective size of a firm by easily passing non-routine problems to a few problem solvers and then disseminating the solutions to the entire firm. Therefore, the dependency of the IT-intense firm on the few problem solvers and thus the sensitivity of the IT-intense firm to those few superstars become greater, leading to a higher compensation for those superstars in IT-intense firms.

Another thread in this literature is that IT often shifts knowledge requirements from the firm-specific or industry-specific to the more general. For example, companies in

virtually every industry have adopted Enterprise Resource Planning (ERP) software, which provides detailed data on operations to executives. According to one estimate, spending on these enterprise IT platforms at one point accounted for more than 50% of all U.S. corporate IT investment in 2001 (McAfee, 2002). ERP stores, manages, and makes accessible a wide range of corporate data on sales orders, inventories, and many other resource data. This data is potentially accessible to anyone, anywhere in the firm, to whom top management grants access . Moreover, computers contain codified technological knowledge and made high-level problem solvers increasingly important in steel, semiconductor and mechanical sectors (Balconi, 2002). Then, optimal resource management can be done by a few data analysts who may not necessarily have firm-specific or even industry-specific knowledge. However, they do have general analytic skills to process and analyze the codified technological knowledge or the stored data. The use of data in management, therefore, increases the demand for general skills more than for firm- or industry-specific skills. Our model can explain greater wage inequality *within* the same firm, but it's also possible that some firms or industries can concentrate many problem-solvers or superstars for the whole economy. The wage inequality between such firms may be an important part of the overall inequality in the economy. Our study focus is the differences of CEO pay across time and industry, not only the wage inequality within the same firm, as a result of the increased IT intensity.

2.3 CEO Pay

The third stream of literature relevant to our study examines CEO compensation in particular. Academia and the mass media have proposed a wide spectrum of competing explanations for the rise in CEO pay. One end of the spectrum is a rent extraction

view in which CEOs can set their own pay and extract rents from their firms (Bebchuk, Fried, & Walker, 2002; Bebchuk & Fried, 2005; Bertrand & Mullainathan, 2001; Hall & Murphy, 2003; Yermack, 1997) while the other end of the spectrum is a competitive market view in which firms optimally compete for managerial talent in an efficient labor market (e.g., Frydman & Jenter, 2010; Gabaix & Landier, 2008). Although there is evidence for both views, and they are not mutually exclusive, in this paper we more closely pursue the second view and refer to other papers to review the first view.

The competitive market view can be further categorized by four sets of theories (Frydman & Jenter, 2010) that we recapitulate here. The first is the scale effects: the rise in CEO pay is due to the increase in firm sizes. The marginal productivity of CEOs is higher for larger firms, and a larger firm is willing to pay higher pay for a CEO even with a smaller incremental talent compared to the next talented CEO (Gabaix & Landier, 2008; Himmelberg, Hubbard, & Palia, 1999; Rosen 1981, 1996; Tervio, 2008).

The second market-based explanation is the change in the type of managerial skills from firm- or industry-specific to general (Frydman & Saks, 2010; Murphy & Zabojnik, 2004). This change increases firms' competition for talent and consequently the pay of top CEOs.

The third set of theories is an agency view (Baker & Hall, 2004; Dow & Raposo, 2005; Himmelberg et al., 1999; Holmstrom & Kaplan, 2001; Hubbard & Palia, 1995; Jensen & Murphy, 1990) In order to give incentives or reward to CEOs to cope with more volatile and uncertain business environments and globalization, raise the CEO's optimal effort, prevent moral hazard, and bring more innovative idea in the business strategy,

firms must link CEO pay more closely to performance, which in turn can lead to higher levels and dispersion of CEO pay insofar as business has become more globalized and turbulent.

Finally, a fourth market-based theory argues that a stricter corporate governance and a better monitoring of CEOs decrease their job stability. Firms optimally respond by increasing the CEO pay (Hermalin, 2005).

Our paper investigates the correlation between IT and CEO pay via three distinct mechanisms. First, IT may affect the size and size distribution of firms and market value. If IT increases the equilibrium size of firms, or the dispersion in the size of firms, then CEO pay may change in a corresponding way.

Second, the marginal productivity of a CEO may depend not on nominal firm size but on its *effective* size. Effective size may be an increasing function of IT intensity of the firm because the more IT-intensive environment a firm has, the lower the cost of both communication and access of knowledge, as articulated by Garicano & Rossi-Hansberg (2006). As a result, the relationship between CEO performance and firm performance becomes stronger; that is, the marginal productivity of a CEO increases. Changes in CEO pay may be, therefore, at least in part a manifestation of changes in the marginal productivity of CEOs.

Third, we consider the idea that IT can facilitate an increase of the generality of managerial skills. CEOs' ability to make decisions based on quantitative data are likely more portable across industries than their ability to make decisions based on local knowledge, experience or intuition. As CEOs' skills become more general, talented

CEOs have more options outside their firms or even their industries. Therefore, firms compete more for such a talent and consequently pay more for top talent.

These three hypotheses have different predictions about the *level* vs. the *dispersion* of CEO pay, as well as CEO mobility. The first hypothesis attributes changes in the levels and dispersion of CEO pay to commensurate changes in the levels of firm size and dispersion of firm size, respectively. In other words, changes in the dispersion among CEO pay and the average CEO pay should be fully explained by changes in the dispersion among firm sizes and the average firm size. The second hypothesis, which considers IT-induced changes in the "effective size," predicts that increases in IT would be important for the level of CEO pay even after controlling for firm size. Finally, controlling for firm size, the general skill hypothesis predicts that the demand for generality skills in the CEO market will increase the dispersion among CEO pay, not just its level. The top CEO will be much more highly paid than less talented CEOs beyond what can be explained by firm size dispersion. Furthermore, it predicts increased mobility for CEOs of IT-intensive industries. These predictions are summarized in Figure 3.

To examine the robustness of our models, we include industry turbulence as a control variable. Some researchers have reported that IT-intensive industries tend to be more turbulent than others (Brynjolfsson, McAfee, Sorell, & Zhu, 2007). It may be that firms in turbulent industries face more competitive business environments and benefit disproportionately from hiring more talented and thus more expensive CEOs. The increased CEO pay due to IT intensity may be a result of the more competitive business environment that the firm faces in the industry, not the increased effective size of the

firm. To address this possibility, industry turbulence, as defined as the average of rank changes of firms from year to year over industry, is included as a control variable.

3 Three Models of IT's Role

We explore three hypotheses that would link IT intensity to CEO pay.

3.1 IT and Firm Size

IT may affect the size or value of firms by changing monitoring costs (Garicano & Rossi-Hansberg, 2004, 2006; Garicano, 2000), increasing the span of control (Guadalupe, Wulf, & Boston, 2008; Rajan & Wulf, 2006) or affecting firm boundaries (Brynjolfsson, Malone, Gurbaxani, & Kambil, 1994). In addition to the value or size of individual firms, the size distribution of firms within an industry may be affected by IT. If IT facilitates winner-take-all markets (e.g., Brynjolfsson, McAfee, & Spence, 2014), then it may increase the Gini coefficient of firm size.

In turn, many researchers have reported that CEO pay is highly correlated with firm size and market value (Barro & Barro, 1990; Gabaix & Landier, 2008; Kostiuk, 1990; Roberts, 1956; Rosen, 1992). Gabaix & Landier (2008) and Tervio (2008) each propose a model in which the best CEO manages the largest firm at competitive equilibrium because this maximizes the CEO impact. In other words, the CEO of the largest firm has the highest marginal productivity and thus receives the highest pay.

A version of this mechanism can be summarized in this simple model inspired by Lucas[3] (1978):

- Managerial talent T is given at birth, and complemented by IT which is denoted by θ
- Each worker can choose to be a manager or a worker (earning w)
- Firms produces value with one manager and L workers according to:

$$V = \frac{1}{1-\alpha}[(\theta T)^{\alpha} L^{1-\alpha}]$$

In the equilibrium of this simple model:

- A manager with talent T will hire $L = \theta T w^{-\frac{1}{\alpha}}$ workers
- She earns $w_m = V^* - wL^* = \boldsymbol{\alpha V}$
- Individuals born with low talent become workers, with $w = (1-\alpha)\frac{V^*}{L^*}$

When IT (θ) is larger:

- Firms are bigger and CEO pay is larger
- Variance of firm size is higher, and so is the variance of CEO pay

Thus, IT may indirectly affect the level and dispersion of CEO pay by affecting the size of firms and dispersion of firm size. In this model, IT influences CEO pay *through* firm size.

---

[3] We thank Lowell Taylor for contributing this model.

3.2 IT and Effective Firm Size

Even if nominal firm size is held constant, IT may still increase the "effective" firm size relevant to CEO pay by increasing the CEO's influence and control. We define the concept of effective firm size as the extent of the firm that the CEO can effectively influence and control. The ability of CEOs to manage large firms depends on technology for communicating instructions, replicating processes, and monitoring employees. The less perfect the communication between top managers and employees is, the less effective the CEO's monitoring and influence becomes. For instance, if a CEO's instructions are propagated throughout only a part of the firm, or with less than 100% accuracy, then the effective size is not as great as it would be if the instructions were reliably and accurately propagated to every part of the firm. Similarly, the ability of the CEO to use IT to better monitor compliance with instructions will also increase the effective size of the firm. The firm's benefits of having a high-quality CEO will be greater for firms with a bigger effective size of the firm. If IT increases the effective size of firm, then IT-intensive firms, and those recruiting from IT-intensive industries, might pay their CEOs more than those of the same nominal size but with less IT.

We think of this channel in the framework of Gabaix & Landier (2008), replacing firm size with effective size, and we extend their model to include the following:

1. CEOs have different levels of pay and managerial talent and are matched to firms competitively;
2. In equilibrium, the best and thus the highest paid CEO manages the largest firms, as this maximizes the CEO's impact and economic efficiency; and
3. CEO pay also increases with the average size of firm in the economy.

In other words, if there are two firms of different sizes and two managers of different talent, the original concept has a competitive equilibrium in which the larger firm hires the more talented manager at a higher pay than the smaller firm does. We extend the concept of firm size in their theory to an IT-enabled "effective" size of firm, defined as the maximum firm size that top managers can control and reach because IT has integrated firm data and enabled a replicable, speedy, and firm-wide business process aided by an enterprise IT system.

We briefly walk through Gabaix & Landier's model here. Consider the problem of hiring a CEO with talent, T, faced by a particular firm. The firm has "baseline" earnings of a_0. At t=0, it hires a manager of talent T for one period. The manager's talent T increases the firm's earnings according to

$$a_1 = a_0(1 + C\times T) = a_0 + a_0\times C\times T \qquad (1)$$

for some $C > 0$, which quantifies the effect of talent on earnings. Consider one extreme case in which the CEO's actions at date 0 impact earnings only in period 1. The firm's earnings are $(a_1, a_0, a_0, \ldots.)$. The other extreme case is that the CEO's actions at date 0 impact earnings permanently. Then the earnings become $(a_1, a_1, a_1, \ldots)$. In both cases, the firm's problem can be written as the following:

$$\max_T S\times(1 + C\times T) - W(T) = \max_T S + S\times C\times T - W(T) \qquad (2)$$

where $S = \frac{a_0}{1+r}$ for the former (where a CEO's talent impacts the firm's earnings only the first period) or $\frac{a_0}{r}$ for the latter (where a CEO's talent impacts the firm's earnings permanently), r is the discount rate, and $W(T)$ is the wage of CEO with talent, T. Eqn (1)

can be generalized as $a_1 = a_0 + Ca_0^{\gamma} + independent\ factors$, for a non-negative γ. The maximization problem of (2) becomes:

$$\max_T(S + S^{\gamma} \times C \times T - W(T)) \tag{3}$$

Let's call $w(m)$ the equilibrium compensation of each CEO with index m, which can be thought of as the CEO's ranking or quantile in talent. The problem of (3) can be re-written as:

$$\max_m(CS(n)^{\gamma}T(m) - w(m)) \tag{4}$$

A competitive equilibrium consists of:

- a compensation function $W(T)$, which specifies the market pay of a CEO of talent T;
- an assignment function $M(n)$, which specifies the index $m = M(n)$ of the CEO heading firm n in equilibrium;
- an assumption that each firm chooses its CEO optimally: $M(n) \in arg\max_m(CS(n)^{\gamma}T(m) - W(T(m))$; and
- the constraint that the CEO market clears; that is, each firm gets a CEO.

As in equilibrium there is associative matching: $m = n$,

$$w(n) = \int_N^n CS(u)^{\gamma}T'(u)du + w(N) \tag{5}$$

Assuming a specific functional form for $T'(u)$ with the use of the extreme value theory, Gabaix & Landier provide the solution for a CEO's pay in terms of the size of a reference firm as well as the CEO's firm. The most relevant equation for our study is expressed in terms of the effective size of firm as following[4]:

$$w = D_* \hat{S}_*^{\alpha} \hat{S}^{\beta} \tag{6}$$

where w is CEO pay, $\hat{S}$ and $\hat{S}_*$ are the effective size of the CEO's firm and a reference firm, respectively, and α and β are positive constants. D_* is a function of the marginal talent of CEO, $T'(n_*)$ of the reference firm and the size of the reference firm. In all equations, the subscript * indicates attributes for a reference firm.

The effective size of a firm in Gabaix & Landier's model is a function of the sensitivity of the firm to CEO talent and the nominal size of the firm. We extend this concept further: CEO may be able to reach a greater portion of her firm if her firm is more integrated through its IT system. In other words, we hypothesize that the effective size (that CEO can affect) increases as a firm is more integrated through an IT system. This hypothesis reflects that IT reduces, to name a few, the cost of communication between top managers and employees, the cost of implementing new business processes, and the cost of monitoring employee performances. For example, a large retailer such as Walmart has adopted an enterprise IT system, and the inventories and sales data from approximately 4,000 retail stores in the USA alone can be accessed and analyzed at its

[4] This is the equation (25) of Proposition 3 by Gabaix & Landier.

headquarters in real time. Therefore, the effective firm size that top managers at its headquarters can reach has been enlarged, with the centralized IT system allowing the global access to data that were previously accessible only to local managers in the absence of such a centralized IT system. Therefore, we assume that the effective size is an increasing function of both IT and nominal size.

$$\hat{S} = cI^{\delta}S \tag{7}$$

where c is a constant, I is the IT intensity, δ is a constant, and S is the nominal size of the firm.

3.2.1 Effective Size and Level of CEO Pay

Various measures may be used for the nominal firm size, such as: the number of employees, sales, market capitalization, and assets. Following Gabaix & Landier, we use market capitalization as the proxy for the nominal size. The equations (6) and (7) yield:

$$w = D_* c_*^{\alpha} I_*^{\alpha\delta} S_*^{\alpha} c^{\beta} I^{\beta\delta} S^{\beta} = A_*^{\mu} I_*^{\varepsilon} S_*^{\alpha} c^{\beta} I^{\rho} S^{\beta} \tag{8}$$

where $A_* = D_* c_*$, and μ, ε, and ρ are constants.

The resulting empirically testable equation is the following:

$$\ln(w_{i,t}) = \beta_1 + \beta_2 \ln(A_{*t}) + \beta_3 \ln(I_{*t}) + \beta_4 \ln(S_{*t}) + \beta_5 \ln(I_{i,t}) + \beta_6 \ln(S_{i,t}) \tag{9}$$

where i and t indicate an index for firm and time, respectively. A_* is a variable relevant to a reference company, such as the marginal talent of the CEO of the reference company or the sensitivity of the reference firm to its CEO talent (not captured by the effective size). This measure captures business environments that the CEO's firm of interest is under, on the assumption that the CEO's firm would face a similar environment as its reference firm. For example, a firm in a highly-competitive industry may pay a premium for its CEO's talent because it takes more talent to win against the competition; a firm in an industry employing more educated workers may also reward its CEO more because it needs a highly educated CEO to understand the complexities of the tasks that the workers of his firm face. This equation implies that the compensation for a CEO this year ($w_{i,t}$) is determined by the effective size of a reference company (I_{*t} and S_{*t}) as well as the CEO's firm ($I_{i,t}$ and $S_{i,t}$) along with other characteristics associated with the reference company (A_{*t}). We can also introduce a time lag to reduce the potential simultaneity problem between CEO pay and the effective firm size.

We test this model in both firm-level and industry-level analyses. The conceptual unit of analysis of the model lies at firm-level; however, as we argue that an IT-intensive firm may increase its effective size and thus the marginal productivity of its CEO but we do not have firm-level but only industry-level IT intensity data, the industry level analysis may be more consistent with our measures of IT intensity.

3.2.2 Effective Size and Dispersion of CEO Pay

Because we model the log of the level of CEO pay, differences in log CEO pay levels (i.e., dispersion of CEO pay) can be written as the ratio of one CEO's pay, w(i) and the

other CEO's pay, w(j), which is, in turn, modeled as a function of the ratios of firm size and the IT intensity:

$$\frac{w(i)}{w(j)} = a\left(\frac{I(i)}{I(j)}\right)^{\rho}\left(\frac{S(i)}{S(j)}\right)^{\beta} \quad (2)$$

where a is a constant.

In the Gabaix & Landier (G&L) framework, which lacks the IT terms, the distribution of income maps directly onto the distribution of firm size. As such, a shift in the log of firm sizes is expected to shift the logs of all CEO salaries by the same amount, leaving inequality among CEOs unchanged. That is, the ratio of CEO pay will be explained only by the ratio of firm size. We hypothesize that IT plays an additional role in the dispersion of CEO pay.

First, IT affects the dispersion in CEO pay indirectly through the distribution of firm sizes. In the other words, firm size is determined partly by IT. In the equation (2), firm size S(i), is a function of IT intensity, I(i). However, it is possible for IT intensity to affect firm size differentially at the top and at the bottom of the size distribution, which would indirectly lead, according to G&L's framework, to an increase in CEO pay dispersion.

Second, IT can have a direct effect on CEO pay inequality. In the equation (2), the CEO pay ratio, $w(i)/w(j)$, is affected by the ratio of the IT intensity, $I(i)/I(j)$, as well as the ratio of firm size. However, an average of IT intensity of the whole economy or the industry would not necessarily affect the ratio of CEO pay. If the two firms, i and j, increase their IT intensity as well as their firm size by the same factor, this framework predicts the ratio of CEO pay should remain unchanged.

To summarize, if IT influences *effective* size, we expect to see the following where IT intensity is higher:

- we expect higher levels of CEO pay, *even after controlling for nominal firm size.*
- We expect higher intra-industry dispersion of CEO pay, *even after controlling for nominal firm size (and dispersion of nominal firm size)*

3.3 IT and the Generality of Managerial Skills

Information technology (IT) may be correlated with the increasing generality of the required managerial skills. In turn, as argued by Frydman (2005), an increase in generality of skills can lead to higher compensation for top executives.

