DOCTORAL DISSERTATION

APPLICATIONS AND ECONOMIC IMPACT OF MACHINE LEARNING AND BLOCKCHAIN TECHNOLOGIES

by

Nikhil Malik

Submitted to the

David A. Tepper School of Business

in partial fulfillment for the requirements for the degree of

DOCTOR OF PHILOSOPHY

in the field of Industrial Administration

at Carnegie Mellon University

DISSERTATION COMMITTEE:

Kannan Srinivasan (co-chair)

Param Vir Singh (co-chair)

Nitin Mehta

Bryan Routledge

Carnegie Mellon University, September 2020

Abstract

Machine Learning (ML) and Blockchains have been two major technology disruptions in the last decade. On the one extreme, Blockchain decentralizes decision making power to a crowd of anonymous participants. On the other extreme, ML centralizes decision making into uninterpretable algorithms. The first chapter uses ML as a tool to study behavioral biases in the labor markets. The second and third chapter deal with strategic interaction of market participant with Blockchain and ML platforms respectively.

The first chapter, co-authored with Prof Param Vir Singh and Prof Kannan Srinivasan, examines biases arising from attractive appearance. We show that Preference Bias contributes to an attractiveness gap of 0.52% per year, adding to a 2.4% gap over a 15-year career. Belief Bias does not have a statistically significant contribution in our sample of 43,533 MBA graduates. This finding is important because Belief Bias, arising from evaluators group-level priors on subjects job fit, can be overcome by providing rich performance information. But, Preference Bias, arising from evaluators taste for social, romantic or marital relationship with attractive subjects, can be harder to eliminate. We make use of ML based image morhping of subject appearance and page ranking of subject career milestones to construct a pseudo random experiment on observational data.

The second chapter, co-authored with Prof Manmohan Aseri, Prof Param Vir Singh and Prof Kannan Srinivasan, examines peer to peer payments on Blockchains. We show that upgrade to Bitcoin payment throughput is rolled back by tacit collusion among Bitcoin miners. We identify an intervention of banning miners beyond a maximum compute power to eliminate collusion. But, such an intervention makes payments less secure from double spend attacks. Thus owing to the dual threat, of collusion and double spend attacks, its untenable to offer a high througput payment ledger to users with widely different willingness to pay fees, bear delay and risk attacks. We advocate miner collusion as a useful mechanism where a chunk of excess collusion revenue endogenously spill over into an investment into platform's security.

The second chapter examines Machine Learning (ML) pricing in housing market. These ML models are revised regularly using recent sample of sales. The recent sales are themselves confounded by previous version of the ML model. We theoretically show how this Feedback Loop creates a self fulfilling prophecy where ML over estimates its own prediction accuracy and market participants over rely on ML predictions. We formulate size of resulting pricing bias. We identify conditions on ML and market characteristics such that participants are worse off after introduction of ML. We use data from Zillow's Zestimate to provide empirical evidence for necessary primitives of our theoretical model.

Table of Contents

Chapter 1:
When does beauty pay?
An ML based measurement of Beauty Bias

Abstract

A rich literature has demonstrated facial attractiveness related discrimination (beauty bias[1]) in a wide range of contexts, such as personal marketing, election voting and employment. Under a controlled lab setting, most of these studies have found that attractiveness is rewarded. However, these studies cannot imitate long-term, real world interactions and thus lack external validity. In this paper, we show the long-term dynamics of bias to discern among sources of attractiveness bias.

We investigate two sources of attractiveness bias, namely, belief based and preference based. Belief based bias against subjects exists because evaluators have group-level priors based on the subjects' attractiveness. These priors are overcome as the evaluator obtains objective signals of performance. Preference based bias exists because evaluators have an inherent taste for a social, romantic or marital relationship with attractive subjects. We use one of the largest archival longitudinal data sets (43,533 MBA graduates) in this area of research to identify these two sources. We find that attractiveness is associated with a 2.4% gap over a 15-year career period. This gap is largely explained by preference bias, which is associated with an attractiveness gap of 0.52% per year. On the other hand, belief bias has no significant role in post-MBA professional careers. This is a significant finding because belief bias toward an individual can be minimized by the individuals' performance information. However, preference based biases are much harder to remove.

In our setting, there are two key challenges in working with unstructured data. First, for an individual, we observe only one current picture, which is taken up to 25 years after the start of the individual's professional career. We build a generative deep learning model to create life-like versions of a face, thus allowing us to emulate the employers' perceptions of how the individual looked at a younger age. Second, individuals move across job profiles, companies and locations, thereby making it difficult to directly compare their career milestones. We construct a preference order for jobs (job rank) based on observed job switching and text-based job title similarities.

[1] In this paper, we refer to individuals whose faces receive an above average attractiveness rating by evaluators as "attractive" and others who do not receive this rating as "plain-looking." Beliefs or preferences based differences are referred to as the following: an "Attractiveness Bias," an "Attractiveness Gap" and an "Attractiveness Premium." Note that some of the prior literature uses phrases such as a "beauty bias" and a "beauty premium" for the same concept.

1. Introduction

Marketing literature has studied the impact of the visual appeal of product aesthetics as well as the impact of product endorsers on product demand (Petty 1983, Richins 1991, Reingen 1993, Henderson-King 1997, Martin and Gentry 1997, Cryder et al. 2017, Claudia and Shu 2010). The literature has repeatedly shown the effective use of attractive solicitors, models and advertising endorsers on potential customers. Abercrombie & Fitch made headlines for hiring "young, attractive, mainstream athletic types and cheerleaders who might be their girlfriends" to work in their stores (Edwards 2003, Bryant 2003).

One would expect looks to have a particularly strong impact when the product and the endorser are merged, such as in personal marketing. Willis and Todorov (2006) show that it takes just seconds of seeing a face to infer personality traits such as competence, trustworthiness and aggressiveness. These perceptions impact election results, crime convictions and mate choices (Olivola & Todorov 2014). Moreover, positively perceived face traits correlate to the attractiveness of the faces (Langlois et al. 2000, Olivola & Todorov 2010). It is not surprising that an attractiveness premium has been found for professionals, such as doctors (Watkins and Jones 2015), lawyers (Biddle and Hamermesh 1998), retail managers (Watkins and Jones 2015), litigants (Kulka and Kessler 1978), and politicians (Olivola and Todorov 2010).

However, this body of research has mostly studied attractiveness bias in static, experimental situations (Mobius and Rosenblat 2005).[2] Individuals are asked to play hypothetical roles, such as a jury member (Kulka and Kessler 1978), a manager (Watkins and Jones 2015), a likely convict or an able worker. In these settings, individuals choose between attractive and plain-looking subjects, based on short interactions. They are given no performance history or significant time to learn about the subjects. Additionally, the evaluator does not expect to spend time with the subject after their evaluation. Experiments thus control for entangled mechanisms in order to perform causal inference of attractiveness bias effect size. These characteristics of experiments

[2] Without revealing the dynamics or sources of bias, other studies (Hamermesh and Biddle 1994, Zebrowitz and Donald 1991) that look at long term archival data show an attractiveness premium over a significant career length.

are in contrast with real world settings where a manager not only has more information but may also expect to have an ongoing social (or even romantic or marital) relationship with an employee. Experiments are thus limited in their external validity.

Further, short, static settings make it difficult to identify the underlying source of bias because different sources may lead to the same patterns of a single-shot outcome (Bohren et al. 2018, Fang and Moro 2011). Progression in professional roles depends on the continuous manager evaluation of the employees' performance. When employees are starting out, there is limited performance history. An MBA graduate with a 3.7 GPA has similar grades and the same classes as dozens of his or her peers. Similarly, the investment banking interns' first year is spent rotating across company divisions in training, producing little of value to differentiate themselves. Discrimination may be in play if attractive individuals are systematically more or less likely to be evaluated higher and to obtain better jobs or early promotions, even in the absence of any objective performance signals. However, suppose that the individuals start producing outputs of similar quality. Would the attractiveness bias persist or disappear? The answer to this question depends critically upon the source of the bias.

We focus on two major sources of discrimination that are proposed in the literature, namely *preference based* (Becker and Geer 1957) and *belief based* (Fang and Moro 2011, Phelps 1972). The two sources would lead to different dynamics of the attractiveness bias. In the preference-based bias, the evaluators have inherent taste (or distaste) for employees who are attractive. This could simply be a result of the evaluators being able to spend time around the attractive individuals. Although performance on the job provides informative signals of an employee's ability, it will have no impact on changing the preference of the evaluator. In contrast, in belief-based bias, the evaluator at the outset perceives attractive individuals to be on average more or less able than their plain-looking counterparts. Dion et al. (1972) suggest a positive perception of social skills, mental health and personality from attractive looks. Observing performance on the job will overcome such perceptions and reduce discrimination. This particular source of discrimination is also known as statistical discrimination (Altonji and Pierret 2001). An evaluator has group-level priors that diminish once the evaluator obtains objective signals of ability from the individual. Thus, if the source of bias is preference based, the bias will persist; otherwise, the bias will diminish.

To discern the dynamics of attractiveness bias, we require longitudinal data on the career progress of individuals. Hence, we collected a detailed panel data set on the career milestones and the educational background of 43,533 individuals selected from a professional social network. Consider an attractive and a plain-looking individual starting at the same job. The attractive individual has a 50.52% probability of exceeding their plain-looking counterpart by the end of year, i.e., the attractive individual accumulates a 0.52% attractiveness gap per year. This bias exists between MBA graduates who had a similar undergraduate education, pre-MBA experience and MBA program. We discuss why this premium is explained by *preference bias* alone. We cannot say that there is no belief bias at all; we can only say that it has no significant role in the post-MBA professional (white collar) job market.

Overall, the small size of preference bias (0.52% gap per year) and the insignificant degree of belief bias seem to suggest a marginal attractiveness premium. However, these yearly premiums add up to a 2.4% attractiveness gap in a 15-year career. In comparison, the gender-pay gap in the US stands at over 1% per year (Stanley and Jarrell 1998) and adds up to an 18%-22% gap when averaged across all jobs. The gender gap is estimated to be smaller at 1%-6% (Hay Group 2016) when the same set of jobs is considered between men and women. Note that these numbers are not directly comparable. We measure success as the quick attainment of a desired job; however, most gender gap studies look at average wages.

Our longitudinal sample presents major challenges in dealing with unstructured data in the shape of images and text. We need to rate the attractiveness of a large number of subjects. Additionally, we only observe a single instance of the subject’s self-posted profile picture. This creates three major challenges: (1) the derivation of a measure of attractiveness from a high-dimensional (unstructured) image pixel-level data, (2) the obfuscation of the underlying facial appearance by the subject’s choice of temporary accessories, such as clothing, hair style and expressions, (3) a single current snapshot that must be used to extrapolate the appearance of an individual at the time of the individual’s MBA program graduation up to 25 years prior. The first is a relatively common obstacle that has been tackled in recent times by using deep learning models supervised by labels from human raters. To ensure that our deep learning model’s performance is not affected by extraneous features, we use a model capable of disregarding temporary features and focusing only on permanent ones. Finally, obtaining a rating of attractiveness for an individual over a number of years from only one profile image poses a serious challenge. It has been established in the literature that attractiveness evolves with age (Zhang et al. 2017). As a result, comparing a 25-year-old individual with a 45-year-old individual in terms of attractiveness could be misleading. More importantly, different individuals age differently. For example, consider a case of two individuals A and B. Let us say that at 25, A is more attractive than B. Now, at 45, A will not necessarily continue to be more attractive than B. Even if A continues to be more attractive than B, the gap between the two could change significantly. Hence, to do a fair comparison over time, we need to model the evolution of attractiveness for people who look like A and who look like B. We use a generative deep learning model (Conditional Adversarial Auto Encoder, Zhang et al. 2017) to calculate the attractiveness of an individual over time.

In addition to building a quantitative measure out of the image data, we need to establish a measure of success comparable between individuals. We observe all career milestones (job title, company, location and timelines) that need to be ranked for each individual. We need to be able to order jobs such that a higher-ranked job is more desired by the job market participants. We rely upon traditional labor economics methods (Baker et al. 1994 and Gayle et al. 2012) that establish a preference order between jobs by utilizing observed pairwise job switches. A large number of job switches from job A to job B implies that job B is more desirable. This approach works well when the number of jobs to be ranked are small, e.g., C-suite executive positions. Our unstructured data is sparse and includes over 10,000 unique employers and 10,000 unique titles. Therefore, we improve upon the basic idea by building a dense representation of jobs and running a variant of a Page Rank algorithm to calculate a measure of a job’s rank. The resulting job ranks are intuitive and can be used by other studies that investigate career transitions by using large scale unstructured data.

Our paper makes three main contributions. First, we discern two sources of attractiveness bias, namely preference based and belief based, by using the dynamics of bias measured over 15 years of post-MBA careers. This is significant because belief bias may be minimized by initially providing rich performance-related information. However, preference-based biases are much harder to remove. The media has highlighted sexual harassment across industries. Such pure taste-based, perverse preference cannot simply be eliminated by small policy changes. It rather requires a social and cultural change. Second, our findings are based on one of the largest longitudinal, archival data sample which provides external validity to the findings of prior experimental, small-sample studies. The study spans 25 years of MBA graduates between 1990 and 2015, thus providing credence to a perpetual phenomenon rather than a quirk in a short-lived setting. Third, our attractiveness ratings and job ranking procedures have wide applicability in any labor economics study on appearances. To our knowledge, this is the first attempt in business research at visual image morphing driven by deep learning. Unlike Computer Science literature, we use these unstructured data sources in modeling a real outcome variable (career success) rather than a pure prediction problem.

2. Literature Review and Theory Development

Our study connects the following streams of literature: (i) sources of (gender and race) bias, (ii) existence of similar sources of attractiveness bias and (iii) dynamics in attractiveness bias.

2.1 Sources of Bias

Literature posits two potential sources of discrimination based on extraneous factors, such as race and gender, namely *preference-based* (Becker 1957) and *belief-based* (Fang and Moro 2011, Phelps 1972). Preference-based discrimination is also known as taste-based discrimination. In preference-based bias, the evaluators have an inherent taste (or distaste) for the subjects' characteristics. In the context of attractiveness, this could be a preference for spending time socially around attractive individuals. Though performance on the job provides informative signals of an employee's ability, it will have no impact on changing the preference of the evaluator. These tastes persist because the evaluators continue to have a preference for social, marital or romantic relationships with attractive individuals.

Belief-based discrimination is also known as statistical discrimination. Belief-based bias occurs in an environment of imperfect information where evaluators rely on group-level beliefs to judge individuals. The evaluator starts with a prior based on group-level beliefs on the ability of individuals. Attractive looks are often associated with health, social skills and positive personality traits (Dion et al. 1972). Observing an individual's performance provides a signal of his or her quality to the evaluator who then relies less on the group-level priors. However, if the evaluators do not have a long duration to judge the subject, the priors play a dominant role. Belief bias may be minimized by the initial provision of rich performance-related information or by allowing longer interaction times. However, preference-based biases are much harder to remove.

2.2 Attractiveness Bias

While a majority of Marketing and Economics attractiveness bias research[3] (Table 1) constructs controlled settings to isolate and measure bias size, they do provide suggestions on potential sources that are likely factors in their specific settings. A majority of the studies in the literature point to beliefs as the underlying source of bias. Dion et al. (1972) suggest a positive perception of social skills, mental health and personality from attractive looks. Olivola & Todorov (2010) show the perception of competence to be correlated with attractiveness. These qualities align well with the requirements of the labor market. Not surprisingly, a majority of this research finds positive premiums from attractiveness. Mobius and Rosenblat (2005) show that in an experimental labor market, employers consider attractive workers as more able, even with equal performance measures. Similar outcomes are shown in a public goods experimental setup (Andreoni and Petrie 2006, Rezlescu et al. 2012). Settings where belief bias is shown as a penalty do exist as well (Heilman and Saruwatari 1979). First, young, attractive individuals (especially

[3] refer to Olivola and Todorov 2017 for a literature review on this topic

young women) may sometimes be wrongly perceived to have advanced by relying on their looks. Second, a perception of a competent personality trait may result in less empathy and in the offering of less help to attractive individuals (Cryder et al. 2017, Fisher and Ma 2014).

Table 1: Prior literature on attractiveness bias

Method	**Effect**	**Context**	
Experimental	Positive	Desirability, Competence	Dion et al. 1972
Experimental	Positive	Plaintiff, Defendant	Kulka & Kessler 1978
Experimental	Negative for women in managerial.	Insurance Company Jobs	Heilman & Sauwatari 1979
Experimental	Positive	Salesperson	Reingen & Kernan 1993
Archival	Positive	Labor Market	Hamermesh & Biddle 1993
Archival	Positive	Lawyers	Biddle & Hamermesh 1998
Experimental	Positive	Ultimatum Game	Solnick & Schweitzer 1999
Meta-Analysis	Positive		Langlois et al. 2000
Experimental	Positive	Maze Solving	Mobius & Rosenblat 2006
Field Exp.	Negative for same sex	Hiring	Luxen & Van De Vijver 2006
Experimental	Positive (negative with information)	Public Goods Game	Andreoni & Petrie 2003
Experimental	Positive	Elections	Olivola & Todorov 2008
Field Exp.	Negative for same sex	Hiring	Agthe et al. 2008
Archival	Positive	Elections	Berggren et al. 2010
Experimental	Positive	Economic Exchange Game	Rezlescu et al. 2012
Experimental	Negative	Charity	Fisher & Ma 2014
Experimental	Negative for women	Hiring	Ruffle & Shtudiner 2014
Experimental	Positive	Retail Managers	Fruhen et al. 2015
Experimental	Negative with deliberation	Charity	Cryder et al. 2017

Some of the attractiveness bias literature also highlights preference-based sources. Luxen and Vijver (2006) show a preference for attractive individuals, especially when there is high expectation of continued contact post-experiment. In addition to the academic literature, instances of widespread sexual harassment in the workplace have recently emerged in the media. These examples provide more than sufficient evidence for the existence of purely taste-based decision-making by employers. Once again, a small handful of literature shows a penalty from taste preferences as well. Agthe et al. (2010) show a penalty for attractive scholarship candidates when judged by same sex evaluators. Ruffle and Shtudiner (2014) suggest envy and jealousy can engender a penalty for attractive young women when evaluated by other young women.

Most of these prior studies involved experimental setups that differ from an actual professional work environment in the following three key ways: (1) the evaluator has little knowledge of past performance, (2) the evaluator has limited time to learn about the subject on the job, and (3) (in most cases) the evaluator does not expect to spend time with the employee after the end of the experiment. The first two naturally mean that experimental setups focus on belief-based bias. The evaluator must decide based on priors or group level beliefs for attractive individuals. The third contrast means that pure taste or preference-based bias is minimized in experiments. The hypothetical manager in an experiment has no expectation of spending time with the subject by hiring them or promoting them. These experiments successfully control for numerous interfering factors in order to isolate a single causal factor. They are able to inform us of impact and direction of bias in specific settings. Over the duration of one year of employment, an individual may go through numerous such interactions where an evaluator's beliefs and preference aid or penalize them. Thus, all these factors are thrown in simultaneously, making an overall effect direction and size hard to anticipate.

There are only a few observational data studies on attractiveness bias. Most establish a positive career impact. This suggests that belief and preference biases combine to create a positive premium. Unfortunately, the existing studies do not clarify the actual positive or negative role of both sources independently. The observational literature is unable to answer this because it does not study the dynamics in attractiveness bias. Biddle and Hamermesh (1998) do show a premium in 5-year and 15-year careers, but they do not model dynamics in career progress or bias. Therefore, they do not compare the size of early versus late bias or discern different sources. Our observational data and unstructured data processing allows us to exploit fine-grained career dynamics and provides external validity through the use of a large sample.

2.3 Attractiveness Bias Dynamics

There are sporadic examples of short-run dynamics simulated in experiments. Rezlescu et al. (2012) show a premium in an economic exchange game, where attractive individuals are perceived to be trustworthy. This attractiveness premium diminishes from 42% to 6% after objective signals are revealed. Andreoni & Petrie (2003) show an attractiveness premium in a public goods experiment, where attractive individuals are perceived to be cooperative. This perception and the resulting attractiveness premium vanish in their experiment when objective information on the contribution to the public good is revealed. These are both examples where belief bias is overcome across multiple runs of the experiment. Preference does not have a strong role to play because subjects don't expect a long-term association with other attractive subjects. Nevertheless, we can draw insights from such settings.

We suggest that the dynamics of attractiveness bias can be used to distinguish the sources and direction of the bias. The preference- or taste-based bias is unaffected by any information regarding the individual's underlying quality. In contrast, the belief-based bias disappears as evaluators receive information regarding the individual's underlying quality. Early career decisions resemble a scenario where the employer has limited information and therefore may unwittingly absorb perceptions from appearance cues. In professional careers, performance signals are of course not objective indicators. Nevertheless, we expect a period of 4-8 years to be a more than sufficient time duration for prior beliefs to be overcome by actual performance signals. Belief-based attractiveness rewards should vanish after this point. We use this critical difference between preference and belief bias to discern their roles.