Managerial decisions are increasingly becoming based on data more than experience. Data-driven decision-making practices are likely to make manager's skills transferrable across firms and industries, increasing the generality of the required managerial skill. For instance, Gary Loveman, CEO of Harrah's Entertainment Corp., successfully transformed that company by bringing to bear a set of analytical methods and quantitative "rocket scientists" from outside the industry. He had no special knowledge of the entertainment industry, and he has said that he could just as easily have brought the same techniques to any other industry.[5] Similarly, executives at GE are trained in a data-driven approach to management, including the concepts of "Total Quality Management" and "Six-Sigma" methods, which apply across a diverse set of industries. The CEOs are

---

[5] Presentation at "Economics of Information" class (15.567) at MIT, October 14, 2009.

rotated through different industries so that they can apply those quantitative methods in a variety of contexts.

There is some evidence that firms practicing data-driven decision-making tend to perform better (Brynjolfsson, Hitt, & Kim, 2011; Mcelheran & Brynjolfsson, 2015). This is consistent with the arguments that the increase in executive pay may be due to the increased generality of the required managerial skills (Frydman, 2005; Murphy & Zabojnik, 2004).

Following Frydman (2005), we formulate the level of CEO pay to reflect the importance of general skill as a function of not only the general human capital of the CEO but also the firm-specific human capital:

$$q_{CEOi,k} = S_k(g^* + g_i + h_i)$$

Where $q_{CEOi,k}$ is the total output by CEO,i, at firm k, s_k is the size of firm k, g^* is the general skill of CEO,i, before entering firm k, g_i is the general skill of CEO, i, acquired at firm, h_i is the firm-specific human capital of CEO,i, acquired at firm k. The firm, k, pays its CEO, i, the following wage:

$$w_{CEOi} = (1 - \rho)S_k(g^* + g_i + h_i)$$

$(1 - \rho)$ is the fraction of the value of the marginal product of labor of the CEO that the firms can have where firms bargain for a share of the output jointly created ($0<\rho<1$).

The CEO, i, leaves firm k for k' if

$$(1 - \rho)S_{k'}(g^* + g_i) > (1 - \rho)S_k(g^* + g_i + h_i)$$

The CEO's wage becomes:

$$w_{CEOi} = (1-\rho)S_{k'}(g^* + g_i) = (1-\rho)S_{k'}(g^* + \alpha a_i)$$

With the assumption that $g_i = \alpha a_i$ where α is the relative importance of general skill and $a_i$is the ability of CEO, i. The difference in pay between CEOs with high and low ability is larger if α is larger.

The variance of CEO pay becomes:

$$Var(w_{CEOi}) = Var[(1-\rho)S_{k'}(g^* + g_i)] = Var[(1-\rho)S_{k'}(g^* + \alpha a_i)]$$

$$= Var[(1-\rho)S_{k'}g^*)] + Var[(1-\rho)S_{k'}\alpha a_i)] + 2cov[(1-\rho)S_{k'}g^*, (1-\rho)S_{k'}\alpha a_i]$$

$$= Var[(1-\rho)S_{k'}g^*)] + ((1-\rho)\alpha)^2 Var[S_{k'}a_i)] + 2(1-\rho)^2\alpha\, g^* Cov(S_{k'}, S_k a_i)$$

$Cov(s_{k'}, s_{k'}a_i)$ is positive because managers with higher ability, a, self-select into larger firms (i.e., higher s) (Frydman 2005). Therefore, the variance of CEO pay increases with increasing α.

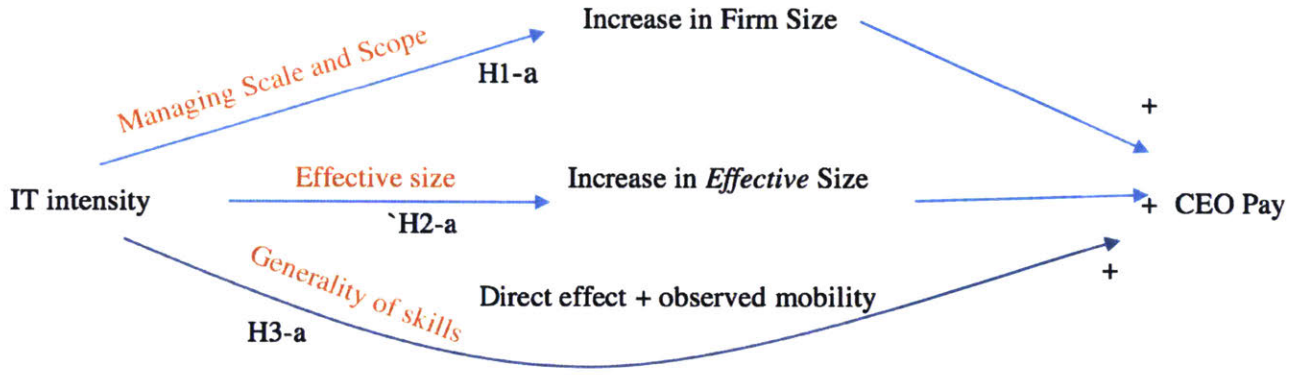
We posit that α is a function of IT: industries with a high IT intensity have a higher value of α because data-driven decision-making is a more general skill.

In other words, if the average IT intensity in the whole economy or industry wide increases the demand for the general skill of CEOs, then the average IT intensity can have a direct effect on CEO pay inequality (i.e., without going through nominal firm size or effective firm size). If IT-intensive firms can be managed by general managers (who are not necessarily experts in the specific sector of the firm), one can expect CEO pay inequality to increase, as hiring boards now have access to a larger external pool of CEO candidates (Frydman, 2005).

3.4 Summary

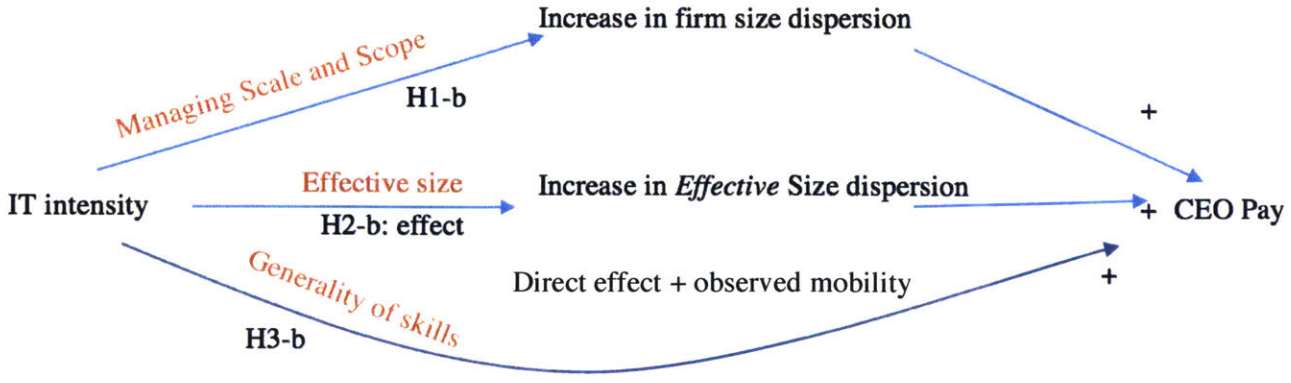
Figure 10 summarizes the three hypothesized causal channels between the average IT intensity and CEO pay levels (H1-a,H2-a, H2-a), and three predictions for the role of IT in the dispersion of CEO pay (H1-b, H2-b, and H3-b).

Figure 10. The role of IT in the change in CEO pay. (a) and (b) illustrate the hypoth-



eses of the role of IT in the

changes in CEO pay level and CEO pay dispersion, respectively.



4 4. Data Sources and Variables

4.1 IT intensity at industry level

We follow a method similar to one described in a previous study (Brynjolfsson, McAfee, Sorell, & Zhu, 2007) to estimate IT intensity at the industry level. In summary, IT intensity is defined as IT capital stock divided by the sum of Structure, Equipment and Intellectual property. The capital stock data for IT, Structures, Plant, and Intellectual Property are available from the Bureau of Economic Analysis's (BEA) "Fixed Assets Table" for 63 industry sectors at approximately three-digit NAICS level from 1947 to 2014. A precise list of asset types used to define "IT" is available in the appendix. We use two variables for IT intensity: one is the IT intensity in the whole economy each year, and the other is the IT intensity of each industry each year. Both of these variables are computed in real terms, taking into account the steep decline in the cost of storage and computing power. For example, the BEA price index for mainframes has declined by a factor of 60 between 1992 and 2014s.[6] The theories of CEO pay that we will explore all argue that it is largely determined at the industry-wide level, so we will focus on industry-level measures of IT.

4.2 Executive compensation and firm-level company data

We use two Compustat databases, *Industrial* and *Executives*, for the period from 1992 to 2014. Compustat provides commercially available databases for public companies. The *Industrial* database provides firm characteristics such as physical assets, employee

[6] Our results are, however, robust to using "constant-dollar" IT figures, i.e., simply deflated by GDP growth or by the CPI.

numbers, common stock, and sales. The *Executives* database provides data on compensation for up to 13 of the top executives from each company. It is compiled from proxy statements filed by the companies in compliance with Securities and Exchange Commission (SEC) regulations and covers S&P 1500 companies starting in 1992.

Executive compensation is taken from Execucomp variable *tdc1*. *tdc1* and includes salary, bonus, other annual, restricted stock grants, LITP payouts, all other, and present value of option grants. We select companies with at least three executives included in the database.

As our firm size variable, we use firm market value. It is computed using the same equation as in Gabaix & Landier (2008), who report it as the best available proxy for firm size. The exact computation is detailed in the appendix of this paper.

Industry turbulence is used as a control, and it is computed as the average rank change of firms within each industry in terms of Sales or EBITDA. A turbulent industry sees more reshuffling (top-ranked firms being ranked lower in subsequent years, and conversely) compared to a non-turbulent industry in which firms' ranks remain relatively constant over time.

Capital stock values are deflated using BEA price indices for each asset type. All other nominal quantities (such as market values) are converted into year 2000 dollars using the GDP deflator from the BEA.

Excluding the observations with missing variables, we examine panel data from 3413 publicly traded firms from 61 industries over 23 years.

5 5. Results

5.1 IT and firm size

This section shows basic correlations between industry-level IT intensity and firm size. The evidence is consistent with IT leading to a winner-take-all effect in firm size: higher IT intensity leads to higher total and mean market value of the industry, but to a lower median market values. When IT intensity increases, the sector as a whole expands, but mainly through an increase in size of top firms and at an expense in firm size of midsized and smaller firms. This is confirmed by column (4) of the table below, showing the IT intensity if positively correlated with Gini in firm size of the industry.

	Dependent variable			
	Total industry market value	Industry mean market value	Industry median market value	Industry Gini of market values
	(1)	(2)	(3)	(4)
IT intensity-in-dustry	2.639***	0.747**	-0.517**	0.346***
	(0.753)	(0.292)	(0.236)	(0.073)
Observations	1,332	1,332	1,332	1,332
R^2	0.080	0.020	0.013	0.114
Adjusted R^2	0.079	0.019	0.012	0.113
Residual Std. Error	1.826 (df = 1330)	1.066 (df = 1330)	0.926 (df = 1330)	0.197 (df = 1330)

Note: *p<0.1; **p<0.05; ***p<0.01
Errors clustered by industry

Table 3 IT intensity and firm size

The results in the above table are robust to introducing year dummies and time dummies, or both (except in column 2, where the results no longer hold when both types of dummies are introduced)

5.2 IT and Level of CEO pay

CEO pay increased dramatically in the 1990s, but Execucomp data suggests that top CEO pay is relatively flat after 2002 (**Figure 11**).

In the following charts, we compare trends in total CEO pay (the sum of the pay of all CEOs in Execucomp) with trends in firm size, defined as the sum of market values of their firms (**Figure 4**) and market value with trends in IT intensity, as computed from BEA data as detailed above.

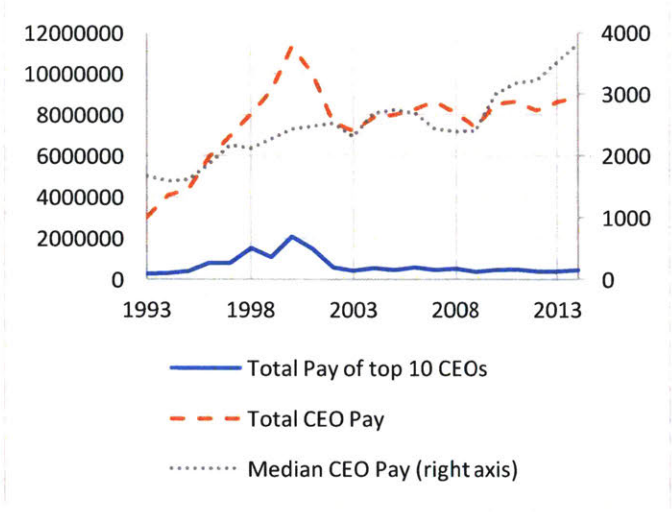


Figure 11: CEO Pay over time

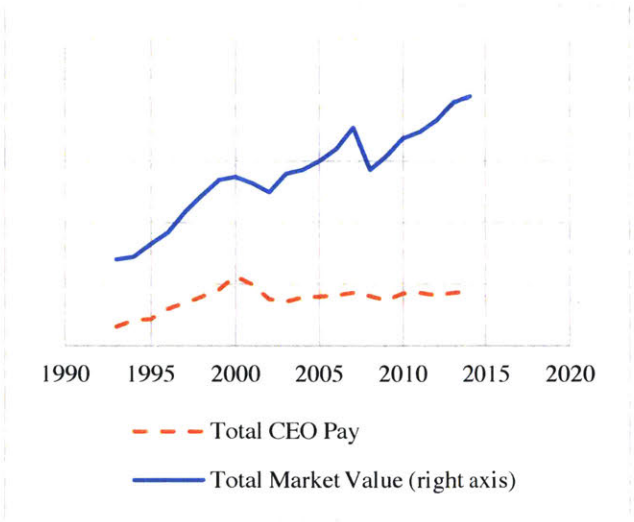


Figure 12 Total CEO Pay vs. Total Market Value

Total CEO pay as well as top CEO pay seems to be leveling off after 2002, whereas median CEO pay has been steadily increasing in real terms (**Figure 11**), as have IT intensity and total market values of firms in the sample (**Figure 11**).

As previous studies have reported (e.g., Gabaix & Landier, 2008), total CEO pay is partly explained by the total market value of firms. We find that the increase in total CEO pay is also partly explained by the growth rate of real IT intensity, even after controlling for market value.

We explore this CEO pay relationship via a variety of regressions using industry data on IT intensity. The effects of market value are significant (Table 4), consistent with previous research (Gabaix & Landier, 2008). In addition to the firm size variables, IT intensity both in the whole economy and at industry level are consistently positive in each of the specifications. Note that our IT intensity measure is at the industry-wide or economy-wide level, while the CEO pay is at firm level.

	log(CEO Pay)			
	(1)	(2)	(3)	(4)
log(Market Value)	0.362***	0.362***	0.367***	0.369***
	(0.027)	(0.027)	(0.026)	(0.026)
log(Market Value of 250th firm)	0.666***	0.363***	0.357***	
	(0.095)	(0.078)	(0.077)	
IT Intensity (Whole Economy)		3.244**	2.643*	
		(1.325)	(1.454)	
IT Intensity (Industry)			0.255**	0.252**
			(0.110)	(0.110)
Year Dummies	No	No	No	Yes
Observations	36,531	36,531	36,531	36,531
R^2	0.410	0.411	0.416	0.422
Adjusted R^2	0.410	0.411	0.416	0.422
Residual Std. Error	0.749 (df = 36528)	0.747 (df = 36527)	0.745 (df = 36526)	0.741 (df = 36506)
Note:	*p<0.1; **p<0.05; ***p<0.01			
	Errors Clustered By Industry			

Table 4 Firm-level regression: CEO Pay vs. IT Intensity

If industry turbulence and median worker wage are included as additional control variables, the IT intensity remains significant (**Table 5**). The effect of industry-level IT intensity is positive and significant except in the industry fixed effect specification, suggesting that most of the relevant variation is between industries, rather than over time. All the results described in this section are robust to considering the average of all C-level executives in each firm, rather than just the CEO. This suggests that the same mechanisms might apply to CEOs and other top executives, and that the positive effects of IT on CEO pay is not the result of a transfer from other executives to CEOs, but rather from a productivity effect affecting all executives.

	Dependent variable:			
	log(CEO Pay)			
	(1)	(2)	(3)	(4)
log(Market Value)	0.368***	0.371***	0.371***	0.417***
	(0.022)	(0.027)	(0.023)	(0.009)
log(Market Value of 250th firm)	0.375***	0.330***	0.342***	0.303***
	(0.094)	(0.078)	(0.122)	(0.066)
IT Intensity (Whole Economy)	2.253	2.931**	2.671	4.052***
	(2.050)	(1.490)	(2.585)	(1.234)
IT Intensity (Industry)	0.274**	0.297**	0.308*	-0.310
	(0.133)	(0.136)	(0.165)	(0.190)
Industry Turbulence	-0.002		-0.001	
	(0.003)		(0.005)	
log(Median Wage in Industry)		-0.158	-0.151	
		(0.112)	(0.119)	
Industry Fixed Effects	No	No	No	Yes
Observations	36,478	35,711	35,658	36,531
R^2	0.417	0.415	0.415	0.488
Adjusted R^2	0.416	0.415	0.415	0.487
Residual Std. Error	0.744 (df = 36472)	0.742 (df = 35705)	0.742 (df = 35651)	0.698 (df = 36466)
Note:	*p<0.1; **p<0.05; ***p<0.01			
	Errors Clustered By Industry			

Table 5. CEO Pay vs. IT intensity with more controls

5.3 IT and Dispersion in CEO pay

On the other hand, average IT intensity in the whole economy or industry wide has a different effect on the dispersion of CEO pay.