First, we measure attractiveness bias in the later part of one's career. If a positive bias (does not) exists (exist), it would imply the presence (absence) of a positive preference bias. Next, we measure the attractiveness bias in the early part of one's career. If the preference bias was absent, any early career bias points to the existence of a belief bias. If a positive preference bias was present, determining the belief bias is non-trivial. One of the following three possibilities exist: (1) positive belief bias, if the early career bias is greater than the late career bias, (2) zero belief bias, if the early career bias is equal to the late career bias, (3) small negative belief bias, if the early career bias is less than the late career bias. Thus, we can use the relative size of the early and late career bias to answer the following three open questions: (i) Is preference bias absent, positive or negative? (ii) Is belief bias absent, positive or negative? (iii) What is the combined effect when preference and belief bias coexist early in professional careers? We ignore possibilities where one or both of the biases are significantly negative, resulting in an attractiveness penalty over the entire career. This would go against all prior observational data research.

2.4 Image and Text Analysis

Gan et al. (2014) proposes self-taught deep architecture for attractiveness ratings. Xu et al. (2017) develop psychologically inspired convolutional neural network (PI-CNN). Liang et al. (2017) build a three-stage architecture that explicitly accounts for face region extraction and face rotation invariance. Gao et al (2018) use a simple CNN but jointly train on attractiveness prediction as well as face landmark labels. These papers have been developed on top of Gan et al. (2014)'s broad approach; we do the same. We use OpenFace Convolution Neural Network (CNN), developed by Amos et al. (2016), to extract features. Then we use Conditional Adversarial Auto Encoder (CAAE), developed by Zhang et al. (2017), to generate heterogeneous attractiveness evolution paths from a single image. With regards to unstructured profile data, we build on Page Rank (Page et al. 1999) to order unstructured, self-reported, free-text job titles by their desirability. Our work adds to this burgeoning marketing and management literature using unstructured text and image (Lu et al. 2016, Zhang and Luo 2018, Netzer et al. 2012, Dzyabura et al. 2018, Liu et al. 2017, Zhang et al. 2018).

3. Data

We use one of the largest professional social networks as our primary data source. This online platform is used by professionals and job seekers to exhibit their curriculum vitae and by employers to post jobs. A typical user curriculum vitae describes work experience and education alongside a personal photo. Members on this platform typically make connections with each other that may represent real world associations. While members self-report their professional information, the social nature of the platform ensures the veracity of the data. This social network provides a profile search functionality that allows searches based on free text as well as on filters by location, industry and the university attended by professionals. We specifically focus on MBA graduates in the United States. This criterion ensures that the individuals in the sample are comparable because of the similarity in their educational background and their professional goals. While these graduates may all desire swift career progression, we do account for peculiarities of job domains. We use the search functionality offered by this professional social network platform to collect profiles. We use “MBA” as a search string along with filters by location (”United States”) and school name. The list of university names is derived from the US News business school rankings.[4] Figure 1 shows the number of profiles collected by MBA school rank.

Table 2: Number of MBA profiles with profile picture and key career history

Reasons	Profiles
Total Profiles Collected	98,868
Profile picture available	82,477
MBA graduation between [1990,2015] and other key covariates available	**43,533**

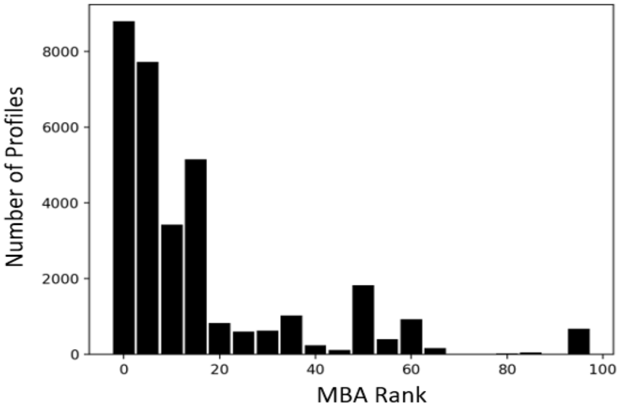


Figure 1: Number of profiles collected across the top 100 ranked MBA programs. Profiles for top-ranked MBA graduates are easier to collect, as individual are more likely to prominently highlight their school name on their profiles.

Unfortunately, the online platform places a high cost on collection and places severe limits on the number of profiles that can be viewed per day. The platform search mechanism is relatively opaque. The search for profiles for a school often returns individuals who were never MBA graduates at the target school. These individuals may have served in teaching, administrative or other capacities in the specific MBA program or in a similar-sounding MBA program. Due to these limitations, we are able to collect a total of 43,533 MBA graduate profiles with profile pictures.

[4] A typical MBA program graduates anywhere between 200 to 1000 students per year, or 5000 to 25000 graduates in the last 25 years. However, the search functionality returns only 100 pages with 10 profiles per page of search results. To cover more profiles, we make use of a platform quirk. We perform up to 10 searches per school by using a different filter for industry each time. This theoretically allows us to capture up to 10,000 profiles per school search. Appendix A.1 provides a breakdown of the collected profiles by school.

Some profiles will always be omitted from our collected sample for the following reasons: (a) some graduates may have exited the job market prior to 2017 and thus never created or deleted their profiles, (b) some may be in the job market but do not use this professional social network, (c) some use this social network but keep their profiles private, and (d) others may have an active public profile but do not mention their MBA degree at all. All such individuals are systematically left out of our sample. To be robust, we collect a secondary data set[5] where some of this systematic selections are observed. We realize that the access to complete data is unrealistic in this domain (on or outside this platform), as a large number of professionals choose to keep their profiles completely or partially private. Beyond a certain limit, as researchers, we are cognizant of not using increasingly intrusive technical methods to gather sensitive data.

Table 3: Sample values for each of the collected profile variables.

Variables	**Sample Values**
Title	Software Developer, Consultant, Intern, CEO.
Employer Name	McKinsey & Company, JPMorgan Chase
Current Industry	Financial Services, Management Consulting, Computer Software
University Name	Harvard University, Stanford University.
Degree Name	B.S., B.A., MBA, MFE
Year From	2009 or January 18, 2007
Year To	2015 or December 29, 2016

Table 3 shows sample values for key profile fields fetched by us. We typically find 2 to 8 career steps, with each step comprising the job title, employer name and industry. We also find 2 to 4 degrees (e.g., high school, undergraduate, Masters, MBA, PhD), each comprising a degree and university name. All of these have a corresponding start year and end year. We identify the attendance at an MBA program using the degree field. Figure 2 shows the number of profiles collected corresponding to each MBA graduation year.

[5] We access the school directory of a top 10 MBA program in the US. This gives us access to names and the graduating year for all 15,640 students between 1990 and 2016. Of course, we do not find professional profiles for all these 15,640 graduates, but now we have visibility into the characteristics of the graduates that have been selected out. In Appendix A.11.3, we describe this secondary data and the robustness tests that were conducted by using this smaller dataset.

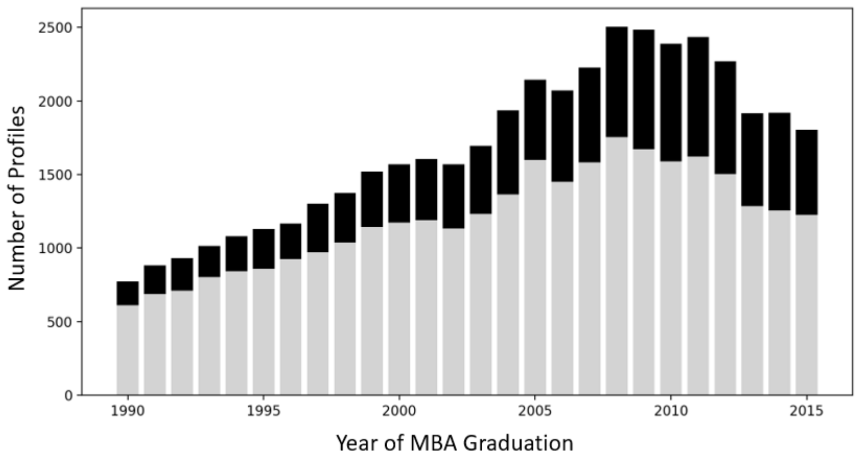


Figure 2: Number of male (light shaded) and female (dark shaded) profiles collected for each graduating class.

4. Calculated Features

Because the profile details are self-reported in free-text fields, we perform significant data cleanup to standardize the names of universities, employers, degrees and job titles. This cleanup ranges from correcting misspellings to matching synonymous names. In addition to the features directly available on the profiles, we also need to calculate new features that will be used in our primary analysis to compare the professional success of attractive and plain-looking individuals while controlling for other characteristics. Table 4 provides a list of all features, their source (or calculation) and sample values. Microsoft Azure Cognitive Face API and Face++ API largely cover face image related features. While these API's do not disclose exact methodology, we believe that they use supervised Deep Convolutional Neural Networks to tag these face features. Website usnews.com is used for education related features such as university names, rankings and fees. Note that some features are sourced from unverified or noisy sources (e.g., ethnicity, age), in which case we perform a sense check by extracting from multiple sources. In Appendix A.6, we discuss calculations for all the features in detail.

Our research goal is to compare career outcome for individuals that differ in attractiveness. In Section 4.1, 4.2 and 4.3 we describe our methodology to assign every individual i with a job rank $rank_{i,t}$ and attractiveness $beauty_{i,t}$ at t years after their MBA graduation. The job rank measures desirability of job attained by the individual e.g., CEO is more desirable than Analyst.

Table 4: Calculated Features, their sources and sample values.