Does the average IT increase the dispersion of firm size and consequently increase the dispersion of CEO pay? (H1-b)

	Dependent variable:			
	Gini in Market Value			
	(1)	(2)	(3)	(4)
IT Intensity (Industry)	0.163***	0.192**	0.145***	0.127**
	(0.045)	(0.076)	(0.047)	(0.053)
log(Industry Turbulence)	0.025***	0.010	0.029***	0.014**
	(0.009)	(0.009)	(0.010)	(0.006)
log(Log Industry Average Market Value)	0.060***	0.041***	0.063***	0.055***
	(0.010)	(0.016)	(0.011)	(0.019)
Industry Fixed Effects	No	Yes	No	Yes
Year Dummies	No	No	Yes	Yes
Observations	1,047	1,047	1,047	1,047
R^2	0.426	0.808	0.475	0.859
Adjusted R^2	0.425	0.797	0.462	0.848
Residual Std. Error	0.103 (df = 1043)	0.061 (df = 990)	0.099 (df = 1021)	0.053 (df = 968)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6. IT Intensity and Dispersion of Firm Size

The regressions shown in **Table 6** provide evidence of a correlation between IT intensity at the industry level and the dispersion of market values within that industry (as measured by the variance of log of market value within industry). This is evidence for our H1-B hypothesis.

Does the average IT intensity increase the dispersion in CEO pay beyond the effect of market value? (H2-b and H3-b)

Does IT explain dispersion in CEO pay above and beyond the effect through dispersion in market value? According to the generality of skills argument (H3-b), it should, while the other hypotheses do not predict such an effect.

In order to gain insight into this, in each industry, we regress the Gini coefficient of CEO pay on the level and the dispersion of CEO Pay (Table 7 and Table 8). At the industry level, higher IT intensity leads to higher dispersion in CEO pay (as measured by the variance of log of CEO pay), even after controlling for dispersion in firm sizes (measured as market value) [see appendix].

Dependent variable: within-industry Gini in CEO Pay			
	(1)	(3)	(4)
Log Industry Average Market Value	0.020**	0.020**	
	(0.009)	(0.008)	
IT Intensity (Whole Economy)		-1.076***	
		(0.198)	
IT Intensity (Industry)		0.155***	0.161***
		(0.019)	(0.022)
Industry Fixed Effects	No	No	No
Year Dummies	No	No	Yes
Observations	1,020	1,020	1,020
R^2	0.031	0.160	0.256
Adjusted R^2	0.030	0.158	0.239
Residual Std. Error	0.108 (df = 1018)	0.101 (df = 1016)	0.096 (df = 996)

Note:

Table 7. IT intensity and dispersion in CEO pay

Dependent variable: within-industry Gini in CEO Pay			
Gini in Market Value	0.341***	0.307***	
	(0.041)	(0.038)	
IT Intensity (Whole Economy)		-1.133***	
		(0.177)	
IT Intensity (Industry)		0.094***	0.161***
		(0.014)	(0.022)
Industry Fixed Effects	No	No	No
Year Dummies	No	No	Yes
Observations	1,020	1,020	1,020
R^2	0.176	0.252	0.256
Adjusted R^2	0.175	0.250	0.239
Residual Std. Error	0.100 (df = 1018)	0.095 (df = 1016)	0.096 (df = 996)

Note:

Table 8. Dispersion of CEO pay vs. IT intensity with more controls

5.4 IT and mobility of executives

Another way to distinguish among the hypotheses is by looking at turnover among CEOs and other executives. The more important the general skill of CEOs becomes, the more executives' turnover there would be (Frydman, 2005; Murphy & Zabojnik, 2004). Researchers have examined the correlation between CEO turnover and firm performance (Kaplan & Minton, 2012) and industry performance (Eisfeldt & Kuhnen, 2013). If IT increases the importance of general versus firm-specific skill, a prediction from Frydman's model is increased executive mobility. In fact, we do find a significant correlation between IT intensity and executives' turnover in this industry.

Let S_t^i be the set of executives employed by firm i at time t. "Inflow" is defined as the number of executives who are part of the firm in year t but were not part of year in year t-1. This value is normalized by the number of executives in the firm in year t. Similarly, "outflow" is the number of executives present in year t-1 but not in year t, normalized by the number of executives at time t. Finally, turnover is the average of these two values. Turnover is computed at the firm level.

$$\text{Inflow}_t^i = \frac{\#(S_t^i \setminus S_{t-1}^i)}{\#S_t^i}$$

$$\text{Outflow}_t^i = \frac{\#(S_{t-1}^i \setminus S_t^i)}{\#S_t^i}$$

$$Turnover_t^i = \frac{\text{Inflow}_i^t + \text{Outflow}_i^t}{2}$$

In our empirical analysis, we focus on executive turnover instead of CEO turnover because Execucomp has few instances of documented CEO mobility (only about 300 CEOs changed jobs in our sample). We averaged the IT intensity and the executives turnovers over 22 years from 1993 to 2014 and examined the correlation (

Figure 13 and **Table 9**). The industries with higher IT intensity have higher degree of executive turnover (

Figure 13

Figure 13). The IT intensity is significantly correlated with executive turnover (**Table 9**).

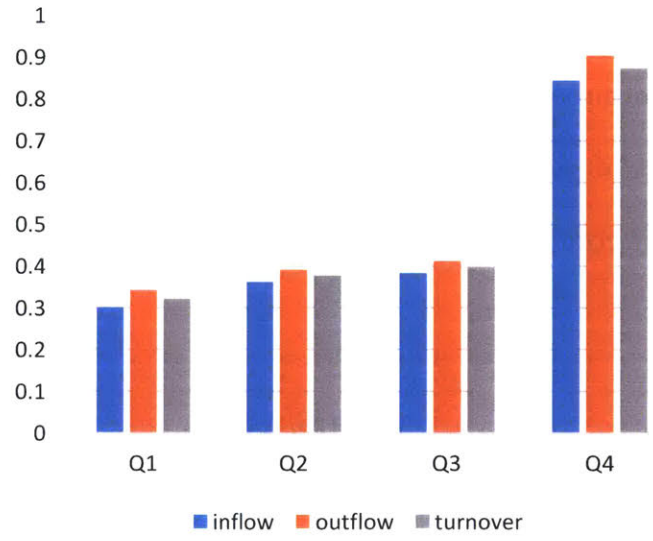


Figure 13. Executive turnover by quartile of IT intensity over the whole period studied

	inflow	outflow	Turnover
IT Intensity	0.077***	0.071***	0.074***
	(0.022)	(0.026)	(0.023)
Constant	0.117***	0.129***	0.123***
	(0.003)	(0.003)	(0.003)
Observations	61	61	61
R^2	0.168	0.115	0.148
Adjusted R^2	0.154	0.100	0.134
Residual Std. Error (df = 59)	0.019	0.022	0.020
Note:	*p<0.1; **p<0.05; ***p<0.01 Both dependent variables and IT intensity variable are the average over the period from 1993 to 2014.		

Table 9. IT intensity and Executive turnover (industry level)

Executive turnover has increased over the period from 1993 to 2014, consistent with other reports (Frydman, 2005; Kaplan & Minton, 2012). We find that the industries with high IT intensity have higher executive turnover over the sample period (Figure 14).

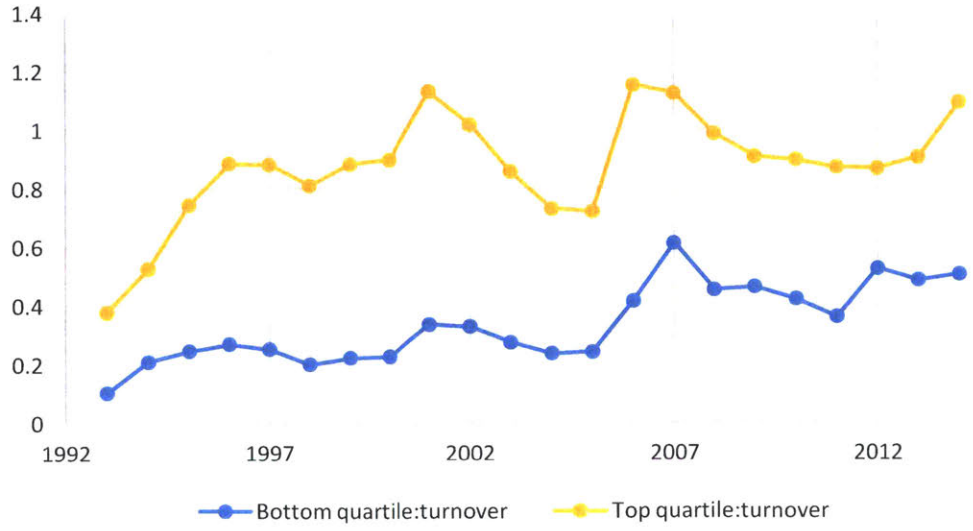


Figure 14. Executive turnover in top and bottom quartile industries in terms of IT intensity from 1993 to 2014. Each quartile of industries based on IT intensity is calculated every year, and an industry may not stay in the same quartile group over the sample period.

Where do executives go?

Note that due to the limitations of Compustat data the above measure of turnover may also include within-firm turbulence, as executives may no longer show up if they get demoted and are no longer in the top 5 salaries collected by Compustat. In order to gain insight into executive mobility patterns, we follow each executive through his/her career and decompose his/her employer changes as follows:

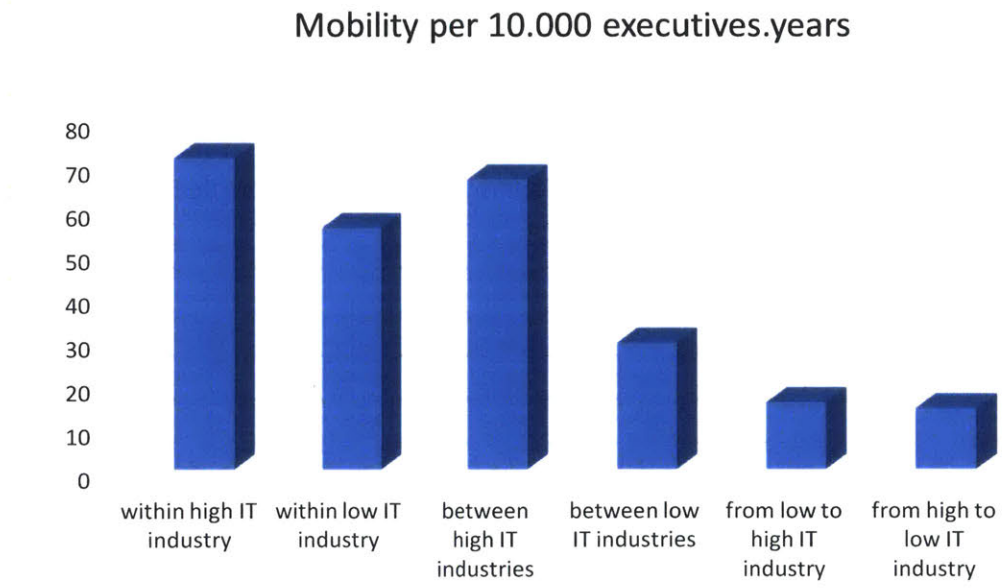


Figure 15: Executive mobility by industry IT intensity, all years pooled.

Movements per 10,000 executive-years.

Figure 15 is constructed by following each executive's career. With each new year that executives are present in the data, one of the following events can happen:

- The executive stays in the same company
- The executive changes companies but stays in the same industry (columns 1 and 2)
- The executive changes companies and industries, but the arrival industry still has above (below) median IT intensity (columns 3 and 4)
- The executive changes companies and industries, and moves from a high IT industry to a low IT industry (columns 5 and 6)

In **Figure 15**, "high IT" and "low IT" simply denote industries that are above or below the median in IT intensity. The classification here is static: IT intensities are computed over the whole time period as the average IT capital present in an industry over the whole 22 years divided by the average total assets over the whole 22 years.[7] Because the median is computed over industries, there is not an equal number of executives on either side of it, which is why columns in **Figure 15** are further scaled. For example, in order to obtain column 1, the raw count of mobility events within high IT industries is normalized by the number of total potential mobility events in high-IT industries, summed over all years (i.e., the sum of movements and non-movements), and then multiplied by 10,000. The columns can then be read as follows: out of 10,000 executives-years in high IT industries, 71 changed companies but stayed in the same industry and 66 changed

[7] Re-classifying industries every year does not affect the results, however.

industries but stayed in a high IT one. By contrast, out of 10,000 executives-years in low IT industries, 55 changed companies within the same industry, and only 29 changed industries (within low IT industries).

The general pattern is more mobility in high IT industries than in low IT industries, both when considering executives staying in the same industry or switching between high (low) IT industries. One can perform a simple t-test to assess the significance in mobility differences between high and low IT industries. High IT industries see 5296 mobility events out of 131,611 executive-years, whereas Low IT industries see 2588 mobility events out of 78,210 executive-years. A t-test rejects the null that the probability of mobility is the same in both groups, with a p-value under 0.01%. Figure 16 below decomposes **Figure 15** by year, and shows that the pattern holds over time.

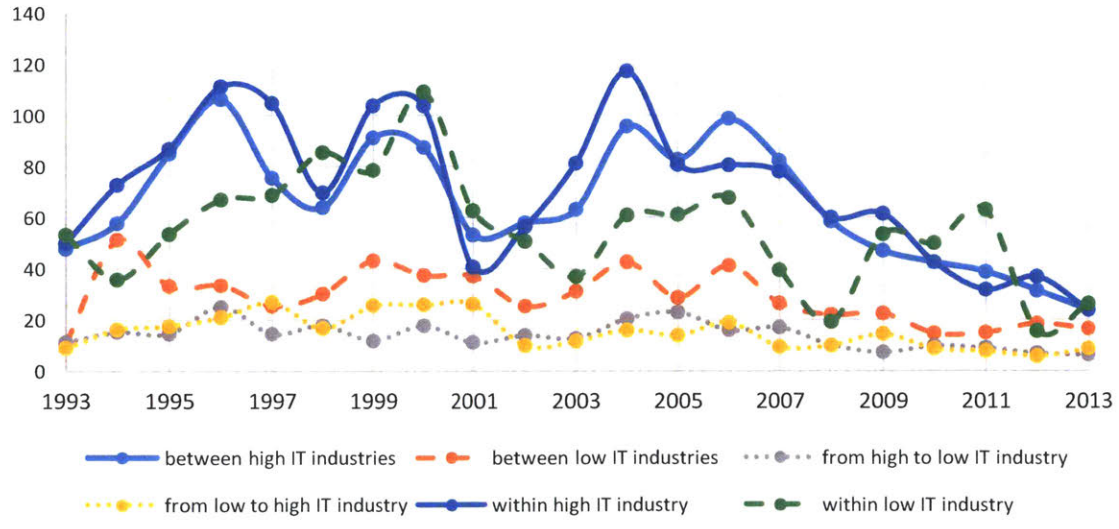


Figure 16 Executive mobility by industry IT intensity by year. Numbers represent movements per 10,000 executives.

Our argument on generality and specificity of skill is consistent with the results of **Figure 15** and Figure 16: if at least a portion of executive skill is industry-specific, executives should generally be better matched to other companies within the same industry, which is why one might expect that within-industry mobility is generally higher than between-industry mobility (comparing columns 1 and 2 to columns 3 and 4). Furthermore, if IT increases the importance of general skills (vs. specific skills), then executives in high IT industries should be able to move across firms more easily than executives in low IT industries, in which firm specific skills matter more. This prediction seems verified by comparing column 1 to column 2 and column 3 to column 4. Finally, high IT firms might be better able to recruit executives from low IT industries than the other way around, because moving to a low IT company requires more firm-specific skill than moving to a high IT company. We find some evidence of this by comparing columns 5 and 6.

Mobility: regression analysis

We then regress executive turnover on IT intensity. IT intensity is still significant if industry and year fixed effects are included (**Table 10**) and other industry characteristics are controlled for (**Table 11**).

	Turnover (industry level)				
	Pooling	pooling (clustered)	Fixed Effects (industry)	Year Dummies	Fixed Effects (industry) +year dummies
	(1)	(2)	(3)	(4)	(5)
IT Intensity	0.091***	0.091***	0.235***	0.073***	0.060*
	(0.010)	(0.015)	(0.032)	(0.013)	(0.029)
Observations	1,286	1,286	1,286	1,286	1,286
R^2	0.024	0.024	0.121	0.150	0.243
Adjusted R	0.024	0.024	0.077	0.135	0.191
Residual Std. Error	0.063 (df = 1284)	0.063 (df = 1284)	0.062 (df = 1224)	0.060 (df = 1263)	0.058 (df = 1203)
Note:					*p<0.1;**p<0.05;***p<0.01

Column (1) uses heteroscedasticity robust SEs. All others are clustered by industry

Table 10. Executive Turnover vs. IT Intensity

	Turnover			
	Pooling (clustered)	Fixed Effects (industry)	Year Dummies	Fixed Effects (industry) +year dummies
	(1)	(2)	(3)	(4)
IT Intensity	0.090***	0.204***	0.067***	0.085**
	(0.016)	(0.026)	(0.015)	(0.036)
log(Industry Turbulence)	-0.003	-0.003	0.0002	0.0001
	(0.002)	(0.004)	(0.002)	(0.006)
log(Median Salary in Industry)	-0.002	0.011	-0.0002	0.010
	(0.004)	(0.009)	(0.005)	(0.012)
log(Mean Market Value in Industry)	-0.0004	0.011	-0.003	0.003
	(0.003)	(0.009)	(0.003)	(0.014)
log(Total Employment in Industry)	0.003	-0.002	0.002	-0.008
	(0.003)	(0.010)	(0.003)	(0.008)
Observations	1,207	1,207	1,207	1,207
R^2	0.044	0.153	0.223	0.320
Adjusted R^2	0.040	0.107	0.205	0.269
Residual Std. Error	0.051 (df = 1201)	0.049 (df = 1143)	0.046 (df = 1180)	0.045 (df = 1122)
Note:	*p<0.1; **p<0.05; ***p<0.01			

Errors clustered by industry.

Table 11. Executive turnover vs. IT intensity with more controls.

We also provide a similar analysis at the firm level. (Executive turnover)$_{ijt}$, (i: industry, j:firm, t: year) is regressed on (IT intensity)$_{it}$ at the industry level and other variables at the firm level (Table 12). We find that the market value of the firm is negatively correlated with executives' turnover (Table 12), corroborating what other researchers reported that poor firm performance increases the executives' turnover (Kaplan & Minton, 2012). However, IT intensity at the industry level still explains a part of increased executives' turnover (Table 12).

	Turnover (firm level)				
	Pooling	Pooling (clustered)	Fixed Effects (firm)	Year Dummies	Fixed Effects (firm) +year dummies
	(1)	(2)	(3)	(4)	(5)
IT Intensity	**0.066*****	**0.066*****	**0.282*****	**0.055*****	**0.041**
	(0.005)	(0.008)	(0.031)	(0.008)	(0.036)
log(Market Value)	-0.006***	-0.006***	-0.007***	-0.006***	-0.015***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
log(Average Pay of top 5 executives)	0.020***	0.020***	0.014***	0.018***	0.008***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
Observations	36,067	36,067	36,067	36,067	36,067
R^2	0.017	0.017	0.201	0.042	0.231
Adjusted R^2	0.017	0.017	0.119	0.042	0.151
Residual Std. Error	0.130 (df = 36063)	0.130 (df = 36063)	0.123 (df = 32697)	0.129 (df = 36042)	0.121 (df = 32676)
Note:	*p<0.1; **p<0.05; ***p<0.01				

Column (1) uses heteroscedasticity robust SEs. All others are clustered by firm.

Table 12. Executive turnover at the firm level vs. IT intensity

5.5 Robustness checks

We conduct a number of robustness checks to examine whether our results are driven by a small subset of industries, particularities of CEOs versus other C-level executives, by the IT price deflator provided by the BEA, or by simultaneity or reverse causality problems. All of the tables and charts presented in this paper analyzing the combinations of the conditions described below are available from the authors.

Excluding IT producing industries. One might be worried that the results may be influenced by a small subset of highly IT intensive industries, such as the IT producing ones. We therefore also run the analysis excluding the four main IT producing industries[8]. The results remain broadly the same: the coefficient associated with industry level IT loses significance for regressions run on CEOs only, but remains significant if all C-level executives are included. The effects documented in dispersion tables and mobility tables remain significant.

Including all C-level executives. Including all C-level executives (between 3 and 5 executives per firm in our dataset), and regressing IT on the firm-average of their pay does not affect our results.

---

[8] These have the following BEA industry codes: 5140 - Information and data processing services; 5415 - Computer systems design and related services; 3340 - Computer and electronic products, 5110 - Publishing industries (including software).

Using constant-dollar IT instead of the "Moore's Law" deflator. Instead of using the strong deflator provided by the BEA that seeks to adjust for the tremendous decline in prices of computing power over the years we studied, it is also possible to simply use nominal values or constant-dollar values. Neither of these options change our results.

Causality concerns: using lagged IT values. IT data as it is currently available from the BEA does not lend itself well to instrumental variable analysis, which is why we cannot formally exclude all possible sources of reverse causation. Accordingly, our analysis focuses on motivated empirical correlations. Analyzing the impact of lagged values of IT intensity on pay and mobility can partially alleviate concerns about reverse causation. Here as well, our results do not seem affected by the use of IT intensity values lagged by a year.

6 Conclusion

The compensation of a CEO will depend in part on the size of the firm, the CEO's information gathering and control capabilities, and the alternative options the CEO has for employment. Information technology can potentially influence CEO pay through all three of these mechanisms.

We find some evidence for all three stories, but the strongest evidence is for the second and the third stories. IT is correlated not only with higher CEO pay, but also with increased dispersion in firm size, which is in turn reflected in increased dispersion in CEO pay. Furthermore, even controlling for changes in firm size, IT remains correlated with increased dispersion in CEO pay.

We also find a strong correlation between IT intensity and CEO mobility. The higher IT intensity an industry has, the more executive turnover it has. This is further evidence that IT increases the generality of skill for top managers and top managers become more mobile. This is consistent with the theory that CEOs in IT intensive industries have more general skills, leading to higher relative pay for top CEOs.

REFERENCES

Attewell, P. and J. Rule. 1984. "Computing and Organizations: What We Know and What We Don't Know." *Communications of the ACM* 27(12):1184–92.

Autor, David H., Lawrence F. Katz, and Alan B. Krueger. 1998. "Computing Inequality: Have Computers Changed the Labor Market?" *Quarterly Journal of Economics* 113(4):1169–1213. Retrieved (http://www.mitpressjournals.org/doi/abs/10.1162/003355398555874).

Baker, G. P. and B. J. Hall. 2004. "CEO Incentives and Firm Size." *Journal of Labor Economics* 22(4):767–98.

Balconi, M. 2002. "Tacitness, Codification of Technological Knowledge and the Organisation of Industry* 1." *Research Policy* 31(3):357–79.

Barro, Jason R. and Robert J. Barro. 1990. "Pay, Performance, and Turnover of Bank CEOs." *Journal of Labor Economics* 8(4):448–81.

Bebchuk, L. A., J. M. Fried, and D. I. Walker. 2002. "Managerial Power and Rent Extraction in the Design of Executive Compensation." *University of Chicago Law Review* 69:751–846.

Bebchuk, Lucian and Jesse Fried. 2005. "Pay Without Performance: Overview of the Issues." *Journal of Applied Corporate Finance* 17(4):8–23.

Bebchuk, Lucian and Yaniv Grinstein. 2005. "The Growth of Executive Pay." *Oxford Review of Economic Policy. Winter* 21(2):283–303.

Bertrand, M. and S. Mullainathan. 2001. "Are CEOS Rewarded for Luck? The Ones without Principals Are*." *Quarterly Journal of Economics* 116(3):901–32.

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt. 2002. "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence." *The Quarterly Journal of Economics* 117(1):339–76. Retrieved (http://www.jstor.org/stable/2696490).

Brynjolfsson, E., T. W. Malone, V. Gurbaxani, and A. Kambil. 1994. "Does Information Technology Lead to Smaller Firms?" *Management Science* 1628–44.

Brynjolfsson, E., A. McAfee, and M. Spence. 2014. "New World Order." *Foreign Affairs*.

Brynjolfsson, E. and H. Mendelson. 1993. "Information Systems and the Organization of Modern Enterprise." *Journal of Organizational Computing* 3(3):245–55.

Brynjolfsson, Erik, Lorin M. Hitt, and Heekyung Hellen Kim. 2011. "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" in *International Conference on Information System*.

Brynjolfsson, Erik, Andrew McAfee, Michael Sorell, and Feng Zhu. 2007. *Scale without*

Mass: Business Process Replication and Industry Dynamics. Cambridge: MIT and HBS.

Coase, R. H. 1937. "The Nature of the Firm." *Economica* 386–405.

Dow, J. and C. C. Raposo. 2005. "CEO Compensation, Change, and Corporate Strategy." *Journal of Finance* 2701–27.

Eisfeldt, AL and CM Kuhnen. 2013. "CEO Turnover in a Competitive Assignment Framework." *Journal of Financial Economics*.

Frydman, C. 2005. *Rising through the Ranks: The Evolution of the Market for Corporate Executives, 1936-2003*. Harvard University.

Frydman, Carola and Dirk Jenter. 2010. "CEO Compensation." *Annual Review of Financial Economics* 2(1):75–102.

Frydman, Carola and Raven Saks. 2007. *Historical Trends in Executive Compensation, 1936-2005*. Cambridge: MIT Sloan School of ManagementFederal Reserve Board of Governors.

Frydman, Carola and Raven E. Saks. 2010. "Executive Compensation: A New View from a Long-Term Perspective, 1936â€"2005." *Review of Financial Studies* hhp120.

Gabaix, Xavier and Augustin Landier. 2008. "Why Has CEO Pay Increased so Much?" *The Quarterly Journal of Economics* 123(1):49–100. Retrieved (http://www.jstor.org/stable/25098894).

Garicano, L. 2000. "Hierarchies and the Organization of Knowledge in Production." *Journal of Political Economy* 108(5):874–904.

Garicano, L. and E. Rossi-Hansberg. 2004. "Inequality and the Organization of Knowledge." *American Economic Review* 94(2):197–202.

Garicano, L. and E. Rossi-Hansberg. 2006. "Organization and Inequality in a Knowledge Economy*." *The Quarterly Journal of Economics* 121(4):1383–1435.

Grossman, Sanford J. and Oliver D. Hart. 1986. "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration." *The Journal of Political Economy* 691–719.

Guadalupe, M., J. Wulf, and M. A. Boston. 2008. "The Flattening Firm and Product Market Competition: The Effect of Trade Liberalization." *NBER Working Paper*.

Gurbaxani, V. and S. Whang. 1991. "The Impact of Information Systems on Organizations and Markets." *Communications of the ACM* 34(1):59–73.

Hall, B. J. and J. B. Liebman. 1998. "Are CEOS Really Paid Like Bureaucrats?*." *Quarterly Journal of Economics* 113(3):653–91.

Hall, B. J. and K. J. Murphy. 2003. "The Trouble with Stock Options." *Journal of Economic Perspectives* 49–70.

Hermalin, Benjamin E. 2005. "Trends in Corporate Governance." *Journal of Finance* 60(5):2351–84.

Himmelberg, Charles P., R. Glenn Hubbard, and Darius Palia. 1999. "Understanding the Determinants of Managerial Ownership and the Link between Ownership and Performance." *Journal of Financial Economics* 53(3):353–84.

Holmstrom, B. and S. N. Kaplan. 2001. "Corporate Governance and Merger Activity in the United States: Making Sense of the 1980s and 1990s (Digest Summary)." *Journal of Economic Perspectives* 15(2):121–44.

Hubbard, R. Glenn and Darius Palia. 1995. "Executive Pay and Performance Evidence from the US Banking Industry." *Journal of financial economics* 39(1):105–30.

Jensen, M. C. and K. J. Murphy. 1990. "Performance Pay and Top-Management Incentives." *Journal of Political Economy* 98(2):225.

Kaplan, SN and BA Minton. 2012. "How Has CEO Turnover Changed?" *International review of Finance.*

Kostiuk, Peter F. 1990. "Firm Size and Executive Compensation." *Journal of Human Resources* 25(1):91–105.

Leavitt, H. and T. Whisler. 1958. "Management in the 1980's." *Harvard Business Review* November-D:41–48.

Lucas Jr, Robert E. 1978. "On the Size Distribution of Business Firms." *The Bell Journal of Economics* 508–23.

Mcelheran, Kristina and Erik Brynjolfsson. 2015. "Data in Action: Data-Driven Decision Making in US Manufacturing."

Mishel, Lawrence and Alyssa Davis. 2014. "CEO Pay Continues to Rise as Typical Workers Are Paid Less." *Economic Policy Institute: Issue Brief #380.*

Murphy, K. J. and J. Zabojnik. 2004. "CEO Pay and Appointments: A Market-Based Explanation for Recent Trends." *American Economic Review* 94(2):192–96.

Rajan, R. G. and J. Wulf. 2006. "The Flattening Firm: Evidence from Panel Data on the Changing Nature of Corporate Hierarchies." *The Review of Economics and Statistics* 88(4):759–73.

Roberts, David R. 1956. "A General Theory of Executive Compensation Based on Statistically Tested Propositions." *Quarterly Journal of Economics* 70(2):270–94.

Rosen, Sherwin. 1981. "The Economics of Superstars." *The American Economic Review* 71(5):845–58. Retrieved (http://www.jstor.org/stable/1803469).

Rosen, Sherwin. 1992. "Contracts and the Market for Executives." Pp. 181–211 in *Contract Economics*, edited by L. Werin and H. Wijkander. Cambridge, MAOxford: Blackwell.

Rosen, Sherwin. 1996. *Review of: The Winner-Take-All Society.*

Tervio, M. 2008. "The Difference That CEOs Make: An Assignment Model Approach." *American Economic Review* 98(3):642–68.

Williamson, O. E. 1973. "Markets and Hierarchies: Some Elementary Considerations."

The American Economic Review 316–25.

Williamson, O. E. 1981. "The Economics of Organization: The Transaction Cost Approach." *American Journal of Sociology* 87(3):548.

Yermack, D. 1997. "Good Timing: CEO Stock Option Awards and Company News Announcements." *Journal of Finance* 449–76.

Appendix: Correlation Matrix of main variables (industry level)

	Industry average IT intensity	Mean CEO pay in industry	Pay of industry top CEO	Total CEO compensation of industry	Pay of median CEO	Mean industry market value	Market value of largest firm	Market value of median firm	Total employment	Industry turbulence	Median wage in industry
Industry average IT intensity	100.0%	11.2%	27.2%	30.9%	-3.2%	6.3%	17.1%	-9.9%	10.1%	7.8%	19.3%
Mean CEO pay in industry	11.2%	100.0%	50.8%	35.1%	63.5%	39.3%	18.7%	34.5%	9.5%	-4.2%	14.5%
Pay of industry top CEO	27.2%	50.8%	100.0%	66.2%	9.0%	18.1%	21.2%	5.4%	18.3%	14.7%	7.9%
Total CEO compensation of industry	30.9%	35.1%	66.2%	100.0%	14.1%	37.3%	48.4%	7.2%	54.1%	23.1%	20.5%
Pay of median CEO	-3.2%	63.5%	9.0%	14.1%	100.0%	36.1%	10.3%	52.1%	5.1%	-8.0%	17.6%
Mean industry market value	6.3%	39.3%	18.1%	37.3%	36.1%	100.0%	81.7%	51.5%	15.4%	6.5%	32.6%
Market value of largest firm	17.1%	18.7%	21.2%	48.4%	10.3%	81.7%	100.0%	18.7%	28.6%	14.0%	17.3%
Market value of median firm	-9.9%	34.5%	5.4%	7.2%	52.1%	51.5%	18.7%	100.0%	-0.9%	-0.8%	26.5%
Total employment of industry (census)	10.1%	9.5%	18.3%	54.1%	5.1%	15.4%	28.6%	-0.9%	100.0%	6.1%	-5.3%
Industry turbulence	7.8%	-4.2%	14.7%	23.1%	-8.0%	6.5%	14.0%	-0.8%	6.1%	100.0%	15.2%
Median wage in industry	19.3%	14.5%	7.9%	20.5%	17.6%	32.6%	17.3%	26.5%	-5.3%	15.2%	100.0%
Executive Turnover	19.3%	6.3%	8.8%	12.3%	3.2%	2.3%	6.0%	-0.2%	9.7%	-10.0%	3.1%

Technical Appendix

Merging Compustat/Execucomp data with BEA data

The final industry classification used in this paper is made of 63 "BEA" industries, whereas Compustat data contains NAICS codes. NAICS codes are therefore converted to BEA industries using the below table:

INDUSTRY TITLE	BEA CODE	1997 NA-ICS Codes	2002 NA-ICS Codes	2007 NA-ICS Codes
Agriculture, forestry, fishing, and hunting	--------	**11**	**11**	**11**
Farms	110C	111,112	111,112	111,112
Forestry, fishing, and related activities	113F	113,114,115	113,114,115	113,114,115
Mining	--------	**21**	**21**	**21**
Oil and gas extraction	2110	211	211	211
Mining, except oil and gas	2120	212	212	212
Support activities for mining	2130	213	213	213
Utilities	**2200**	**22**	**22**	**22**
Construction	**2300**	**23**	**23**	**23**
Manufacturing	--------	**31-33**	**31-33**	**31-33**
Durable goods	--------	--------	--------	--------
Wood products	3210	321	321	321
Nonmetallic mineral products	3270	327	327	327
Primary metals	3310	331	331	331
Fabricated metal products	3320	332	332	332
Machinery	3330	333	333	333
Computer and electronic products	3340	334	334	334
Electrical equipment, appliances, and components	3350	335	335	335
Motor vehicles, bodies and trailers, and parts	336M	3361-3	3361-3	3361-3
Other transportation equipment	336O	3364-9	3364-9	3364-9

Furniture and related products	3370	337	337	337
Miscellaneous manufacturing	338A	339	339	339
Nondurable goods	--------	--------	--------	--------
Food, beverage, and tobacco products	311A	311,312	311,312	311,312
Textile mills and textile product mills	313T	313,314	313,314	313, 314
Apparel and leather and allied products	315A	315,316	315,316	315, 316
Paper products	3220	322	322	322
Printing and related support activities	3230	323	323	323
Petroleum and coal products	3240	324	324	324
Chemical products	3250	325	325	325
Plastics and rubber products	3260	326	326	326
Wholesale trade	**4200**	**42**	**42**	**42**
Retail trade	**44RT**	**44-45**	**44-45**	**44-45**
Transportation and warehousing	**--------**	**48-49**	**48-49**	**48-49**
Air transportation	4810	481	481	481
Railroad transportation	4820	482	482	482
Water transportation	4830	483	483	483
Truck transportation	4840	484	484	484
Transit and ground passenger transportation	4850	485	485	485
Pipeline transportation	4860	486	486	486
Other transportation and support activities	487S	487,488,492	487,488,492	487,488,492
Warehousing and storage	4930	493	493	493
Information	**--------**	**51**	**51**	**51**
Publishing industries (including software)	5110	511	511, 516 (pt.)	511
Motion picture and sound recording industries	5120	512	512	512
Broadcasting and telecommunications	5130	513	515, 517	515, 517
Information and data processing services	5140	514	516 (pt.), 518, 519	518, 519
Finance and insurance	**--------**	**52**	**52**	**52**

Federal Reserve banks	5210	521	521	521
Credit intermediation and related activities	5220	522	522	522
Securities, commodity contracts, and investments	5230	523	523	523
Insurance carriers and related activities	5240	524	524	524
Funds, trusts, and other financial vehicles	5250	525	525	525
Real estate and rental and leasing	--------	**53**	**53**	**53**
Real estate	5310	531	531	531
Rental and leasing services and lessors of intangible assets	5320	532,533	532,533	532,533
Professional, scientific, and technical services	--------	**54**	**54**	**54**
Legal services	5411	5411	5411	5411
Computer systems design and related services	5415	5415	5415	5415
Miscellaneous professional, scientific, and technical services	5412	541 ex. 5411,5415	541 ex. 5411,5415	541 ex. 5411,5415
Management of companies and enterprises	**5500**	**55**	**55**	**55**
Administrative and waste management services	--------			
Administrative and support services	5610	561	561	561
Waste management and remediation services	5620	562	562	562
Educational services	**6100**	**61**	**61**	**61**
Health care and social assistance	--------	**62**	**62**	**62**
Ambulatory health care services	6210	621	621	621
Hospitals	622H	622	622	622
Nursing and residential care facilities	6230	623	623	623
Social assistance	6240	624	624	624
Arts, entertainment, and recreation	--------	**71**	**71**	**71**

Performing arts, spectator sports, museums, and related activities	711A	711,712	711,712	711,712
Amusements, gambling, and recreation industries	7130	713	713	713
Accommodation and food services	--------	**72**	**72**	**72**
Accommodation	7210	721	721	721
Food services and drinking places	7220	722	722	722
Other services, except government	**8100**	**81**	**81**	**81**

Note: to make this process easier, the authors have built an R package (NAICStoBEA), which supports most variants of NAICS, available upon request.

Measures derived from Execucomp and Compustat

We restrict our attention to full-year CEOs [CEOANN='CEO'] in Execucomp. We then merge the Execucomp dataset with the Compustat Fundamentals database, downloaded in 2015. Compustat offers various levels of aggregations. We use **C**, the highest available level of aggregation. Compustat and Execucomp are straightforwardly merged using the GVKEY variable.