Category	**Feature**	**Source or Calculation**	**Range or Sample Values**
Outcome $r_{i,t}$	Job Rank ($rank_{i,t}$)	Section 4.3	[-9.23,-5.58]
	Std. Job Rank ($r_{i,t}$)	Standardized among peers at t	$\mu(r_{i,t}) = 0, \sigma(r_{i,t}) = 1$
Treatment $b_{i,t}$	Attractiveness Rating ($beauty_{i,t}$)	Sec 4.1, 4.2	[1,7]
	Attractiveness Class ($b_{i,t}$)	Binary classification among peers at t	{0,1}
Demographics D_i	Gender	Microsoft Azure Cognitive Face API	Male, Female
	Ethnicity	Face++ API (on picture)	White, Asian, Black, Others
		Text map API (on name)	European, Asian, Others
	Age	2017 – (UG Graduation) + 22	23-60
		Microsoft Azure Cognitive Face API	18-65
Education and Training E_i	UG University	Self reported	
	UG Ranking	US News	1-200
	UG Degree Area	Mapping Table in Appendix A.6.2	Business, Arts, Science, Others
	Pre MBA Training	(MBA Start) – (UG Graduation)	[0,16]
	MBA School	Self reported	
	MBA Ranking	US News	1-100
	MBA Graduation Year	Self reported	1990-2015
Wealth W_i	Undergraduate Fees	US News	[$14,937, $55,506]
	MBA Fees	US News	[$32,328, $65,988]
Job domain $JD_{i,t}$	Industry Category	Mapping Table in Appendix A.6.3	Finance, Consulting, IT, Others
	Job Type	Mapping Table in Appendix A.6.3	Management, Technical, Others
	Large Employer	Binary Classification on frequency of employer	{0,1}
	Large Location	Binary Classification on frequency of location	{0,1}
Photograph Quality PQ_i	Blur	Microsoft Azure Cognitive Face API	[0,1]
	Exposure	Microsoft Azure Cognitive Face API	[0,1]
	Noise	Microsoft Azure Cognitive Face API	[0,1]
	Resolution	Face++ API	[0,1]
	Face Quality	Face++ API	[0,1]
	Roll	Microsoft Azure Cognitive Face API	[0,1]
	Pitch	Microsoft Azure Cognitive Face API	[0,1]
	Yaw	Microsoft Azure Cognitive Face API	[0,1]
Visible Characteristics VC_i	Lip Makeup	Microsoft Azure Cognitive Face API	[0,1]
	Eye Makeup	Microsoft Azure Cognitive Face API	[0,1]
	Beard	Microsoft Azure Cognitive Face API	[0,1]
	Moustache	Microsoft Azure Cognitive Face API	[0,1]
	Sideburns	Microsoft Azure Cognitive Face API	[0,1]
Invisible Characteristics IC_i	Hair color	Microsoft Azure Cognitive Face API	[0,1]
	Earrings	Microsoft Azure Cognitive Face API	[0,1]
	Necklace	Microsoft Azure Cognitive Face API	[0,1]
	Other accessories	Microsoft Azure Cognitive Face API	[0,1]
	Background	Microsoft Azure Cognitive Face API	Formal, Informal, None
	Clothing	Microsoft Azure Cognitive Face API	Formal, Informal, None
Health H_i	Skin Health	Face++ API	[0,1]
	Skin stain	Face++ API	[0,1]
	Dark Circles	Face++ API	[0,1]
	Eye Glasses	Microsoft Azure Cognitive Face API	[0,1]

4.1 Attractiveness

We focus only on attractiveness of facial appearance. Given the professional nature of the profiles, people often post clear head shots. We detect and remove profiles where pictures either do not capture a face or capture more than one face. Most studies in the literature measure attractiveness as an average of ratings by a group of human raters (Biddle and Hamermesh 1998). Assuming at least 5 human raters per image, it would require approximately 200,000 ratings (43,533 profiles * 5 ratings) for our dataset. In order to scale our analysis, we want to build an attractiveness prediction model.

Golden ratios and neoclassical canons (Schmid et al. 2008) were used in past for machine calculated attractiveness ratings. This was superseded by Machine Learning methods, e.g., Eigen faces (Turk and Pentland 1991), that disassociate feature extraction using PCA and attractiveness label prediction using supervised learning on PCA features. This has been superseded by Deep Learning methods. Gan et al. (2014) proposed self-taught deep architecture to extract features and SVM to map these features to attractiveness ratings. We broadly follow this approach. Contemporaneously, numerous improvements on this broad approach have come up (Xu et al. 2017, Liang et al. 2017, Gao et al. 2018). We follow the evaluation metric used in this literature and contrast our performance with these contemporary papers.

Labeled dataset: We get human raters to judge the attractiveness of a random set of 659 profile pictures. This experiment was executed on the AMT (Amazon Mechanical Turk) platform. The raters were selected to participate based on the following three criteria: 1) geographical location in United States, 2) completion of at least 500 human intelligence tasks (HITs) in the past on the AMT platform and 3) the receipt of approvals on 95% or higher of all their completed HITs. These conditions were applied to ensure that the raters understood our requirements in English, that they were experienced in using the AMT platform and that they have showcased good quality of work in the past. Additionally, limiting the raters to the United States is likely to help in mimicking how a subject's appearance would be judged in a typical US professional work environment. Our training data consists of images labeled as attractive based on a 1 to 7 scale, where 1 represents the lowest value of attractiveness and 7 represents the highest value of attractiveness.Table 5 provides the mean and standard deviation of the attractiveness ratings, with women being rated on average higher than men. Each image is labeled by 5 coders, and we use the average of the five ratings for an image as its true rating. Attractiveness is a subjective measurement, and multiple raters are unlikely to exactly agree in their ratings. We calculate a standardized Cronbach alpha (Santos 1999) value of 0.73, which is a typical psychometric standard used for consensus between the raters. The relatively high value (0.7 typically being an acceptable standard) gives us confidence that using an average would capture at least a coarse variance in the attractiveness of the profile pictures. These ratings would next act as training labels for our machine learning model.

Table 5: Mean and standard deviation of attractive ratings

Gender	Rating Mean	Rating Std. Dev.
Male	4.02	0.88
Female	4.46	1.06

Image Features: We need to ensure that the machine learning model picks up unalterable appearance differences instead of alterable differences such as hairstyle, clothing and photograph quality. To ensure this, we first use the dlib library (dlib.net/python) to obtain a bounding box for the face on the profile picture. This removes picture background, clothing and hairstyle, from the images. Next, we normalize color histogram, overall pixel intensity values and resolution across images. We standardize the head pose by using a 2-D affine transformation (Zhang and Gao 2009) to align the head rotation in three dimensions, i.e., roll, tilt and yaw. Once front-facing comparable face images are obtained, we compress a typical 400x400x3 (RGB) image into a handful of key features that capture a majority of differences between appearances. We use a deep neural network implementation by the Open Face project (Amos et al. 2016). This architecture, while trained for face recognition task, provides a 128-dimensional intermediate layer. This layer represents a low-dimensional embedding of any face image. Figure 3 lays out a summary of various transformations performed on profile pictures to arrive at a set of 128 features for every face. Using features otherwise used for face recognition has a side benefit of ignoring superficial temporary characteristics, e.g., room lighting, facial hair, expressions, etc. This is because face recognition deep architectures are fine tuned to recognize the same face across different settings. The 128-Dim features learn to ignore superficial temporary characteristics.

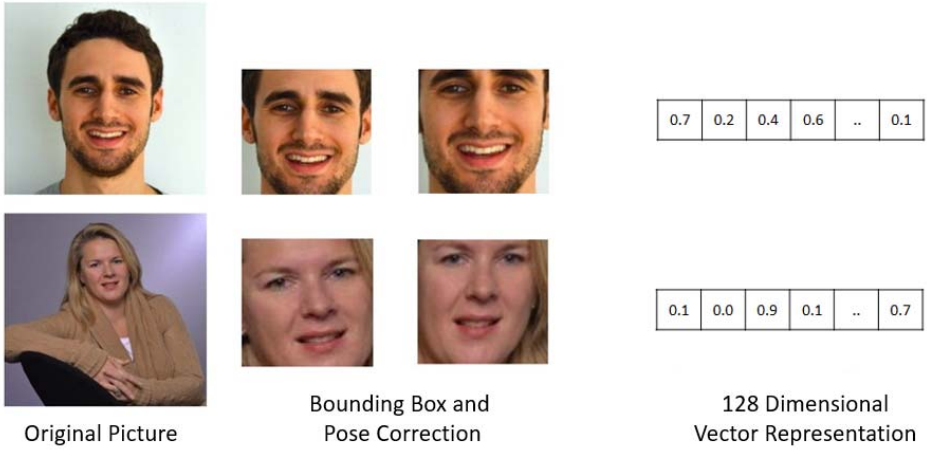


Figure 3: For two samples, the figure shows (left to right) the original picture, the bounded picture, the pose corrected versions and the 128-dimensional feature representation.

Predictive Model: We train a support vector regression model that learns the relationship between 128 image features and the attractiveness label. We validate model performance by holding back some of the labeled images from the training step and evaluating the model performance on this held-out set. Our predicted ratings are off from the actual average human rating by a rmse (root mean squared error) of 0.81 on a 1 to 7 scale. Attractiveness prediction literature (Gan et al. 2014, Xu et al. 2017, Liang et al. 2017, Gao et al. 2018) evaluates model prediction performance using a variety of metrics such as mean squared error, Pearson correlation, adjusted R-squared, accuracy and ROC-AUC. We provide more details on all relevant evaluation metrics in Appendix A.2 to compare the selected model against alternative hyper parameters and alternative models (Support Vector Classification, Random Forest, Logistic Regression, K-NN, Adaboost, Gaussian Naïve Bayes and Quadratic Discriminant Analysis). There is also a practical concern that the model has learnt to pick up on some unrelated correlated features e.g., attractive individuals always post higher quality photographs. So, the training and validation performance have not picked up on causal factors of attractive looks, but rather alterable features chosen by individuals on their self-posted profile pictures. In Appendix A.2, we provide details on model evaluation on test dataset, which is drawn from experimental settings where individuals have no control on their face images.