As our measure of CEO pay, we use TDC1 in Execucomp, which includes the current value of non-pay compensation (such as stock options).

As our measure of firm value, we use the same formula as in Gabaix & Landier, 2008:

data199 x *abs(data25)* + *data6* – *data60* – *data74*, where *data199* is the share price of closing at fiscal year, *data25* is Common Shares Outstanding, *data6* is Total Assets, *data60* is Total Common Equity, and *data74* is Deferred Taxes. Note that using 2015 Compustat variable names, this equation becomes:

csho x abs(prcc_f) + at - ceq- txdb

Industry turbulence: We use the SALE and the EBIDTA variables from Compustat. For each year and within each industry, we rank firms by their sales and EBITDA. We then compute, for each industry, the average absolute value rank change from one year to the next. This serves as our measure of industry turbulence, which is used as a control in some tables.

All nominal quantities are converted into 2000 dollars using the GDP deflator from the BEA.

In our regression tables, we restrict our attention to CEOs with pay > $200,000 (in 2000 dollars). This removes a negligible fraction of the sample.

Building Industry-Level IT measures

As of 2015, the BEA reports the following asset classes in its survey of tangible wealth. For each class, we report whether it is included in measure of IT spending. The denominator used to convert IT capital into IT intensity is the sum of the equipment and structures categories below. HW indicates the asset code is included in "hardware only" variables, and SW indicates it is included in "software only" variables. The general "IT" variable in our paper includes both hardware and software.

Asset Codes	NIPA Asset Types	Included in IT variable
EQUIPMENT	**TOTAL EQUIPMENT**	
EP1A	**Mainframes**	**Yes-HW**
EP1B	**PCs**	**Yes-HW**
EP1C	**DASDs**	**Yes-HW**
EP1D	**Printers**	**Yes-HW**
EP1E	**Terminals**	**Yes-HW**
EP1F	**Tape drives**	**Yes-HW**
EP1G	**Storage devices**	**Yes-HW**
EP1H	**System integrators**	**Yes-HW**

EP20	**Communications**	**Yes-HW**
EP34	**Nonelectro medical instruments**	
EP35	**Electro medical instruments**	
EP36	**Nonmedical instruments**	
EP31	**Photocopy and related equipment**	
EP12	**Office and accounting equipment**	
EI11	**Nuclear fuel**	
EI12	**Other fabricated metals**	
EI21	**Steam engines**	
EI22	**Internal combustion engines**	
EI30	**Metalworking machinery**	
EI40	**Special industrial machinery**	
EI50	**General industrial equipment**	
EI60	**Electric transmission and distribution**	
ET11	**Light trucks (including utility vehicles)**	
ET12	**Other trucks, buses and truck trailers**	
ET20	**Autos**	
ET30	**Aircraft**	
ET40	**Ships and boats**	
ET50	**Railroad equipment**	
EO11	**Household furniture**	
EO12	**Other furniture**	
EO30	**Other agricultural machinery**	
EO21	**Farm tractors**	

EO40	**Other construction machinery**
EO22	**Construction tractors**
EO50	**Mining and oilfield machinery**
EO60	**Service industry machinery**
EO71	**Household appliances**
EO72	**Other electrical**
EO80	**Other**
STRUCTURES	**TOTAL STRUCTURES**
SOO1	**Office**
SB31	**Hospitals**
SB32	**Special care**
SOO2	**Medical buildings**
SC03	**Multimerchandise shopping**
SC04	**Food and beverage establishments**
SC01	**Warehouses**
SOMO	**Mobile structures**
SC02	**Other commercial**
SI00	**Manufacturing**
SU30	**Electric**
SU60	**Wind and solar**
SU40	**Gas**
SU50	**Petroleum pipelines**
SU20	**Communication**
SM01	**Petroleum and natural gas**

SM02	**Mining**	
SB10	**Religious**	
SB20	**Educational and vocational**	
SB41	**Lodging**	
SB42	**Amusement and recreation**	
SB43	**Air transportation**	
SB45	**Other transportation**	
SU11	**Other railroad**	
SU12	**Track replacement**	
SB44	**Local transit structures**	
SB46	**Other land transportation**	
SN00	**Farm**	
SO01	**Water supply**	
SO02	**Sewage and waste disposal**	
SO03	**Public safety**	
SO04	**Highway and conservation and development**	
IPP	**TOTAL INTELLECTUAL PROPERTY PRODUCTS**	
ENS1	**Prepackaged software**	**Yes-SW**
ENS2	**Custom software**	**Yes-SW**
ENS3	**Own account software**	**Yes-SW**
RD11	**Pharmaceutical and medicine manufacturing**	
RD12	**Chemical manufacturing, ex. pharma and med**	
RD23	**Semiconductor and other component manufacturing**	**Yes-HW**
RD21	**Computers and peripheral equipment manufacturing**	**Yes-HW**

RD22	**Communications equipment manufacturing**	**Yes-HW**
RD24	**Navigational and other instruments manufacturing**	
RD25	**Other computer and electronic manufacturing, n.e.c.**	**Yes-HW**
RD31	**Motor vehicles and parts manufacturing**	
RD32	**Aerospace products and parts manufacturing**	
RDOM	**Other manufacturing**	
RD70	**Scientific research and development services**	
RD40	**Software publishers**	**Yes-SW**
RD50	**Financial and real estate services**	
RD60	**Computer systems design and related services**	**Yes-SW**
RD80	**All other nonmanufacturing, n.e.c.**	
RD91	**Private universities and colleges**	
RD92	**Other nonprofit institutions**	
AE10	**Theatrical movies**	
AE20	**Long-lived television programs**	
AE30	**Books**	
AE40	**Music**	
AE50	**Other entertainment originals**	

These IT, Software, and Hardware measures are built at the industry level for each year, and at the level of the whole economy for each year.

THIS PAGE INTENTIONALLY LEFT BLANK

Networks and Income: Evidence from Individually Matched Income and Mobile Phone Metadata

The role of social ties has been investigated in the context of job search, job mobility, income, and numerous other outcomes. However, measuring the relationship between income and various properties of one's social network has proven difficult because it requires data on income and social ties to be matched at the individual level. This paper offers the first large-scale investigation of this question using data that is both large scale and individually matched.

We investigate the relationship between income and characteristics of ego-networks. How are ego-networks different across income levels? Are there measurable differences in degree, reciprocity, diversity and centrality?

We use a dataset of Call Detail Records from a southeast Asian country containing over 100M individuals and income surveys sent out to over a hundred thousand individuals. This allows us to use fine location data to control for location effects, rather than rely on it to match incomes.

Introduction

The impact of social relationships in individuals' economic outcomes has been the subject of much interest. The role of social ties has been investigated in the context of job search, employee search, job mobility, income, and a number of other outcome varia-

bles. However, precisely measuring the relationship between income and various properties of one's social network has proven difficult so far, principally because it ideally requires data on income and social ties to be matched at the individual level.

Given the unavailability of such data, most research so far has relied on geographical matching: using individuals' most frequent location as a proxy for their income, assuming that their most frequent location is their home, and using local real estate prices or local government statistics on income and development to impute incomes. This may pose a number of challenges: first, the effects obtained may be confounded by location effects; second, such an approach does not allow us to focus on individual networks, relying on local averages instead.

We investigate the relationship between characteristics of ego networks (i.e. social networks centered around individual of interest). How are ego networks different between high and low-income individuals? Are there measurable differences in degree, reciprocity, diversity and centrality?

We set out to explore these questions using a novel dataset of Call Detail Records (CDR) from a southeast Asian country containing more than a hundred million individuals and income surveys sent out to over a hundred thousand individuals. To our knowledge, this is the first investigation of these questions on such a large scale relying on individually matched data rather than local averages. This allows us to use fine-grained location data to control for location fixed effects as the core of our analysis, rather than rely on them to match individuals to an approximate income.

Literature Review

Our approach is at the intersection of three main branches of the literature: literature that used CDR data to predict local development level, theory literature on social networks and income, and the developing literature on this topic.

An emergent field of research named "computational social science" seeks to use new sources of data, mainly produced by the use of recent technologies such as information technology, in order to better understand individual behavior at a large scale (Lazer et al. 2009), and a number of international organizations call for better collection of big data for public policy use (Group 2014).

Using Call Details Records data to predict development

A number of papers have shown the potential of CDR data in order to predict development at the local level by showing that predictions obtained from cell phone data can match official government statistics on development. For example, cell phone use data has been used to reconstruct unemployment statistics in 340 Spanish regions (Llorente et al. 2015). It has been used to track population densities at the local level (Deville et al. 2014), as well as a variety of measures of education, demographic variables and ownership variables (Frias-Martinez and Virseda 2012a). Once trained, such models can be used to generate poverty maps at higher resolution than the ones available from government statistics (Pokhriyal et al. 2015; Smith et al. 2013). It should be noted that development is not the only domain CDR data prediction has been applied to: researchers have sought to apply it to public health issues, such as malaria control (Enns and Amuasi

2013) or dengue fever (Wesolowski et al. 2015). It has also been applied to prediction of crime (Bogomolov et al. 2014) or loan repayment (Bjorkegren and Grissen 2015).

It should be noted that the above papers do not use outcome variables measured at the individual level. Rather, they merge CDR data with their variables of interest using location. A notable exception to this is (Blumenstock et al. 2015), which also makes use of a small number of income surveys (on the order of 700). In contrast to that study and (Sundsøy et al. 2016), which aimed to predict poverty based on an individual's own phone communication behavior, here we attempt to understand the relationship between income and the structure of the immediate network surrounding the individual.

Social Networks and income: theory

The role of social networks in determining individual income has originally been linked to job search: "weak ties" may play a role in individuals becoming aware of various opportunities, including job opportunities (Granovetter 1973), and a number of studies have shown that the majority of jobs are obtained through social contacts (Granovetter 1995; Rees 1966). More recently, theoretical models have emerged showing how the social network mechanisms underlying job search can lead to exacerbated inequality (Calvo-Armengol and Jackson 2004). Similarly, various models of network structure and knowledge diffusion articulate a relationship between network structure and the steady-state repartition of knowledge (Cowan and Jonard 2004). Finally, networked digital technologies may exacerbate inequality (Saint-Jacques and Brynjolfsson 2015).

Social Networks and income: empirical investigation

Beyond their importance on job search, social networks have also been documented as relevant to job mobility (Wegener 1991), pay (Seibert et al. 2001) and negotiation power (Brass and Burkhardt 1993).

Social network diversity within regions has been shown to be correlated with income using landline phone records (Eagle et al. 2010). A more recent investigation of the role of social network connections on job finding can be found in (Reis and Ferreira 2015).

Our main methodological difference with most of the extant literature is our ability to match income data individually rather than at the neighborhood level.

Data

We use an anonymized mobile phone dataset containing one month of standard metadata in a developing country in South Asia. Our goal is to study the relationship between local and global network characteristics and individual income. In particular we focus on a local view of the network called *ego network*. The focal node of interest is called the *ego* whereas all ego's connections are called *alters*. In addition to ego-alter edges, the ego network includes all the edges between the alters, thus enabling us to study structural factors that are not directly controlled by the ego.

Income Data

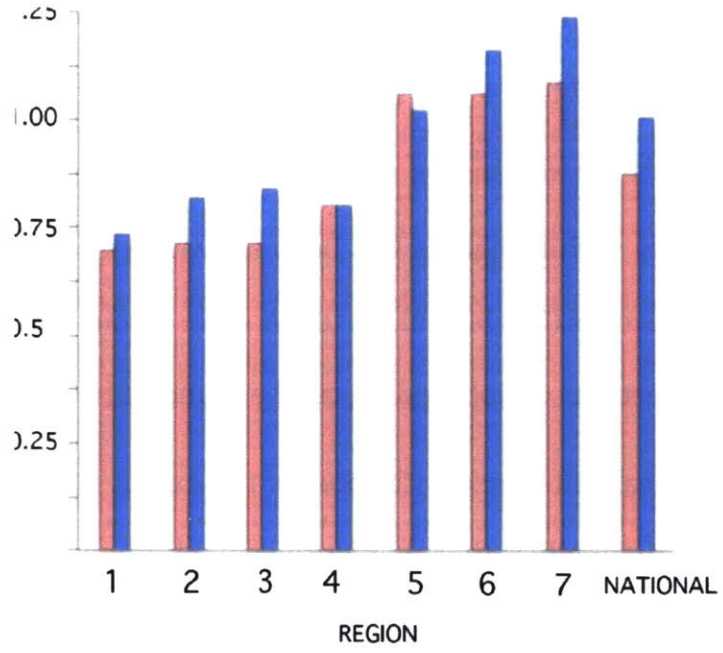


Figure 17 Average household income per administrative region

Metrics are normalized as a fraction of official national average. Blue bars represent the official statistics, while pink bars represent the combined household survey income.

The income categories for a random selection of more than 100,000 subscribers were obtained through three sequential large-scale market research household surveys. Approximately 270,000 individuals were surveyed across all mobile operators in the country, out of which 111,128 customers of our carrier. We treat this random set of subscribers as our *egos*.

Information about income was directly asked from the respondents, who were requested to place themselves within pre-defined income bins. The country was stratified in over 220 sales territories by the phone company, and for every territory, an equal number of

sub-territories were randomly selected. Survey participants were distributed across these sub-territories proportional to their population so that there were overall about 400 surveyed households in each sales territory. Systematic sampling was undertaken by selecting every fourth household, starting from a randomly selected geographic reference point and direction within each sub-territory. In the case of more than one household present in the complex or building, the fourth household was selected. In cases of non-response, the next household was selected. Non-response rate was approximately 10% of households. Respondents within the household were selected via the Kish grid method (Kish 1949) among those who were eligible. Eligibility was defined as individuals with their own phone, between 15 and 65 years of age. Figure 17 compares our projected average income per region based on the survey results and their actual values published in official statistics (Pearson correlation 0.925).

The monthly income values were coded as ordinal categories from 1-13. Table 13 summarizes the correspondence between the income categories and their actual monetary value after conversion to US dollars. Figure 18 illustrates the income distribution among our egos. To the best of our knowledge, this is the first large-scale study on the link between networks and income on an individual level with more than 100,000 data points.

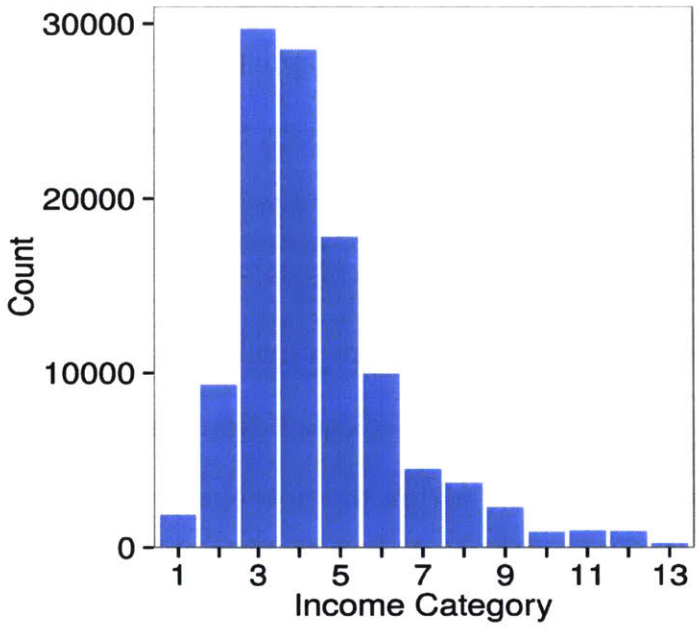


Figure 18 Income distribution of egos

Income bin	Monthly Household Income (USD)	Number of individuals in bin
1	0 - 33	1895
2	33 - 78	9351
3	78 - 130	29718
4	130 - 195	28532
5	195 - 260	17841
6	260 - 325	9995
7	325 – 390	4536
8	390 - 455	3752
9	455 - 520	2341
10	520 - 585	929
11	585 - 651	999
12	651 - 1301	966
13	1301+	274

Table 13 Survey relationship between household income categories and corresponding range in US dollars

Social Network Data (Phone records)

We used one month of raw Call Detail Records (CDR) for all carrier subscribers to construct a large-scale call graph from which we extracted individual ego networks. Raw CDR records for each user contain the following metadata:

- Interactions type (SMS or Call)
- Correspondent ID (The unique identifier of the contact)
- Direction (Incoming or Outgoing)
- Date and time of the interaction
- Duration of Interaction (Only valid for calls)
- Location of cell tower serving the subscriber (Latitude and Longitude)

In addition to the ~100,000 individual ego networks, we construct the full country-wide call graph to compute measures of global centrality for our egos. The full graph consists of 113 Million nodes and 2.7 Billion edges. Each node is tagged with auxiliary information such as phone type (basic, feature or smartphone) and home location (Location of the most frequent tower at night). Edge attributes consist of total call duration in minutes, total count of phone calls and SMS messages exchanged between the two parties in each direction. The edge attributes allow us to construct meaningful metrics weighted by the strength of the link between the ego and alters. The table below summarizes some descriptive statistics on our full network dataset.

Number of nodes	113 Million
Number of edges	2.7 Billion
Total number of calls	11 Billion
Total call duration (Hours)	3.47 Million
Total number of SMS	1.16 Billion
Number of Towers	10306
Number of egos with income information	111,128
Median number of nodes in ego networks	46
Median Number of edges in ego networks	88

Table 14: Descriptive statistics on the full dataset used to construct the country-wide call graph and individual ego networks.

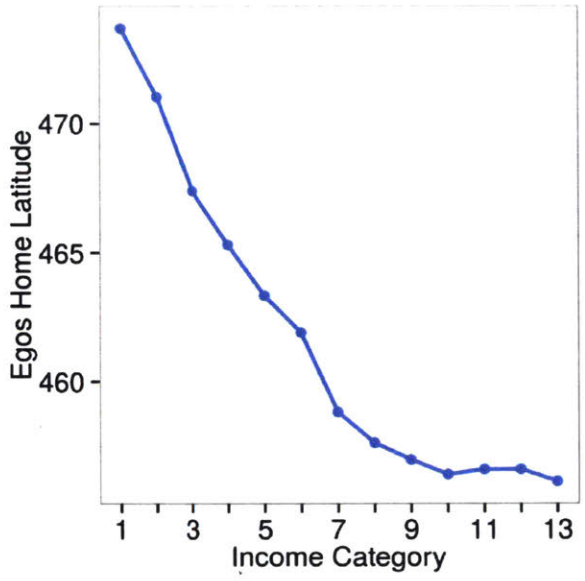


Figure 19 Median of egos home latitude grouped by income cate-

Our independent variables, extracted from node attributes and network structure, capture structural information about the *ego* but also about the *alters*. In particular, they incorporate social signals present in the ego network such as density, reciprocity, centrality and alter diversity.(Frias-Martinez and Virseda 2012b).

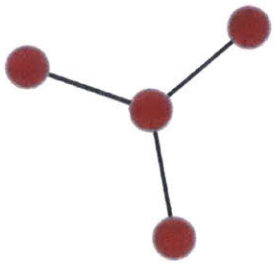
Many of our variables reveal the existence of stark differences between the rich and the poor. For example, Figure 19 above illustrates the mean latitude of respondents' home location among different income categories. It shows income segregation across the country which justifies the inclusion of home location as a fixed effect in our analysis.

Degree, density and reciprocity

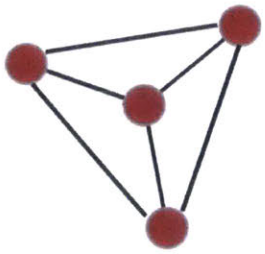
Main concepts and variables

With the dataset assembled as described above, we now seek to document the relationship between measures of degree, local density and link reciprocity and income. Our unit of analysis is always the "ego", i.e. an individual who responded to an income survey. The variables are then computed on the "ego network", i.e. the network that immediately surrounds them. One exception to this is eigenvalue centrality, which is computed on the whole network of cell phone subscribers (and the non-subscribers they are in contact with) in our country of interest. We use the following variables:

- *Out degree*, defined as the number of individuals *ego* has sent a SMS to or called
- *In degree*, the number of individuals ego has received SMSs or calls from
- *Density* measures the completeness of the local network, and is defined as the number of ties existing in the ego network divided by the number of ties that would exist if the network were fully connected. In other words, if few of ego's alters are connected with each other, density is low. If all of ego's alters are in contact with each other, density is 100%.



A low-density ego network



A high-density ego network

- *Tie Reciprocity* measures the percentage of ties that are reciprocal (i.e. ties with individuals that have both called or texted the *ego* and that have also been called or texted *by the ego.)*
- *Eigenvalue centrality* measures how central an individual to the nationwide social network. For computational tractability (over billions of ties), it is operationalized using the unweighted PageRank algorithm.
- *Local centrality* measures how central an individual is to their immediate network. This is computed as the sum of an individual's phone calls (made or received) divided by the sum of calls observed from the alters. It is the answer to the question: out of all calls happening in the local network, what percentage directly involved the *ego?*
- *Smartphone ownership (ego)* and *Smartphone ownership (most alters)* are indicator variables that denote *ego* owning a smartphone and *the majority of ego's alters* owning a smartphone. Note that in our country of interest, smartphone penetration is much lower than in the United States.

Summary statistics and income group comparisons

Degree

We now turn to a brief description of the correlation between our variables of interest and income. For clarity, we first comment on charts of comparing income group averages (i.e. not featuring any controls), and then move on to comment on the correlations within our regression framework (Table 15).

Graphically, many of these variables show a surprisingly monotonic relationship with income. For example, both incoming and outgoing degree are positively correlated with income category in our data (see Figure 20). This is the simplest way to illustrate that wealthier individuals are "better connected": they have a higher number of social connections (In this sense, one can treat degree as a measure of global centrality). Furthermore, one can see that the gap between their average of incoming connections versus outgoing ones broadens with income, consistent with the idea that wealthier individuals are being reached and solicited more than they reach others.

Reciprocity and Density

On the other hand, tie reciprocity and tie reciprocity are negatively correlated, as can be seen in Figure 21. In other words, it appears that wealthier individuals have a lower proportion of their contacts with whom they have two-way communication. This is even more striking when taking into account the fact that outgoing calls are charged, but incoming calls are free: one would expect less wealthy individuals to have lower reciprocity because they may be able to receive calls but not to return them. However, this is not the case in the data. A possible explanation may be that wealthier individuals are more frequently solicited by others because they command larger amounts of social and economic capital that others may want access to. In other words, even though wealthier individuals can afford to make calls, they are the ones disproportionately being called, and not the other way around. A good description of the relationship between income and social network may therefore be "it's who knows you", rather than the famous saying that "it's who you know".

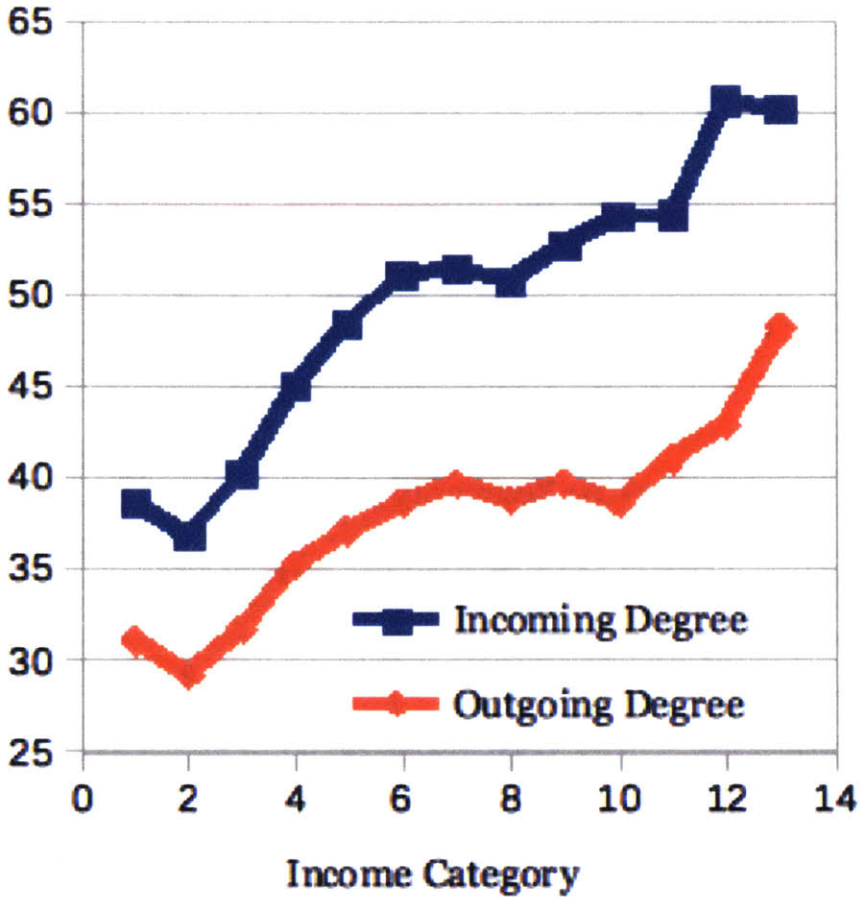


Figure 20 average incoming and outgoing degree by income category

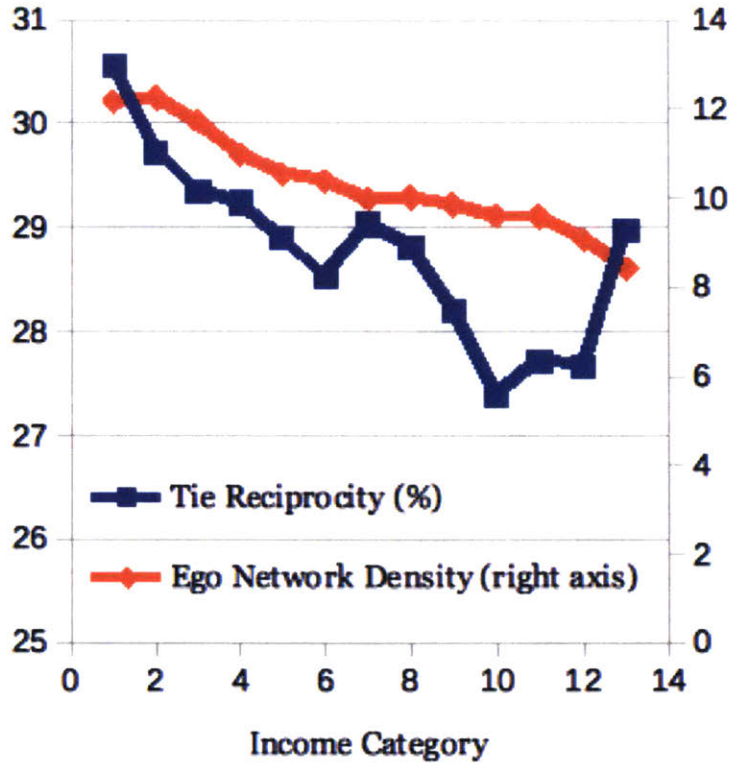


Figure 21 Tie reciprocity and Network density by income category

An equally interesting result lies in the relationship between income and ego network density (Figure 21): they are negatively correlated. In other words, the alters of a wealthier individual are less likely to know each other than the alters of a less wealthy individual. This lends credibility to the idea that wealthy individuals may occupy "special" spots in their local social networks, and may act as information hubs: if all alters know each other, then getting access to information does not require going through the *ego;* however, if this is not the case, then getting access to resources or information of another alter is more likely to require leveraging the *ego*.

Centrality

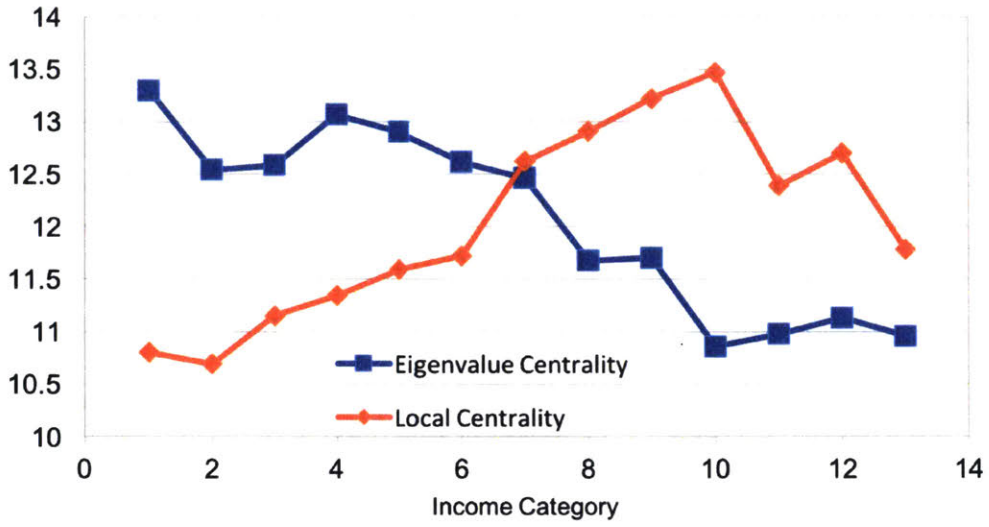


Figure 22 global (eigenvalue) and local centrality by income category

If richer individuals act as local hubs, then one might expect them to have higher structural local centrality. This is what we find in Figure 22, though the relationship may seem more inverted-U-shaped than monotonous. Still, on the major part of the income distribution, local centrality is positively correlated with income, suggesting once again that richer individuals may act as information hubs. The U-shaped behavior may be explained by the idea that very high local centrality values could indicate a lack of any novel information in the immediate network as alters stop acting as a bridge to the outside world. This would happen, for instance, if the high-income egos as a class had automatic, dependable sources of income.

Much more surprising is the seemingly negative relationship between global centrality and income: it seems that individuals that are more central to the nationwide network are, on average, poorer. What makes this fact even more puzzling is that, in a regression framework, it seems to survive the addition of geographical fixed effects. A possible explanation may be that wealthy people are few compared to poor individuals, and that they may be segregated socially (our country of interest has large levels of economic inequality). It is also possible that poor people are more involved in activities that depend on with transportation networks that link different communities together, whereas wealthy people have administrative jobs that mostly involve the local population. Another possible factor is the fact that eigenvalue centrality here has been computed in an unweighted fashion, putting equal weight on all ties, and not weighting them by the number or duration of calls (or the number of text messages exchanged). This is mainly done for computational reasons, as computing global centrality metrics on a network of over billions of edges remains, at the time of this writing, a difficult and computationally expensive process. Still, the negative relationship between centrality and income seems fairly robust in our data, and should be the subject of more investigation by future research.

Interval Regression Analysis

The above charts provide an overview of our results regarding degree, reciprocity, diversity and centrality. However, a regression framework can usefully be added, for four main purposes:

- Seeking out interpretable correlation numbers (i.e. what is the dollar value of an additional connection?). This is best done looking at Table 15, where degree variables are not scaled.
- Checking that the reciprocity, density and centrality results are still present when degree is controlled for.
- To get a sense of the "relative importance" of different factors in their correlation with income. This is best done looking at Table 17 in the appendix, where all regressors have been scaled to unit variance)
- Perhaps most importantly, to check whether at least some of the identified effects survive the addition of fine-grained geographical controls. This may alleviate concerns about previous research that used location as a way to match income, and may have suffered from confounding factors linked to location.

Table 15 shows a simple regression of the variables discussed above on income. As a reminder, income is not observed as a continuous variable, but as 13 possible categories, which is why we use interval regression. Interval regression is estimated with Maximum Likelihood, and therefore usual regression statistics such as the R-squared are not available (and would not make much sense in the presence of a categorical dependent variable). However, other regression models have been used as robustness checks (such as replacing the income category by the expectation of income within that category under the assumption that income is broadly log-normally distributed), and give findings consistent with interval regression. For clarity, we only discuss interval regression results.

Table 15 is structured as follows: the first column shows regression results in the whole sample, without any fixed effects. The second column uses the same variables as column

(1), but includes 100 geographical controls, in the form of indicator variables that have been built as a 10x10 grid of GPS coordinates (only the most common location of a user is considered when computing these indicator variables). Finally, column 3 uses cell-tower fixed effects: each cell tower has its own dummy variable. For computational tractability reasons, estimation of the fixed effects in column (3) are performed using the method of alternating projections present in the *lfe* package in R, which provided an approximate method of computing a large number of fixed effects.

	Monthly Income in U.S. Dollars		
	(1)	(2)	(3)
Incoming Degree	0.336***	0.210***	0.161***
	(0.013)	(0.013)	(0.015)
Outgoing Degree	0.750***	0.426***	0.283***
	(0.021)	(0.021)	(0.025)
Tie Reciprocity (%)	-47.135***	-27.463***	-21.585***
	(3.337)	(3.286)	(3.737)
Ego Network Density	-75.670***	-52.269***	-35.372***
	(6.151)	(6.010)	(6.813)
Eigenvalue Centrality	-36.259***	-17.695***	-10.880***
	(0.658)	(0.729)	(0.900)
Local Centrality	0.355***	0.049	-0.035
	(0.035)	(0.034)	(0.040)
Smartphone Ownership (ego)	17.340***	10.963***	5.264***
	(1.136)	(1.111)	(1.291)
Smartphone Ownership (most alters)	51.932***	30.341***	16.034***
	(3.403)	(3.306)	(3.891)
Location Fixed Effects	**No**	**10x10 GPS Grid**	**Towers (10,306)**
Number of observations	110247	105896	105896
Log Likelihood	-243,420.900	-229,282.300	-236,350.400
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 15: interval regression of degree, diversity, reciprocity and centrality variables on income. Note: a complete table with scaled regressors (to unit variance) can be found in the appendix.

As can be seen in the table, all but one result keep their significance even in the presence of controls, coarse geographical control and very fine-grained geographical controls.

This provides some reassurance that the results discussed above were not driven by correlation with another variable or by geographical confounding effects. For interpreting the scale of the effects, it is useful to consider that the average hourly wage in our country of interest is around 25 cents (in US dollar terms) an hour. This sheds some light on magnitudes: for example, having one additional incoming connection predicts an income higher by the equivalent of 1.3 worked hours every month (0.33/0.25), whereas an additional incoming connection predicts an income higher by the equivalent of three worked hours (per month). Owning a smartphone is associated with an income $17 higher; however, having the majority of one's alters be smartphone owners is associated with a greater income increase, at almost $52. Of course, no causal claims can be made with the data we are using: we simply set out to compare ego networks for low and high-income individuals, as this is still an open venue for social science research. Causal identification would likely require exogenous variation in network structure, which is beyond the scope of this paper.

A stylized illustration of the main findings

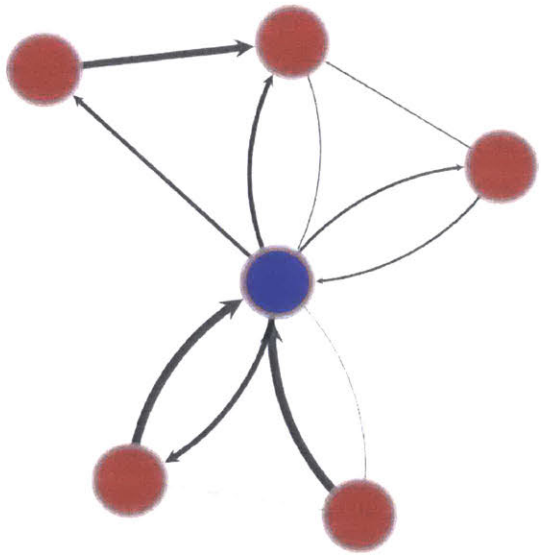


Figure 23 a stylized illustration of a low-income ego network

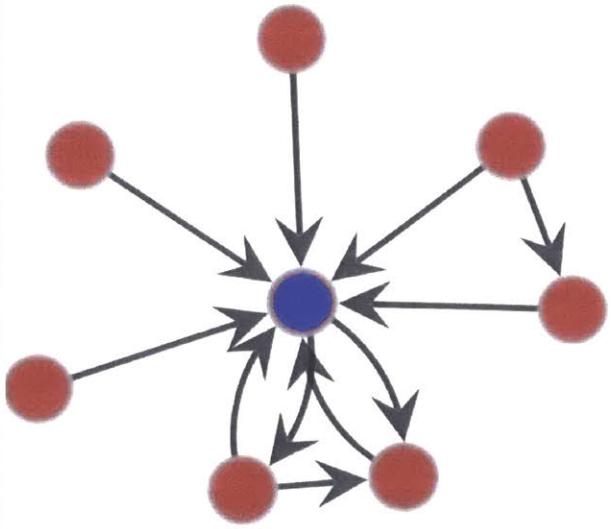


Figure 24 a stylized illustration of a high-income ego network

Figure 23 and Figure 24 provide a visual (and highly stylized) illustration of the main findings of the paper. The left network represents a low-income ego network and has lower degree (outgoing and incoming), higher reciprocity, higher density, and higher variation in tie strength. On the other hand, the right network represents a typical high-income network and has higher degree (outgoing and incoming), lower reciprocity, lower density, and lower variation in tie strength. It should be noted that the number of alters has been scaled for legibility. In reality average number of alters are about ten times as large (see Figure 20). The results regarding tie strength are now discussed in the next section:

Diversity and Habitual Behavior

Previous studies have empirically shown the link between regional access to socioeconomic opportunities and heterogeneity of social ties held by individuals living within that region. For example, Eagle, Macy, and Claxton (2010) showed that diversity of the phone communication network is strongly correlated with economic development of regions in the UK. In contrast, we are interested in studying the phenomenon on an individual level rather than on a regional level. We are interested in two main kinds of diversity of ego networks: *spatial diversity* and *tie strength* diversity.