4.2 Attractiveness Evolution

Model

Our research question requires us to measure an individual's attractiveness over their career. However, our primary profile picture dataset has a single picture per person. Thus, we need a model to project attractiveness evolution over age. We have access to two secondary datasets to guide our attractiveness evolution approach. The Cross Age Celebrity Dataset (CACD) contains 104,595 images for 1282 celebrities between age of 25 and 60. The FGNET dataset contains 217 images for 42 subjects between age of 21 and 34. Unlike our profile pictures, these datasets contain multiple pictures for a subject. Figure 4 shows that the celebrity photographs in CACD dataset tend to be more attractive on average. However, all three datasets exhibit a similar declining evolution of attractiveness over age. Figure 4 also explores the heterogeneity in attractiveness evolution among four celebrities within the CACD dataset. As an example, Claire Danes has a higher attractiveness rating but declines more rapidly with age compared to Anthony Mackie. More generally, if person A is 1 rating point higher than person B at our data collection in 2017, it does not necessarily mean that this gap was preserved at their MBA graduation date. This heterogeneity is also supported by past studies on the evolution of looks over age. Zhang et al. (2017) suggest that looks evolve over a complex manifold, thus taking unique paths for every individual. Following these examples, the equation 1 models the attractiveness of individual i at age a. At the population level, attractiveness evolves linearly over age (λ_1). Individuals are heterogeneous in attractiveness evolution (ρ_i^1).

$$beauty_{i,age} = \lambda_0 + (\lambda_1 + \rho_i^1) * age + \rho_i^0 + e_{i,a} \qquad (1)$$

Estimation

If attractiveness evolution were homogenous ($\rho_i^0 = 0$), then estimating ($\lambda_0, \lambda_1, \rho_i^0$) would be relatively trivial using one profile picture per individual in our dataset. The heterogeneous

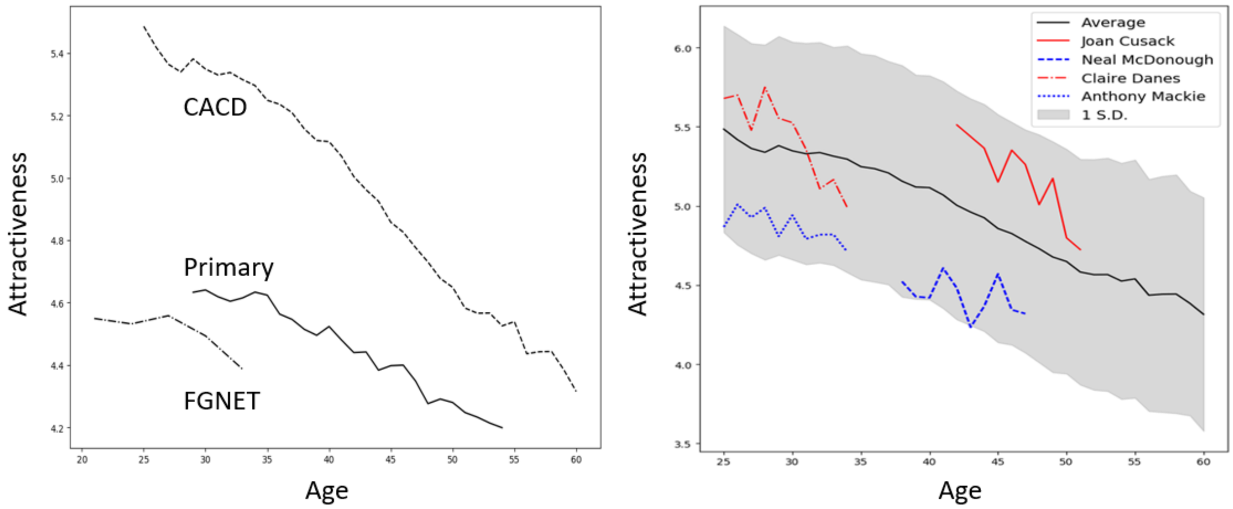


Figure 4: (Left) Average attractiveness of subjects at different age for three datasets – CACD, FGNET and our primary professional social network profile pictures. (Right) Attractiveness evolution of four celebrities compared against the average attractiveness evolution of CACD dataset (± 1 std. dev.).

attractive evolution $\rho_{1,i}$ can only be estimated if we have multiple observations per individual. Zhang et al. (2017) develop a CAAE model that uses a single picture to generate multiple photo-realistic morphs at sequence of age milestones. These morphs are expected to match the individual's looks at those age milestones. We use a pre-trained CAAE model, using multiple pictures corresponding to different ages of the same individual. This data is completely independent of our profile pictures dataset. The training data for the CAAE model is limited; it cannot learn a unique evolution path for every single face. It must generalize. Therefore, it tries to learn evolution paths for a large number of representative faces seen during training. For a new test face not seen during training, where only a single profile picture is available, the model predicts an evolution path. This is based on evolution seen for training faces that were similar to the new test face. It does so by creating a low-dimensional representation for an input face picture such that similar faces are close together in this low-dimensional space. We provide more detailed intuition for the CAAE architecture in Appendix A.3.

Figure 5 depicts an example we are able to generate from this model. It is particularly interesting to note that the model learns to ignore alterable face features (e.g., hairstyle, eye glasses). This happens because, at different ages, easily alterable features vary across images. For example, during training, the model uses pictures that may show a clean-shaven individual at the age of

25 or a bearded individual at the age of 30. In the beginning, model may try to extrapolate, i.e., it may try to show all evolutions from 25 to 30 as an individual with a beard. Over training iterations, it learns that such an extrapolation is inaccurate for the vast majority of faces. Therefore, it drops information pertaining to beards from its low-dimensional representation of an input image. Thus, a reconstruction of an input image with a beard, even when the reconstruction is one representing the individual at the same age, ignores the beard. This provides a side benefit in line with our research objective.

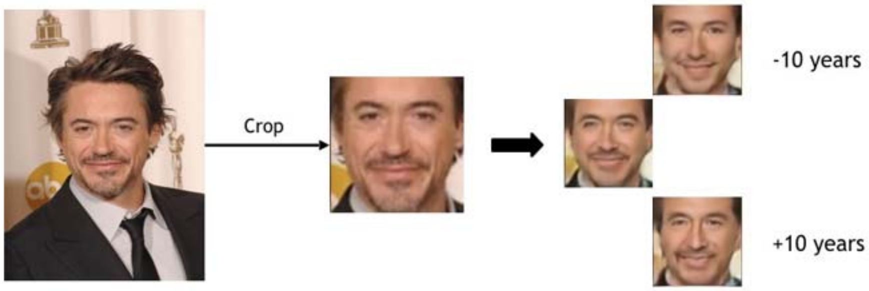


Figure 5: (from left to right) The original face image, the aligned version, and the generated (reconstructed) images at the current age, 10 years prior to the current age and 10 years past the current age. Note how facial hair is minimized even on the current age picture.

$$beauty^*_{i,age} = beauty_{i,age} + \eta_{i,age} \tag{2}$$

Applying the attractiveness prediction model on the CAAE morphed pictures gives us a sequence of attractiveness predictions $beauty^*_{i,age}$. Owing to unknown properties of error $\eta_{i,age}$ in $beauty^*_{i,a}$, we choose not to use these predictions directly. Instead, we use a regularized ridge regression for estimation of equation 1[6]. The regularization is meant to penalize (ρ^1_i), i.e., large gap between individual specific evolution and the population level evolution. Without this regularization we would have some individuals following outlandish attractiveness evolution due to one or two bad morphs. Once we have estimates ($\widehat{\lambda_0}, \widehat{\lambda_1}, \widehat{\rho^0_i}, \widehat{\rho^1_i}$), we can predict $\widehat{beauty_{i,age}}$ at any time age for individual i. Note that we have combined image morphing (CAAE), attractiveness prediction (OpenFace + SVM) and regularized heterogeneous evolution (RHE) model (Figure 6). We do so to benefit from the capacity of deep learning models to learn accurate

[6] This implicitly assumes that $\eta_{i,age}$ is uncorrelated with $beauty_{i,age}$

high-dimensional pixel-level distributions while still relying on linear model for interpretable, unbiased and low variance heterogeneity estimates.

Table 6: Ridge Regression Results separately for men and women. We discard the quadratic model since λ_2 turns out to be statistically insignificant.

	(1) Male	(2) Male	(3) Female	(4) Female
Constant	4.593**	**4.756****	5.576**	**6.076****
	(0.148)	(0.024)	(0.281)	(0.045)
Age	-0.0036	**-0.0117****	-0.0055	**-0.031****
	(0.007)	(0.001)	(0.014)	(0.001)
Age^2	-0.0001		-0.0003	
	(0.0001)		(0.000)	
Obs.	16,216	16,216	6,022	6,022
R-squared	0.024	0.024	0.108	0.107
F-statistic	197.6	393.9	362.9	722.4

Standard errors are in parenthesis
** p<0.01, * p<0.05

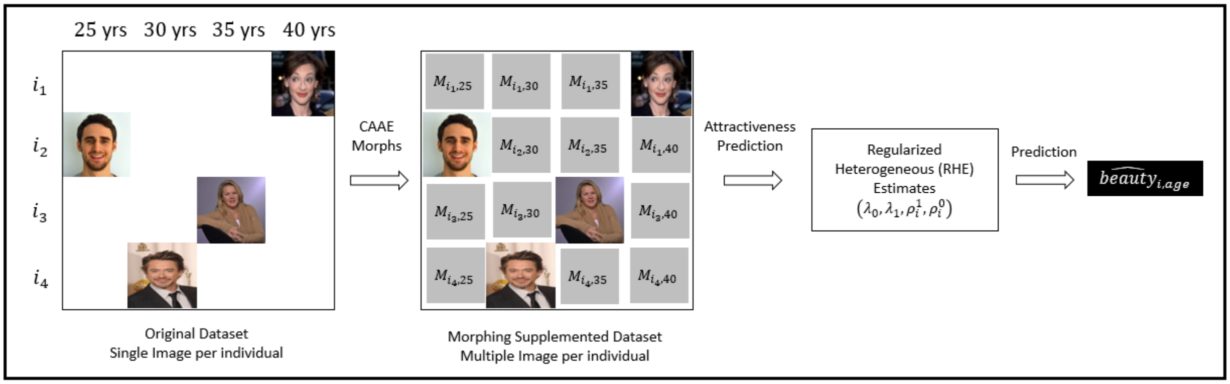


Figure 6: We first create morphed picture $M_{i,age}$ at all age milestones where actual picture is not available for individual i. Then RHE model is estimated using predicted attractiveness of morhped pictures.

Testing

Before we use these predictions $\widehat{beauty}_{\iota,age}$ in our econometric model, its important to quantify the accuracy of this attractiveness evolution approach. Zhang et al. (2017) perform a survey to validate CAAE face morphs with greater than 20-year age gap between available ground-truth pictures and target morph age. A total of 48.4% of the respondents indicate that the generated face image depicts the same person as the ground truth image, 29.6% of the respondents indicate that it does not, and the rest of the respondents are not sure. The CAAE morphs are just one component in our attractiveness evolution approach. We formally test our attractiveness evolution approach on the CACD dataset. We randomly sample a single image for every celebrity. This acts as the training sample, which is akin to our professional social network setting with one picture per individual. We compare the predicted attractiveness $\widehat{beauty}_{i,age}$ with the actual attractiveness $beauty_{i,age}$. We attain a root mean square error of 0.245 on an attractivness scale of 1 to 7. In Appendix A.4, we elaborate further and show performance comparison against alternative models. In Appendix A.10.2, we also report robustness of primary econometric model with various alternative attractiveness evolution models. Readers should note that our primary econometric model is robust even if we simply assume that attractiveness does not evolve at all. This strongly suggests that the main findings are not driven by measurement errors in this admittedly cumbersome attractiveness evolution approach, nor by over simplification of attractiveness evolution.

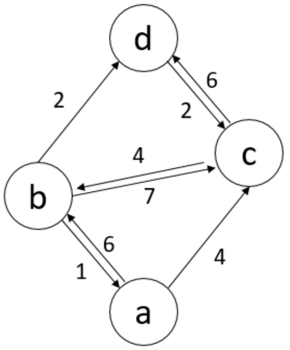
4.3 Job Rank

Any study on career progression in professional roles needs a preference order of jobs. We will build a measure that allows us to compare an individual's job, say 5 years after MBA graduation, with that of his or her peers, i.e., the jobs of everyone at 5 years after their MBA graduation. We utilize the approach suggested by Baker et al. (1994) and Gayle et al. (2012) that relies on observed job switches to establish a preference order. This formulation presumes that a large number of transitions from job A to job B compared to the number of backward transitions imply that job B is more desirable than job A and is therefore ranked higher. This approach works well with a handful of jobs, which is the case for the executive roles (CEO, CFO, principal consultant, board member) studied by Baker et al. (1994) and Gayle et al. (2012). However, the majority of milestones in our data corresponds to early and mid-career jobs in significantly dissimilar firms and industries; therefore, fewer pairwise transitions are observed. We observe more than 10,000 distinct jobs in our data because of the unstructured nature of the profiles. A majority of these jobs appear in 5 or fewer profiles and therefore do not afford enough information on pairwise transitions for us to accurately infer their preference order. Therefore, we extend the approach of Baker et al. (1994) by (i) building a preference ranking for the 1000 most frequent jobs and then (ii) approximating the ranking for the remaining jobs, based on text similarity with the 1000 ranked frequent jobs.

We define a job as a combination of title, employer size and employer industry. This ensures that positions, such as “CEO at very small internet” company and “CEO at large finance” company, are not ranked together simply because the title “CEO” remains the same in otherwise dissimilar jobs. The former is likely a position at a 5-person startup company, while the latter may be a position leading thousands of employees in a multi-billion dollar company. Once we have a set of jobs that occur relatively frequently in our dataset, we count all the directed transitions between any two pairs of jobs. The matrix M represents these transitions. We initialize all jobs to have equal rank (r_0) and then run the PageRank algorithm (Page et al. 1999) to arrive at a differentiated rank score for each job (*r*).

The Page Rank algorithm was originally used to determine the web page’s importance based on its in-links and out-links. In our case, jobs are equivalent to web pages, and links are equivalent to job transitions. A top position is unlikely to have the largest number of incoming transitions; instead, it receives incoming transitions from relatively senior jobs (e.g., CFO to CEO, Principal to CEO). This algorithm iteratively propagates rank along the directed edges, therefore eventually transferring weight from a junior position (intern, analyst) to top positions (CEO, board member). The rank (*r*) corresponds to the likelihood of ending up in a given job if an individual spends an infinite time in the job market switching between the jobs, with the probability of each step drawn from the transitions matrix (*M*).

Next, we deal with the less frequent jobs. As an example, “principal consultant at a large consulting firm” and “software developer at a large IT firm” are two common jobs that can be used to infer the ranking for positions, such as the “principal consultant for the energy-oil industry at a large consulting firm” and the “Agile software developer at a large IT firm,” respectively. We convert job strings into a tf-idf (Mihalcia 2006) representation. We chose 2000



$$\underbrace{\left[\begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 0.6 & 0.4 & 0 \\ b & 0.1 & 0 & 0.7 & 0.2 \\ c & 0 & 0.4 & 0 & 0.6 \\ d & 0 & 0 & 0.2 & 0.8 \end{array}\right]^{n}}_{\boldsymbol{M^n}} \times \underbrace{\begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}}_{\boldsymbol{r_0}} = \underbrace{\begin{bmatrix} 0.0086 \\ 0.087 \\ 0.2 \\ 0.69 \end{bmatrix}}_{\boldsymbol{r}}$$

Figure 7: (Left) Job switches observed between a set of four jobs (a, b, c, d). The number next to every edge represents the number of job switches. (Right) The resulting matrix M, the initial ranks r_0 and the final ranks r are shown.

as the size of the vocabulary, therefore describing every job as a 2000-dimensional vector. The ith entry for a job vector represents whether the ith vocabulary word occurs (1) or not (0) in this job. This choice of vocabulary size keeps words informative of rank, such as “vice,” “associate,” and “chief,” while eliminating rare words, such as “agile,” “rockstar-coder,” and “energy-oil.” As rare vocabulary words such as “agile” are ignored in the vector notation, the two jobs “software developer at large IT” and “Agile software developer at large IT” are exactly superimposed in this tf-idf vector space. More generally, the tf-idf representation allows us to calculate the distance between any two jobs. We use the weighted k-Nearest Neighbor (Cover et al. 1967, k = 10) to approximate the ranking for the less frequent jobs, based on the 10 nearest-ranked jobs.

Table 7 reports a random sample of jobs and their respective rankings. For every individual, we find the job performed at t years after MBA graduation and the corresponding job rank $rank_{i,t}$. Figure 8 plots $rank_{i,t}$ trend over time averaged across profiles. This shows that our job ranks grow almost linearly and overall appear to be sensible. While this provides some sense check, we need a more rigorous validation for the entire job rank calculation module. We perform **validation** on a held out sample of job switches. We assign job ranks to the two jobs in every switch. Our job rank model is good if a high proportion of job switches go from a lower-ranked job to a higher-ranked job. If the job ranks are uninformative (or random), this proportion will be 0.5. We are able to achieve a ratio of 0.69.

We perform **testing** on externally collected job hierarchy samples from heirarchystructure.com and salary data from salarylist.com. Note that job hierarchies from heirarchystructure.com are manually crafted and limited to a single career path within an industry. Thus it is not very apt for cross industry switches observed in our dataset. Similarly, order by salary captures only one facet of a job's desirability. In-fact empirically observed switches over 100,000 MBA graduate profiles is perhaps a better approximation of average desirability then an industry specific hierarchy or salary data. We expect our desirability measure to be loosely correlated with industry hierarchies or salary data but we do not expect a high degree of match. Therefore we report the correlation purely as a sense check. We are not aware of any large scale (more than 100 jobs) ground truth job desirability dataset. The job hierarchy samples dataset provides 13 ordered job lists. On average, one list orders 61 distinct jobs. The salary dataset provides a single ordered list with 817 jobs. For all 14 lists, we also create an alternative order as predicted by our job rank model. We evaluate our job rank ordering against the original ordering using rank correlations metric (Spearman, Kendall and Pearson). We are able to achieve a spearman correlation of 0.44 and Kendall correlation of 0.36 averaged over all these lists.

Table 7: A small sample of jobs and the respective log of page ranks.

Title	**Company Size**	**Industry Category**	**Log Rank**
Investment Banking Summer Analyst	Large	IT	-8.581
Systems Analyst	Very Small	IT	-8.371
Marketing Coordinator	Small	IT	-8.265
Senior Research Analyst	Medium	Finance	-8.172
Product Director	Small	IT	-7.979
Regional Director	Small	Others	-7.847
VP Finance	Medium	Consulting	-7.611
Chief Revenue Officer	Very Small	Finance	-7.322

We use these validation and testing metrics to select from alternative choices for job definition, ranking procedure, extrapolation procedure and hyper parameters (N = 1000 frequent jobs for ranking, k = 10 neighbors for extrapolation, v = 2000 tf-idf dimensions). In Appendix A.5, we provide more details on these alternative model specifications and resulting evaluation metrics.

5. Empirical Strategy

We first present some model-free evidence and then describe our primary model.

5.1 Model-Free Evidence

We divide individuals in our dataset into attractive and plain-looking groups by using their attractiveness at the time of MBA graduation ($t = 0$). For every individual, we find the job performed at t years after MBA graduation and their corresponding job rank $rank_{i,t}$. Figure 8 presents the average $rank_{i,t}$ for both groups over time. Figure 8 shows that from the beginning, there appears to be a small gap between the attractive and plain-looking groups. The gap grows over time. It is likely that attractiveness bias plays a role, even before the start of the post-MBA career. It may help individuals to gain better education and skills. This cumulative effect of bias in the years prior to MBA graduation will appear on the $rank_{i,0}$, i.e., the first job after MBA graduation. We perform propensity score matching (Appendix A.7) on pre-MBA factors to mitigate this effect. The gap immediately at the start of the post-MBA career vanishes for matched samples. This is reasonable since any significant gap within a single year would likely signal an extraordinary degree of discrimination.

The gap appears to widen gradually for both unmatched and matched cases, developing into a large gap by the end of the career. This second observation can be reasoned in two ways: (i) a positive attractiveness bias exists throughout the career or (ii) the small gap in the early career places attractive individuals in slightly better roles. However, these small early differences have far reaching effects, as they are propelled up faster throughout their career. This may happen because of high-quality on-the-job training, e.g., a McKinsey Consulting junior associate who gets first hand industry-wide experience. To test (ii), we perform propensity score matching on job rank attained at the end of 5 years in addition to pre-MBA factors. In the matched samples, the two groups have the same education, skills, MBA degree and career outcome at 5 years. Figure 9 shows a large gap (albeit smaller than before) still develops between the two groups. This strongly suggests the presence of a preference bias since it’s the only bias source late in the career.