Spatial Diversity

A more spatially diverse ego network may be indicative of access to multiple sources of information and entrepreneurship opportunities. Shannon entropy, a measure of a distribution randomness, is a natural choice for quantifying an ego's spatial exploration. We consider two ways of computing this indicator for each ego. First, we construct the discrete distribution of towers serving the ego and computed its normalized Shannon entropy as:

$$H(j) = \frac{-\sum_{i=1}^{k} p_{ij}\, log p_{ij}}{log k} \quad (1)$$

where j indicates the ego, k is the number cell towers ever serving the ego, and p_{ij} is the proportion of ego's calls ever served from tower j. This entropy is effectively a measure of ego's mobility and captures how diverse are the locations visited by the ego. Large

values of entropy indicate that the ego distributes her time evenly across all the locations.

In contrast to our first measure of spatial diversity which incorporates behavioral information only about the ego, our second measure captures the spatial diversity of alters in an ego network. In particular, we are interested in geographical distances between ego's and her alters' home locations, as large and diverse distances suggest access to novel information, typically unavailable in ego's own living area. Using distances to alters as a continuous variable allows to capture their true geographic dispersion centered around the ego over short and long distances. This information might not be available simply based on the home cell tower of alters as a categorical variable. Similar to the first measure of entropy, we can compute the differential entropy of distances between ego and alters home locations as follows:

$$H(j) = -\int_0^\infty p_j(y) log p_j(y) dy \tag{2}$$

where y indicates the distance between ego and alter home locations and $p_j(y)$ is the probability of such an ego-alter link. In contrast to our first measure of entropy, distance is not a categorical but a continuous variable and since we don't observe pj(y), we need to estimate H(j). The algorithm proposed by (Kozachenko and Leonenko 1987) is a non-parametric estimator applicable to a wide range of applications. We used this estimator for quantifying our second measure of geographic diversity. Figure 25 and Figure 26 show that wealthier individuals have higher values in both measures. This verifies the existing theory and the results of (Eagle et al. 2010) but on an individual-level.

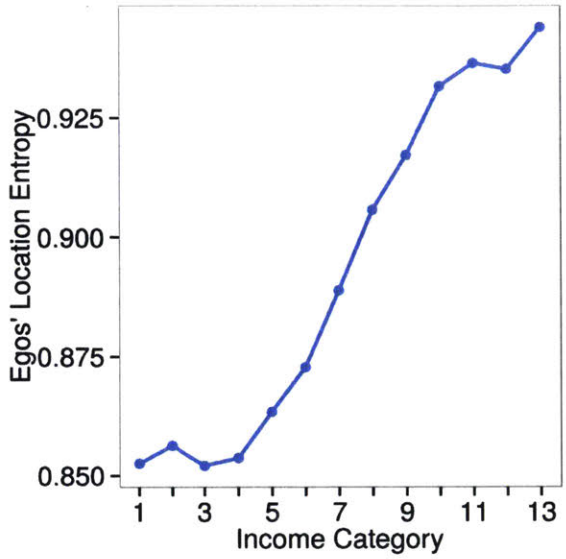


Figure 25 Median entropy of egos' locations grouped by income category

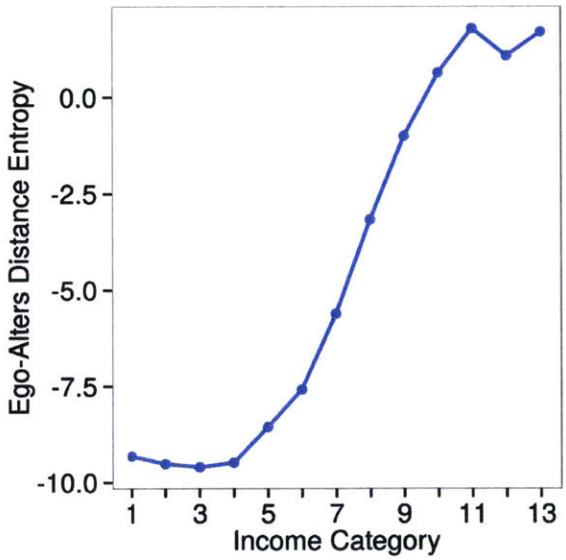


Figure 26 Median entropy of egos distance to alters' homes grouped by egos' income category

Tie Strength Diversity

In addition to geographical diversity, we study the relationship between ego's income and her social diversity which characterizes heterogeneity of the social ties. Social (or topographic) diversity is computed in a similar fashion to equation Text, the only difference being that i refers to an alter and p_{ij} is the proportion of egos calls made to or from the alter. This indicator measures the diversity of the types of relationships held by the ego. Low diversity may be interpreted as routine, habitual behavior in social relationships.

In contrast to findings of (Eagle et al. 2012), our Figure 27 suggests that wealthy individuals have lower diversity in tie strength. These divergent findings may be explained by an important difference between our study and (Eagle et al. 2010). Social diversity in (Eagle et al. 2012) is aggregated on a regional level which similar to location diversity captures access to novel sources of information for a typical individual in each region. However, our social diversity is computed on an individual level which may make it a measure of habitual behavior rather than access to information. Our findings suggest that as individuals get richer, they engage in less "foraging" behavior and become more habitual in their relationships due to their increased financial stability. This is may be associated with the decline in local centrality for the wealthiest shown in Figure 22. If wealthy have fixed, regular incomes then they need not engage in diverse exploration.

Finally, we treat the total number of phone calls made by each alter in the ego network as a continuous random variable and measure its differential entropy using the estimator for equation Text. If we treat number of phone calls as a proxy for social engagement, this variable is also a measure of social diversity, as higher values indicate more variety

of individuals in terms of status and activity in ego's immediate network. Figure 28 suggests that wealthier individuals are generally connected to a more diverse population in terms of network activity.

Table 16 summarizes these findings by regressing the interval of actual income values in US dollar on our four diversity metrics. The other ego network variables described in the previous section are also included in the regression but not shown for simplicity. The results generally match our expectations as explained above. However, there are some surprising findings. After controlling for location fixed effects, the location entropy of the ego loses its significance suggesting that perhaps the contrast we observe in mobility of egos are due to intrinsic differences between urban and rural areas. Individuals in urban areas are on average richer and visit a more diverse set of unique locations (quantified as the number of cell towers served). Furthermore, social diversity of alters becomes significant only after controlling for location fixed effects, even though Figure 28 shows its median is positively correlated with income.

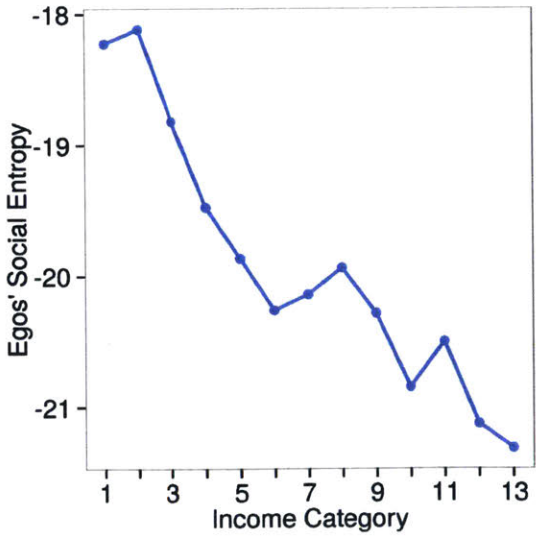


Figure 27 Median entropy of egos' phone calls grouped by income category

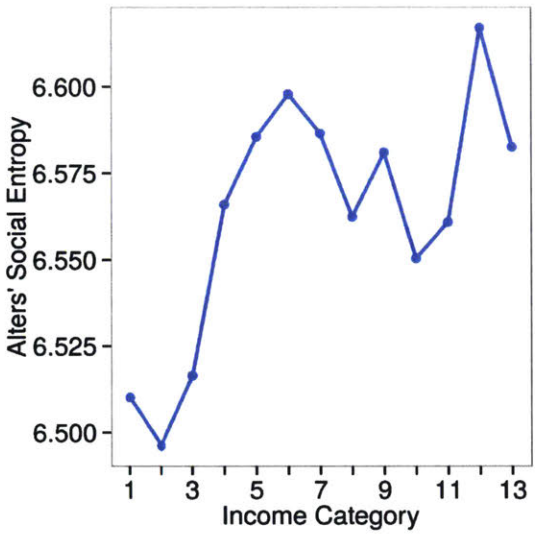


Figure 28 Median entropy of alters' phone calls grouped by income

	Monthly Income in US Dollars		
	(1)	(2)	(3)
Distance Diversity	17.806***	11.552***	5.404***
	(0.680)	(0.673)	(0.779)
Location Diversity	2.158***	-0.005	-0.764
	(0.765)	(0.755)	(0.877)
Diversity in number of calls (alters)	-0.370	2.662***	2.354***
	(0.652)	(0.653)	(0.764)
Diversity in number of calls (ego)	-5.752***	-4.409***	-3.587***
	(0.601)	(0.589)	(0.667)
Location Fixed Effects	**No**	**10x10 GPS Grid**	**Towers (10,306)**
Degree, reciprocity, centrality and control variables	**Included**	**Included**	**Included**
Number of observations	108737	104432	104432
Log Likelihood	-239,256.800	-225,775.600	-233,034.900
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 16: Interval regression of diversity and habitual variables on income.

This evidence seems to suggest that wealthier individuals have networks that are spatially dispersed and diverse in terms of their alters' overall social activity, but regular and predictable in terms of their own social engagement.

Conclusion

This paper investigates the differences between social networks of individuals with low and high income, based on individually matched income data. We find that wealthier individuals have higher degree, but lower density and reciprocity in their individual networks. They seem to have higher local centrality but lower global centrality. Finally, network diversity seems to be strongly correlated with income, with spatial diversity showing a positive correlation. Our analysis also suggests that as income grows individuals become more regular in their social engagement, perhaps as a consequence of having larger networks that become focused on fewer social ties.

It should be noted that our goal is not that of income prediction. Instead, we aim to investigate how typical or average ego-networks from low and high-income categories are different. In particular, our analysis only captures the mean effects of income variation and therefore our ego-network variables are not powerful enough for making accurate individual predictions. Prediction tasks could benefit from wide range of variables that reflect individual's own behavior rather than the structure of her ego-network. Variables such as mean duration of calls, phone type, top-up pattern and radius of gyration are highly predictive of income but were not the focus of our study.

A potential next step in this area of research is to seek causal identification of the effects we document here. This, however, would likely require exogenous shocks to network structure, and is left for future research.

Together with individual matching of income data, CDR data appears a very promising way of studying the economics of inequality, development, and income. We hope that

the economic magnitude of the correlations documented here will illustrate the potential of future research in this area.

Appendix

	income		
	(1)	(2)	(3)
Incoming Degree	15.128***	9.780***	7.651***
	(0.628)	(0.620)	(0.725)
Outgoing Degree	24.818***	14.348***	9.330***
	(0.756)	(0.760)	(0.892)
Tie Reciprocity (%)	-2.794***	-1.806***	-1.782***
	(0.498)	(0.491)	(0.557)
Ego Network Density	-6.335***	-3.255***	-1.440*
	(0.664)	(0.656)	(0.747)
Eigenvalue Centrality	-29.244***	-14.492***	-9.732***
	(0.679)	(0.742)	(0.917)
Local Centrality	2.825***	0.296	-0.247
	(0.415)	(0.409)	(0.471)
Distance Diversity	17.806***	11.552***	5.404***
	(0.680)	(0.673)	(0.779)
Location Diversity	2.158***	-0.005	-0.764
	(0.765)	(0.755)	(0.877)
Diversity in number of calls (alters)	-0.370	2.662***	2.354***
	(0.652)	(0.653)	(0.764)
Diversity in number of calls (ego)	-5.752***	-4.409***	-3.587***
	(0.601)	(0.589)	(0.667)
Smartphone Ownership (ego)	5.603***	3.598***	1.740***
	(0.398)	(0.391)	(0.456)
Smartphone Ownership (most alters)	5.748***	3.643***	1.966***
	(0.406)	(0.396)	(0.468)
Location Fixed Effects	**No**	**10x10 GPS Grid**	**Towers (10,306)**
Degree, reciprocity, centrality and control variables	**Included**	**Included**	**Included**
Number of observations	108737	104432	104432
Log Likelihood	-239,256.800	-225,775.600	-233,034.900
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 17: Full interval regression with scaled regressors (scaling to unit variance)

References

Bjorkegren, D., and Grissen, D. 2015. "Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment," *Available at SSRN 2611775*.

Blumenstock, J., Cadamuro, G., and On, R. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata," *Science* (350:6264), American Association for the Advancement of Science, pp. 1073–1076. (https://doi.org/10.1126/science.aac4420).

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A. 2014. "Once upon a Crime: Towards Crime Prediction from Demographics and Mobile Data," in *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 427–434.

Brass, D. J., and Burkhardt, M. E. 1993. "POTENTIAL POWER AND POWER USE: AN INVESTIGATION OF STRUCTURE AND BEHAVIOR.," *Academy of Management Journal* (36:3), Academy of Management, pp. 441–470. (https://doi.org/10.2307/256588).

Calvo-Armengol, A., and Jackson, M. O. 2004. "The Effects of Social Networks on Employment and Inequality," *American Economic Review*, JSTOR, pp. 426–454.

Cowan, R., and Jonard, N. 2004. "Network Structure and the Diffusion of Knowledge," *Journal of Economic Dynamics and Control* (28:8), pp. 1557–1575. (https://doi.org/10.1016/j.jedc.2003.04.002).

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., and Tatem, A. J. 2014. "Dynamic Population Mapping Using Mobile Phone Data," *Proceedings of the National Academy of Sciences* (111:45), pp. 15888–15893. (https://doi.org/10.1073/pnas.1408439111).

Eagle, N., Macy, M., and Claxton, R. 2010. "Network Diversity and Economic Development.," *Science (New York, N.Y.)* (328:5981), American Association for the Advancement of Science, pp. 1029–31. (https://doi.org/10.1126/science.1186605).

Eagle, N., Macy, M., and Claxton, R. 2012. "Network Diversity and Economic Development," *Science* (335:March), pp. 1215–1220.

Enns, E. A., and Amuasi, J. H. 2013. "Human Mobility and Communication Patterns in Côte D'ivoire: A Network Perspective for Malaria Control," *Mobile Phone Data for Development: Analysis of Mobile Phone Datasets for the Development of Ivory Coast* (1), D4D Challenge sponsored by Orange.

Frias-Martinez, V., and Virseda, J. 2012a. "On the Relationship Between Socio-Economic Factors and Cell Phone Usage," in *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ICTD '12, New York, NY, USA: ACM, pp. 76–84. (https://doi.org/10.1145/2160673.2160684).

Frias-Martinez, V., and Virseda, J. 2012b. "On the Relationship Between Socio-Economic Factors and Cell Phone Usage," in *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ICTD '12, New York, NY, USA: ACM, pp. 76–84. (https://doi.org/10.1145/2160673.2160684).

Granovetter, M. 1995. *Getting a Job: A Study of Contacts and Careers*, University of Chicago Press.

Granovetter, M. S. 1973. "Granovetter - 1973 - The Strength of Weak Ties," *American Journal of Sociology*, pp. 1360–1380. (https://doi.org/10.1037/a0018761).

References

Bjorkegren, D., and Grissen, D. 2015. "Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment," *Available at SSRN 2611775*.

Blumenstock, J., Cadamuro, G., and On, R. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata," *Science* (350:6264), American Association for the Advancement of Science, pp. 1073–1076. (https://doi.org/10.1126/science.aac4420).

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A. 2014. "Once upon a Crime: Towards Crime Prediction from Demographics and Mobile Data," in *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 427–434.

Brass, D. J., and Burkhardt, M. E. 1993. "POTENTIAL POWER AND POWER USE: AN INVESTIGATION OF STRUCTURE AND BEHAVIOR.," *Academy of Management Journal* (36:3), Academy of Management, pp. 441–470. (https://doi.org/10.2307/256588).

Calvo-Armengol, A., and Jackson, M. O. 2004. "The Effects of Social Networks on Employment and Inequality," *American Economic Review*, JSTOR, pp. 426–454.

Cowan, R., and Jonard, N. 2004. "Network Structure and the Diffusion of Knowledge," *Journal of Economic Dynamics and Control* (28:8), pp. 1557–1575. (https://doi.org/10.1016/j.jedc.2003.04.002).

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., and Tatem, A. J. 2014. "Dynamic Population Mapping Using Mobile Phone Data," *Proceedings of the National Academy of Sciences* (111:45), pp. 15888–15893. (https://doi.org/10.1073/pnas.1408439111).

Eagle, N., Macy, M., and Claxton, R. 2010. "Network Diversity and Economic Development.," *Science (New York, N.Y.)* (328:5981), American Association for the Advancement of Science, pp. 1029–31. (https://doi.org/10.1126/science.1186605).

Eagle, N., Macy, M., and Claxton, R. 2012. "Network Diversity and Economic Development," *Science* (335:March), pp. 1215–1220.

Enns, E. A., and Amuasi, J. H. 2013. "Human Mobility and Communication Patterns in Côte D'ivoire: A Network Perspective for Malaria Control," *Mobile Phone Data for Development: Analysis of Mobile Phone Datasets for the Development of Ivory Coast* (1), D4D Challenge sponsored by Orange.

Frias-Martinez, V., and Virseda, J. 2012a. "On the Relationship Between Socio-Economic Factors and Cell Phone Usage," in *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ICTD '12, New York, NY, USA: ACM, pp. 76–84. (https://doi.org/10.1145/2160673.2160684).

Frias-Martinez, V., and Virseda, J. 2012b. "On the Relationship Between Socio-Economic Factors and Cell Phone Usage," in *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ICTD '12, New York, NY, USA: ACM, pp. 76–84. (https://doi.org/10.1145/2160673.2160684).