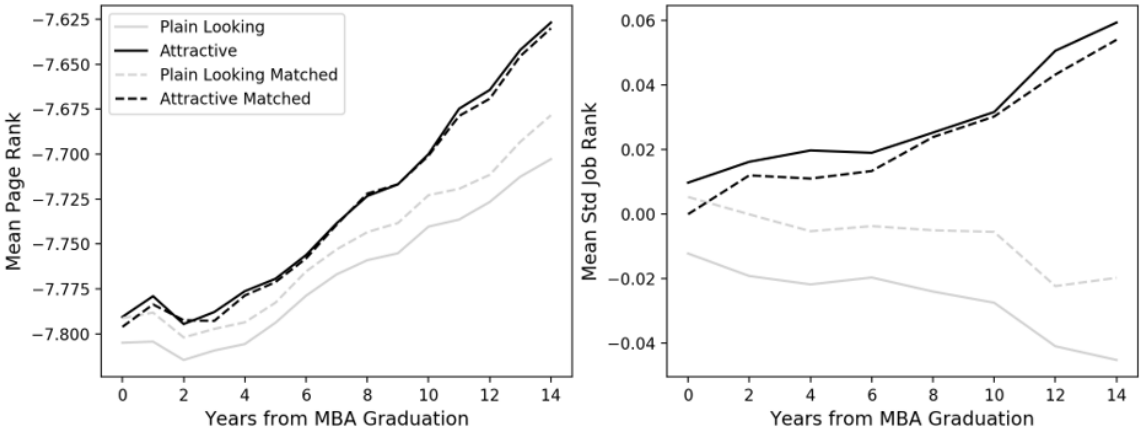


Figure 8: The dark and gray lines are attractive and plain-looking groups respectively. The groups start their career with a gap in the first year after MBA Graduation. The dashed lines are matched on all pre-MBA graduation characteristics. The left figure represents raw job ranks (from page rank), while the right figure represents the proportion of individuals in each group that are successful

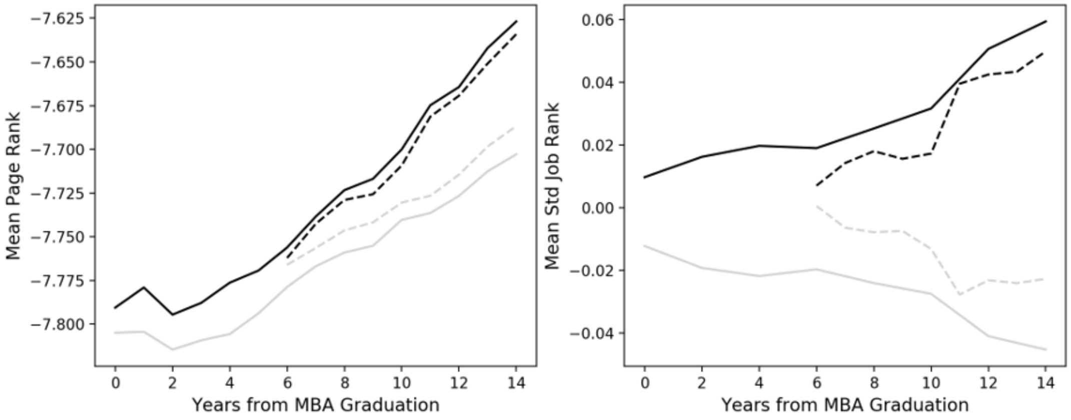


Figure 9: The dark and gray lines are attractive and plain- looking groups, respectively. The groups start their career with a gap in the first year after MBA graduation. The dashed lines are matched on all pre-MBA graduation characteristics as well as job rank attained after 5 years after MBA graduation.

The role of early career belief bias cannot directly be inferred here. A small difference that develops early in a career can be explained by positive preference bias combined with (i) positive belief bias, (ii) zero belief bias, or (iii) small, negative belief bias. The distinction among the three depends of relative size. Next, we develop a full model to precisely measure the effect size.

5.2 Model

Our primary research goal is to measure belief and preference biases. In order to do so, we need to be able to measure career progression between any two milestones (e.g., $t = 2$ years to $t +$

$\Delta t = 3$ years after MBA graduation) for attractive and plain-looking individuals. Then we can compare this attractiveness premium over the same duration in different career phases, e.g., year $2 \rightarrow 3$ vs year $8 \rightarrow 9$. The relative size of the effect sizes would allow us to infer role of beliefs and preferences. If premium in year $8 \rightarrow 9$ is much smaller then $2 \rightarrow 3$, it would indicate belief bias that has been resolved over time. An ideal hypothetical experiment to determine bias or premiums $\beta^{t,t+\Delta t}$ would be to assign attractive or plain looks randomly to a representative sample of professionals at t years and then measure the difference between their average career outcome at the end of $\mathrm{t} + \Delta\mathrm{t}$ years.

$$\beta^{t,t+\Delta t} = E[rank\ at\ t + \Delta t/attractive] - E[rank\ at\ t + \Delta t/plain\ looking] \qquad (3)$$

In this example, attractiveness assignment is a randomized treatment. Naturally, such an experiment is unrealistic. Instead we must rely on a quasi-experiment using observational data. Any quasi-experiment is confounded by the presence of systematic confounders between the two groups. Therefore, we build propensity to be treated model to predict attractiveness assignment from these observed confounders. Controlling for propensity score would be equivalent to quasi-randomized assignment of attractiveness. First, we formally discuss the outcome variable, treatment and the confounders. Then the estimation of belief and preference biases.

Outcome ($\boldsymbol{r_{i,t}}$): We standardize the job ranks for all individuals i whose job transition is observed between $t \rightarrow t + \Delta t$. Standardization is useful because the same job may be considered to be successful in the early career but unsuccessful later. For example, at year 1, a senior associate may be a successful job outcome, while an intern is likely to be an unsuccessful state. By year 15, a senior associate may be an unsuccessful state while a managing partner is likely to be successful. Standardization allows us to compare treatment effect across job periods, e.g., year $2 \rightarrow 3$ vs year $8 \rightarrow 9$. In Appendix A.10.3, we show that our results are robust to alternative measures of job rank beside this standardized measure.

$$r_{i,t} = \frac{rank_{i,t} - \mu_t(rank_{i,t})}{\sigma_t(rank_{i,t})} \tag{4}$$

Treatment ($\boldsymbol{b_{i,t}}$): We divide all individuals i whose job transition is observed between $t \rightarrow t + \Delta t$, into attractive ($b_{i,t} = 1$) and plain-looking ($b_{i,t} = 0$) groups.

$$b_{i,t} = beauty_{i,t} - Median_t\left(beauty_{i,t}\right) \tag{5}$$

Demographics ($\boldsymbol{D_i}$): Individuals' demographics may have two simultaneous effects – (i) attractiveness ratings by human raters and therefore our attractiveness prediction model may systematically rate individuals in certain demographics higher or lower, (ii) individual may face bias in their career because of their demographics. As a result, it would be unclear if career success is the outcome of demographic (gender or ethnicity or ageism) bias or attractiveness bias. Therefore, it is important that attractive and plain looking groups are randomized with respect to these demographics in order to cleanly separate the effect of attractiveness bias. We derive ethnicity[7] both from face image and name. Similarly, we derive age both from face image and undergraduate graduation year. These are inherently noisy guesses. This is not a problem since human attractiveness raters or employers also make their judgments based on noisy guess of individuals' ethnicity and age based on their looks or name. Once matched on D_i, the treated and control groups will have similar gender, similarly ethnic sounding names, similarly ethnic looks, similarly young looks and similar undergraduate graduation year.

$$D_i = (gender_i, \underbrace{faceEthnicity, nameEthnicity}_{ethnicity_i}, \underbrace{faceAge, ugAge}_{age_i})$$

Education ($\boldsymbol{E_i}$): While we want to study attractiveness bias in post-MBA careers, the individuals may not have started on an equal footing. They may have faced bias at admissions to high quality undergraduate or MBA programs. They may also have chosen certain types of educational degrees based on knowledge of their looks, e.g., more attractive individuals may

[7] We do not make any attempt to discern nuanced definitions of "race," "ethnicities" and skin colors. We use the term "ethnicity" throughout the paper to refer to all factors that may cause systematic job market differences.

choose performing arts degrees. Even though these individuals eventually complete the same MBA degree, the average career progression is likely to be significantly different based on quality and type of education. The effects of this pre-MBA stratification is not the focus of our research. Therefore, we randomize attractive and plain-looking groups over education and training. We do so using undergraduate university, undergraduate university rank, undergraduate degree (Science, Arts, Business, Others), length of experience between undergraduate and MBA[8], MBA School and MBA school rank.

$$E_i = (ugSchool_i, ugRank_i, ugDegree_i, mbaSchool_i, mbaRank_i, mbaYear_i,$$
$$preMbaExperience_i, ugFees_i, mbaFees_i)$$

We also randomize the year of MBA graduation. For estimation, we pool together MBA graduates between 1990 and 2015. This potentially leads to the following two concerns: (i) Typical career progression in the 2000s may have changed significantly due to structural shifts in job markets. Employers may have become less hierarchical and more accepting of young, highly talented individuals taking on critical roles early in their career. As a result, the 1990s graduates may have moved more slowly through their early careers. (ii) Even if the structure of the job market is assumed to have remained static, the 1990s graduates are less likely to accurately record their early career progression on this professional social network. Thus, there is a greater likelihood of early graduate profiles having missing career steps. Both (i) and (ii) result in a perception of slower (or at least different) career growth for early graduates (i.e., the 1990s) relative to more recent graduates (i.e., the 2000s). Randomizing years of MBA graduation eliminates these concerns.

It does not play a direct role in productivity, but it could be argued that wealth can be utilized for better education. Education quality is already randomized by features discussed above. It may also be argued that wealth plays an indirect role in career progression by opening up

[8] Thus, we consciously choose length of experience between undergraduate and MBA instead of pre-MBA job rank, because the former is more predictive of post-MBA career success. In our dataset, average job rank over the entire sample grow uniformly from $t = 0$ to 15 years after MBA graduation. However, the average job rank has no upward growth prior to MBA. We interpret this as individuals engaged in career exploration, diversified training and internships, unlike MBA graduates who are focused on growth in their chosen career.

networking opportunities or providing easy access to capital for young individuals to run their own companies. While we do not have evidence for this, let us assume that individuals with wealthier childhoods are permanently more attractive in their adulthoods. Thus, wealth may be driving both career success and attractiveness. We do not have any direct access to individuals' family wealth. We use tuition fees at various undergraduate and MBA universities as an indirect measure. The underlying assumption is that individuals have an array of university choices with similar rankings and there is some natural sorting of wealthier individuals into universities with higher tuition fees.