Granovetter, M. 1995. *Getting a Job: A Study of Contacts and Careers*, University of Chicago Press.

Granovetter, M. S. 1973. "Granovetter - 1973 - The Strength of Weak Ties," *American Journal of Sociology*, pp. 1360–1380. (https://doi.org/10.1037/a0018761).

Group, D. R. 2014. *A World That Counts. Mobilising the Data Revolution for Sustainable Development.*

Kish, L. 1949. "A Procedure for Objective Respondent Selection within the Household," *Journal of the American Statistical Association* (44:247), pp. 380–387. (https://doi.org/10.1080/01621459.1949.10483314).

Kozachenko, L., and Leonenko, N. 1987. "Sample Estimate of the Entropy of a Random Vector," *Problemy Peredachi Informatsii* (23:2), pp. 9–16. (http://www.mathnet.ru/eng/ppi797).

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. 2009. "Computational Social Science," *Science* (323:5915), pp. 721–723. (https://doi.org/10.1126/science.1167742).

Llorente, A., Garcia-Herranz, M., Cebrian, M., and Moro, E. 2015. "Social Media Fingerprints of Unemployment.," *PloS One* (10:5), Public Library of Science, p. e0128692. (https://doi.org/10.1371/journal.pone.0128692).

Pokhriyal, N., Dong, W., and Govindaraju, V. 2015. "Virtual Networks and Poverty Analysis in Senegal," *CoRR* (abs/1506.0). (http://arxiv.org/abs/1506.03401).

Rees, A. 1966. "Information Networks in Labor Markets," *The American Economic Review* (56:1/2), American Economic Association, pp. 559–566.

Reis, F., and Ferreira, P. 2015. "Understanding the Role of Social Networks on Labor Market Outcomes Using a Large Dataset from a Mobile Network," *ICIS 2015 Proceedings*.

Saint-Jacques, G., and Brynjolfsson, E. 2015. "Information Technology and the Rise of the Power Law Economy," *ICIS 2015 Proceedings*. (http://aisel.aisnet.org/icis2015/proceedings/EconofIS/8).

Seibert, S. E., Kraimer, M. L., and Liden, R. C. 2001. "A Social Capital Theory of Career Success," *Academy of Management Journal* (44:2), Academy of Management, pp. 219–237.

Smith, C., Mashhadi, A., and Capra, L. (2013). *Ubiquitous Sensing for Mapping Poverty in Developing Countries*.

Sundsøy, P. R., Bjelland, J., Iqbal, A. M., and Jahani, E. 2016. *Deep Learning Applied to Mobile Phone Data for Individual Income Classification*, (Icaita), pp. 96–99.

Wegener, B. 1991. "Job Mobility and Social Ties: Social Resources, Prior Job, and Status Attainment," *American Sociological Review*, JSTOR, pp. 60–71.

Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., Engø-Monsen, K., and Buckee, C. O. 2015. "Impact of Human Mobility on the Emergence of Dengue Epidemics in Pakistan," *Proceedings of the National Academy of Sciences* (112:38), National Acad Sciences, pp. 11887–1189

THIS PAGE INTENTIONALLY LEFT BLANK

The Strength of Weak Ties: Causal Evidence using People-You-May-Know Randomizations

Coauthored with:
Edoardo M. Airoldi, Harvard
Sinan Aral, MIT
Erik Brynjolfsson, MIT
Ya Xu, LinkedIn

Abstract

The causal relationship between tie strength and labor market outcomes is of interest to a large variety of actors, from individual workers seeking to optimally allocate their resources as they develop their own social network to firms seeking to leverage candidates' networks in their recruitment process. It is also of interest to a social planner or a professional social network platform interested in increasing efficiency in labor market matching processes or increasing equality of opportunity. Using a number of "People You May Know" experiments (testing recommendation algorithms) conducted at LinkedIn between 2014 and 2016, we seek to identify the sign of the causal relationship between tie strength and labor market mobility in two different ways. First, by conducting an edge-level regression of job transmission on tie strength using a PYMK randomization as an instrument. Then, with an individual-level regression of number of jobs reported on individual network clustering coefficient, using over 700 past treatments as instruments with regularization. Both sets of results point to decreasing returns in the relationship between structural tie strength and mobility. These results indicate that a strong tie is not always individually more useful than a weak one, and that the most useful ties are likely not the weakest or the strongest, but the ones that strike a good compromise between strength and diversity.

Introduction

Tie strength may impact worker's labor market outcomes in a variety of ways. A commonly studied phenomenon is job transmission over social ties. If an individual works at a certain firm, she may inform her social ties of job openings, or leverage her personal knowledge of her friends and their abilities to help her firm quickly (and cheaply) identify promising candidates - something many companies encourage through a referral bonus. This results in an individual's connections having a higher likelihood of ending up working at the same firm as her. In a survey, (Granovetter 1973) finds that more people report obtaining their job through a weak social tie than through a strong one. This observation, however, leaves open the question of whether weak ties are cited more often by respondents simply because they are more numerous, or because they are also individually more valuable (for the purposes of finding a job) than strong ones. Two main challenges stand in the way of answering this question. First, until recently, the data required to observe individual's networks and job transmission was not available. Estimating the strength of a tie requires rich data, measuring the intensity of interactions between any pair of individuals, and complete social network data, allowing to compute structural measures of tie strength (like number of friends in common).

Furthermore, as job transmission is a rare event, large-scale data collection is necessary.

The second, more fundamental challenge is one of simultaneity and endogeneity. Individual's labor market outcomes both determine and are determined by their social networks, and the evolution of both is very likely to be correlated with a number of unobservable factors confounding correlational analyses. An individual's network will likely influence the individual's job market options, but an individual's endeavor toward

switching jobs may also lead the individual to grow his/her network in a certain way. In this paper, we address this second challenge by using experimental data, rather than observational data.

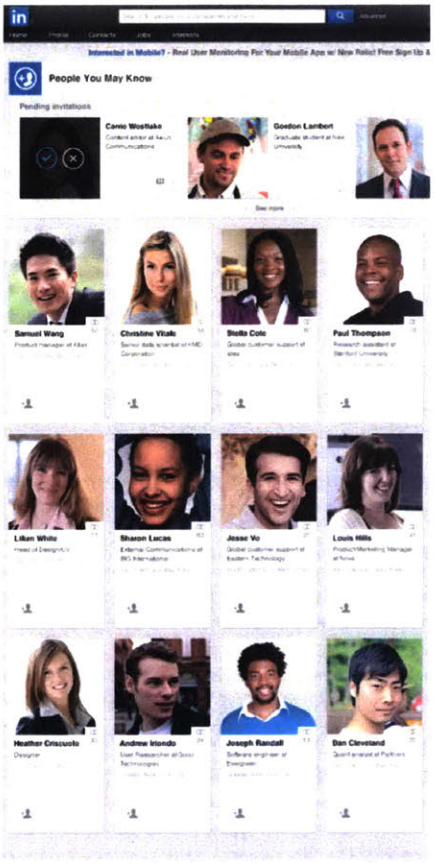


Figure 29 Screen Capture of People-You-May-Know page on LinkedIn

Strategy no. 1: Edge-level regression with one past randomization as an instrument

An in-depth observational investigation of the relationship between tie strength and job transmission can be found in (Gee, Jones, and Burke 2016; Gee et al. 2017) construct a proxy measure of job transmission based on three criteria:

- User A reports working at company *c* at date D_1.

- User B report working at that same company *c* at a later date D_2, with D_2 and D_1 being at least one year apart.
- User A and user B were friends on the social network at least one full year before D_2.

When these three criteria are met, a tie is then tagged as a "sequential job" tie. The paper then shows a series of edge-level regressions (where the unit of analysis is the tie itself, as opposed to the individual), using the above measure as the dependent variable and various measures of tie strength as regressors. The authors find a positive correlation between tie strength (as measured by interactions on Facebook and number of friends in common) and job transmission. We are interested in a somewhat different question: causally, at the margin, when adding a new tie, does tie strength have positive or negative impact on the probability of job transmission? Where is one's energy better spent: developing strong ties or weak ties? To answer this question, we rely on a past randomized experiment as a source of exogenous network variation.

Data

To carry out the edge-level analysis, we choose to remain close to Gee et al's (2017) approach of using a regression, in which the unit of analysis is the social tie, the dependent variable is a binary indication of a *sequential job tie*, as defined above, and various quantifications of tie strength are included among the regressors. In particular, we contrast two quantifications of tie strength: interaction intensity and number of connections in common. Departing from a purely observational approach, in order to obtain causal identification of the role of strong ties on job transmission, we rely on a two-week ex-

periment conducted by LinkedIn's People You May Know (PYMK) service[6], which recommends possible new ties to users when they log in to the site, conducted in early 2015. The experiment we exploit was carried out to test several different (randomly allocated) tie recommendation algorithms. We construct a sample of edges two years after the experiment, and compute a *sequential job* binary indicator using the same definition as in Gee et al. . However, we also include in the sample all individuals with ties that were created as a result of the experiment, and not only the ones that ended up resulting in a job transmission[9].

As a first measure of tie strength, we compute the intensity of interactions among users along various dimensions, such as interactions through the feed, messaging, recommendations, and others. These scores along these dimensions are then averaged and scaled to produce a variable labeled *interaction intensity*. As a second measure of tie strength, we count the number of friends any two connected individuals had in common when the tie was created. Individual's degrees are also entered as controls in our regression.

In the randomized experiment that provides exogenous variation in our analysis, several tie recommendation variants were allocated using Bernoulli randomization across users of the platform. Relatively to the control variants of the recommendation algorithm, treatment variants recommend more triad-closing ties. Millions of invitations were sent which ended up being accepted. We restrict our analysis to these accepted edges, and

---

[9] Because of this, the frequency of sequential jobs observed in the data set we construct is about ten times smaller than in Gee et al (2017-1). If we were to restrict the sample to only individuals with at least one recorded sequential tie, the estimated frequency of sequential job ties would be about 3%, which is close to the one reported by the authors.

use, as an instrument for our two measures of tie strength, the several treatment variant assigned to the invitation sender.

Results

Table 18 shows an excerpt of the regression results. The first column shows the results of a simple OLS regression of the indicator of whether each tie resulted in a sequential job on our measures of tie strength and a number of control variables. Controls, such as whether both individuals went to the same school, or whether they are located in the same region or city, are not shown; neither is their age difference. The second column shows the results of the same model specification, but employing a Probit model instead. Finally, the third column shows the results of a two-stage least square (2SLS) estimation. The results of the first two columns are broadly consistent with the observational results of (Gee et al. 2017): more interaction and more structural closeness (as defined by the percentage of friends in common) is associated with a higher likelihood of job transmission. Also similar with their observational results are the coefficients on the controls: similarity between members increases the likelihood of job transmission, and dissimilarity decreases that likelihood. Having gone to the same school, or living in the same region or the same city is associated with more job transmissions, whereas greater age differences are associated with fewer.

We conduct a number of tests in order to check whether our instrument is weak or not. Our 2SLS estimates are very similar to other available instrumental variable estimators, such as Limited Information Maximum Likelihood or Fuller. Our first-stage F-statistic, which is well over 10, and an Anderson-Rubin test for weak instruments rejects the null that the instrument is weak.

Table 18 Dependent variable: Sequential Job indicator variable

Dependent variable: Sequential Job dummy			
	OLS	Probit	2SLS
Interactions (messages)	**.0002*** **	**.018*** **	**-.005**
% friends in common	**.081*** **	**8.779*** **	**1.882*** **
...High % friends in common (dummy)	**.002*** **	**.178*** **	**-.061*** **
... Interaction term	**-.079*** **	**-8.740*** **	**-.884** **
Same school	.0003***	.019***	-.002
Age difference	-.00001***	-.001***	-.00001
Same region	.001***	.130***	.001
Same city	.001***	.088***	.001
Ego connection count control	Included	Included	included
Alter connection count control	Included	Included	included
Constant	included	included	included
Same-gender controls	included	included	included
Observations	19,763,317	19,763,317	19,763,317

Nonlinear relationships.

Figure 31 and Figure 31 (axes scales hidden for confidentiality) illustrate nonlinear relationships between the two measures of tie strength we consider and the probability of

job transmission. Figure 1 shows the relationship between interaction intensity and the probability of observing a sequential job. The black dots show the unconditional relationship as present in the raw data, and the blue dots show the interaction intensity as predicted from the first stage equation (all edges in our dataset are binned into ten groups which are then shown on this scatterplot).

Figure 31 shows the relationship between interaction intensity and job transmission: the more people interact on the platform, the more likely it is that one will end up working for the same company as the other one.

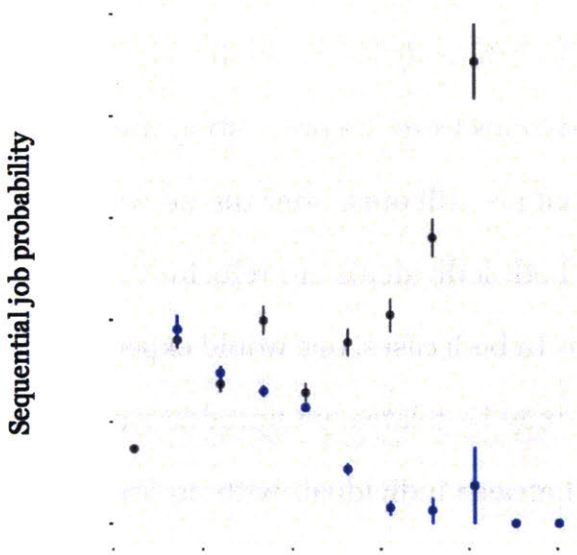


Figure 31 Interaction Intensity

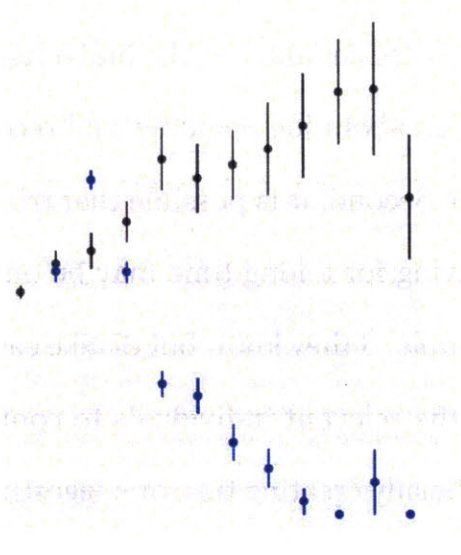


Figure 31 Number of connections in common

Figure 31 shows the relationship between probability of sequential jobs and number of connections that the ego and the alter had in common before the tie was created; i.e. the

number of open triangles the tie closed when it was created. In both figures, the difference between the blue dots and the black ones is striking. When looking at a pure correlation (black dots), we see a positive relationship; descriptively, when people have more friends in common, a sequential job is more likely to be observed. However, leveraging the experimental data and using fitted values from the first stage (blue dots) reveals a more complex relationship. When disconnected individuals already have many friends in common, the tie that results from nudging them to connect has a lower probability of leading to a sequential job than if they had few friends in common. One can hypothesize two mechanisms that might lead to this result. First, it is possible that such a tie is redundant, i.e., the individual most likely already has access to most of the information about job openings and recommendations from the preexisting friends in common. Second, it is possible that triangles that are still open after the network has been evolving for a long time may be ones that both individuals are reluctant to create, for example, if they know but dislike each other. In both cases, one would expect that nudging the relevant individuals to connect on may be relatively unproductive. Similarly, artificially creating ties, or suggesting ties, between individuals with no friends in common also seems to have low value. This may be because high network distance is the sign of large differences between individuals, so that they have little to gain from becoming connected.

Strategy no. 2: Individual-level regression with many past randomizations as instruments.

In this approach, we shift approaches from edge-level regression to individual-level regression. The LinkedIn teams developing new tie recommendation algorithms carry out randomized experiments to evaluate the performance of the new algorithms routinely.

Here, we use many of these experiments as instruments. For the purpose of this analysis, we restrict our attention to a subset of the LinkedIn graph, namely all users reporting a location within the San Francisco Bay area. At the end of each PYMK experiment, we construct a graph consisting of the edges that existed at that date, and compute the clustering coefficient for each member. The clustering coefficient, computed at the individual level, is the ratio of the number of ties existing in an individual's 1.5-out ego network to the number of ties that would exist if the 1.5-out ego network were fully connected (i.e. if all of the focal individual's peers were connected to each other). In other words, this gives us the proportion of closed triads around the focal individual. We use this as our variable of interest. Each experiment has many variants, leaving us with over 700 experiment-variant combinations, which are all potential instruments.

Table 19 Individual-level regression coefficient, with and without instrument selection

	All Instruments	**With Instrument Selection**
Clustering coefficient	**-10.25**	**-15.83**
s.d.	5.07	8.66

P-value	0.0433	0.0682

As a dependent variable, for each member, we count the number of different positions that are listed on her profile. We collect this value at the end of each experiment. Using different outcome variables, such as progress in seniority levels, or number of firms worked for, does not significantly change the outcome. We follow the approach of instrumental variable cross validation proposed by (Peysakhovich and Eckles 2017) in order to select the strongest instruments. The procedure selects 493 instruments, and reveals a negative effect of increasing clustering coefficient (closing triangles around individuals) on number of jobs reported. In this specification, increasing the clustering coefficient by 10% would decrease the number of positions reported by 1.5. It is likely that closing too many triangles around an individual fills her ego-network with relatively less useful ties, and reduce her exposure to novel information and to different firms, therefore reducing her labor market mobility options, resulting in a lower number of reported jobs and positions over time.

References

Gee, Laura K., Jason J. Jones, Christopher J. Fariss, Moira Burke, and James H. Fowler. 2017. "The Paradox of Weak Ties in 55 Countries." *Journal of Economic Behavior and Organization*. https://doi.org/10.1016/j.jebo.2016.12.004.

Gee, Laura K, Jason Jones, and Moira Burke. 2016. "Social Networks and Labor Markets: How Strong Ties Relate to Job Finding on Facebook's Social Network." *Journal of Labor Economics* 35 (2). The University of Chicago Press:485–518. https://doi.org/10.1086/686225.

Granovetter, Mark S. 1973. "Granovetter - 1973 - The Strength of Weak Ties." *American Journal of Sociology*. https://doi.org/10.1037/a0018761.

Peysakhovich, Alexander, and Dean Eckles. 2017. "Learning Causal Effects from Many Randomized Experiments Using Regularized Instrumental Variables," January. http://arxiv.org/abs/1701.01140.