Job Domain (JD_i): The job attained and thus the individual's state in year $t + \Delta t$ depends on the individual's state in the previous year $r_{i,t}$, which in turn depends on $r_{i,t-1}$ and so on. Thus, career state at year $t + \Delta$t is not simply an outcome of performance or bias in year $t + \Delta t$. Early success (say outcome in year 1, $r_{i,1}$) due to early career performance or bias may have a far-reaching effect. Therefore when studying bias in career phase $t \rightarrow \ t + \Delta t$ we must randomize attractive and plain looking groups over their job rank $r_{i,t}$ at the start of this period. Further, individuals at the same stage of their career and with similar abilities may progress differently because of the typical progression in their domain. $JD_{i,t}$ also captures the individuals' current industry, job type, location size and employer size. As a hypothetical example, a Google software developer may start at $150,000, but growth may be slower. In contrast, a Goldman Sachs investment banker starts with a long training program, producing little, and earning $60,000. However, he or she may have much faster executive growth opportunities.

$$JD_{it} = (r_{i,t}, industry_{it}, jobType_{it}, employerSize_{it}, locationSize_{it}, locationChange_{it})$$

Picture Characteristics: The attractiveness measure ($b_{i,t}$) comes from a predictive machine learning model applied on subjects' self-posted pictures. Ideally, the subject should not be altering their own attractiveness (treatment). There is a plausible risk that some alterable characteristics seep into the attractiveness measure. We divide such characteristics into four categories – (i) Photograph quality PQ_i, (ii) Visible Characteristics VC_i e.g,. makeup within the face area that is cropped in and visible to human raters and our attractiveness prediction module. (iii) Invisible Characteristics IC_i e.g., clothing outside the face area that is cropped out and invisible to our attractiveness prediction module. (iv) Health H_i e.g., dark circles which may be altered cheaply on a photograph using Photoshop or altered in reality over long term costly investment in healthy lifestyle and cosmetic products.

$$PQ_i = (blur_i, exposure_i, noise_i, faceQuality_i, resolution_i)^{9}$$
$$VC_i = (lipMakeUp_i, eyeMakeUp_i, beard_i, moustache_i, sideburns_i, eyeGlasses_i, smile_i)$$
$$IC_i = (bald_i, background_i, clothing_i, Necklace_i, Hat_i, Earrings_i, otherAccessories_i,)$$
$$H_i = (stain_i, health_i, darkCircles_i)$$

Forward-looking individuals may be more strategic, posting outdated pictures of their younger selves. Note that we already randomize age derived from face images $faceAge_i$ as part of demographics D_i. Thus, we are already dealing with this concern by only comparing individuals who look to be of similar age from their pictures. A more tricky argument to deal with would be that forward-looking (or ambitious, prosperous or professional) individuals are habitually or strategically more likely to pay attention and invest in picture characteristics. Thus, a forward-looking attitude may be driving both career success and attractiveness ratings. In Appendix A.6.4 we show that picture characteristics (PQ_i, VC_i, IC_i, H_i) are in fact positively correlated among each other. For example, individuals who wear lipstick are also more likely to dress professionally and to post high resolution pictures.

Fortunately, invisible characteristics (IC_i) do not have any way of seeping into our econometric model because we crop out hair, clothing and background from the photograph. We minimize concerns with remaining picture characteristics (PQ_i, VC_i, H_i) by - (i) Re-orienting (roll, tilt, yaw) the face and standardizing picture quality (pixel value distribution, resolution). (ii) Extracting 128-Dim face features using a face recognition module trained to devalue temporal characteristics. (iii) Image morphing that evolves faces based on average population level evolution paths thus devalue unique temporal features, e.g., unique sideburns or sunglasses.

[9] $noise_i$ refers to random variation of pixel color or brightness caused to image capturing instrument. $faceQuality_i$ refers to how well the profile picture captures the face without occlusion, rotation etc.

All of these are meant to ensure that a subject's choice of profile picture does not change the attractiveness measure. Even with these design elements in place, we further randomize attractive and plain-looking groups on (PQ_i, VC_i, H_i). We would effectively be comparing counterparts who post pictures with similar picture characteristics, so that any systematic picture choices by subjects do not confound our attractiveness bias results.

Note that attractiveness is made up of both un-alterable and alterable face characteristics.[10] Our randomization over alterable characteristics $(faceAge_i, PQ_i, VC_i, H_i)$ must not be interpreted to mean that these factors do not have any impact on common perception of attractiveness or bias in the workplace. Instead we are saying that we cannot identify the relationship between career success and attractiveness from alterable face characteristics using observational data. While we expect that posting professional photographs, makeup, fashionable facial hair, etc. has an impact on career advancement, these effects are easy to manipulate for subjects and thus not the focus of our observation data study. In fact, these effects are actually much easier to ascertain in experimental studies, where researchers could perform randomized assignments of makeup or hairstyle. Such a randomized experiment is not possible for unalterable face structure (except through plastic surgery).

Note that all the features described above attempt to randomize the attractive and plain-looking groups among the available sample of collected profiles. Naturally our sample covers only a fraction of all MBA graduates between 1990-2015. It excludes individuals (i) who left the job market entirely and therefore do not keep an active professional social profile, (ii) who are active in the job market but choose to not keep an active searchable profile or (iii) who keep an active profile but do not post a picture. We discuss resulting selection biases from this in Section 7.

[10] After randomizing over (D_i, PQ_i, VC_i, H_i) does the remaining *unalterable features* even have any significant role left to play in our attractiveness measure? In Appendix A.6.4 we show that (D_i, PQ_i, VC_i, H_i) have a combined attractiveness explanatory power of 13.2% (r-squared), a big chunk of remaining attractiveness explanatory power likely comes from these *unalterable features*.

5.3 Estimation

We want to create propensity score matched (PSM) sample of attractive and plain looking groups to identify effect of attractiveness in a Δt year period from $t \rightarrow t + \Delta t$ after MBA graduation. $pScore_{i,t}(\phi)$ parameterized by ϕ denotes individual i's propensity to be assigned in attractive group instead of the plain looking group at t years i.e. start of the $t \rightarrow t + \Delta t$ period. We find a plain-looking ($b_{j,t} = 0$) counterpart for every attractive individual ($b_{i,t} = 1$) with closest propensity score $pScore_{i,t}(\phi) \simeq pScore_{j,t}(\phi)$. We repeat this procedure 15 times to create propensity score matched samples for $0 \rightarrow \Delta t, 1 \rightarrow 1 + \Delta t, 2 \rightarrow 2 + \Delta t, \dots, 14 \rightarrow 14 + \Delta t$. Note that the propensity score models are different every time because JD_{it} is time variant. All these 15 samples stacked together make up our PSM estimation sample.

$$pScore_{i,t}(\phi) = \frac{1}{1 + \exp(-[D_i, E_i, JD_{it}, PQ_i, VC_i, H_i] * \boldsymbol{\phi})} \quad (6)$$

We can now estimate size of attractiveness bias (β_2) as,

$$r_{i,t+\Delta t} = \beta_0 + \beta_1 D_i + \boldsymbol{\beta_2 b_{i,t}} + \beta_3 JD_{i,t} + \beta_4 E_i + u_{i,t} \quad (7)$$

$$r_{i,t+\Delta t} = \beta_0 + \beta_1 D_i + \mathbf{1}\left(t \leq T^{cutoff}\right)\boldsymbol{\beta_2^B b_{i,t}} + \boldsymbol{\beta_2^P b_{i,t}} + \beta_3 JD_{i,t} + \overrightarrow{\beta_4} E_i + u_{i,t} \quad (8)$$

$$D_i \equiv (gender_i, nameEthnicity_i)$$
$$JD_{i,t} \equiv (industry_{it}, jobType_{it}, employerSize_{it}, locationSize_{it})$$
$$E_i = (ugRank_i, ugDegree_i, mbaRank_i)$$

The equation 7 above assumes that attractiveness bias is constant throughout the career, i.e., $\beta_2 = \beta_2^{t,t+\Delta t} \; \forall \, t$. We originally hypothesize dynamics in attractiveness bias, i.e., belief and preference bias in early career ($\beta_2^{t,t+\Delta t} = \beta_2^B + \beta_2^P \; \forall \, t \leq T^{cutoff}$) and only preference bias in late career ($\beta_2^{t,t+\Delta t} = \beta_2^P \; \forall \, t > T^{cutoff}$). Thus, the equation 8 estimates the size of belief bias (β_2^B) and preference bias (β_2^P) through a 15 year post-MBA career. The remaining covariate explain demographic bias (D_i), idiosyncracies of job domain ($JD_{i,t}$), and impact of education (E_i) on career success.[11]

A similar model specification has been used in income and earning equations. The literature (Zhang and Zhou 2017, Altonji et al. 2013, Browning and Ejrnaes 2013) on individual earning dynamics captures the dynamics by using a time series ARMA process and time-invariant heterogeneity, such as schooling, ethnicity, experience, employment duration and job tenure. We have made numerous choices in the overall PSM based estimation above:

(i) A logistic (instead of SVM or decision tree classifier) propensity score model with linear (instead of interaction or higher degree) covariate effects.

(ii) Nearest neighbor (instead of exact) matching of treated units with untreated ones without replacement (instead of with replacement).

(iii) Single parameter estimates across the matched sample instead of parameter estimates for different propensity score stratas (Imbens 2015, Eckles and Bakshy 2017).

These choices closely follow recent Marketing Science literature, such as Gordon et al (2019), Yoon (2019) and Cao and Sorescu (2013), for estimation on propensity score matched samples. We compare these PSM design choices with alternatives in Appendix A.9 (i and ii) and A.10.1 (iii).

6. Results

6.1 Bias Dynamics

We estimate our model using 472,934 career snapshots (i, t combinations) for 43,533 individuals. That is an average 10.86 yearly snapshots per individual.[12] After propensity score matching we are left with 275,858 career snapshots. The average percentage standardized bias

[11] Note that controls $(D_i, JD_{i,t}, E_i)$ were already used in propensity score matching. Their role in the regression is purely to add explanatory power and facilitate supplementary analysis. These are also not the full set of available features that were used for propensity score matching. Some of the features have significant correlation among each other with little additional explanatory power. In Appendix A.10.7, we report Variance Inflation Factor (VIF) and robustness of our main findings if the full set of controls were included.

[12] While we study 15 years of an individual's career, this average of 10.86 is naturally smaller because: (i) individuals who graduate after 2002 do not have 15 years of experience history by 2017, and (ii) some individuals may leave gaps in their career history, e.g., job 1 from 2002-2006 and in job 2 from 2010-2013.

across all confounders reduced from 9.04% to 0.36% after PSM (Table 8). This measures how well PSM solves imbalance in confounders. Appendix A.9 reports the percentage standardized bias before and after PSM for all confounders $(D_i, E_i, JD_{it}, PQ_i, VC_i, H_i)$. All the covariates have an imbalance of less than 2% after matching. This is typically considered a successful match. Figure 10 also shows histogram of propensity scores of treatment and control groups before and after matching.

$$Std.\,Bias\ for\ covariate\ x = \frac{|E[x/b_{i,t}=1] - E[x/b_{i,t}=0]|}{\left(0.5 * Var(x/b_{i,t}=1)\ + 0.5 * Var(x/b_{i,t}=0)\right)^{0.5}} \quad (9)$$

$$Rubin's\ B = \frac{|E[pScore/b_{i,t}=1] - E[pScore/b_{i,t}=0]|}{\left(0.5 * Var(pScore/b_{i,t}=1)\ + 0.5 * Var(pScore/b_{i,t}=0)\right)^{0.5}}$$

$$Rubin's\ R = \frac{Var(pScore/b_{i,t}=1)}{Var(pScore/b_{i,t}=0)}$$

Table 8: Treatment and Control statistics before and after PSM. A Rubin's R between 0.5-2.0 and a Rubin's B less than 25 indicate a good match.

Statistic	Before Matching	After Matching
Mean % Std. Bias	9.04%	0.36%
Rubin's R	0.73	1.01
Rubin's B	98.07%	0.74%

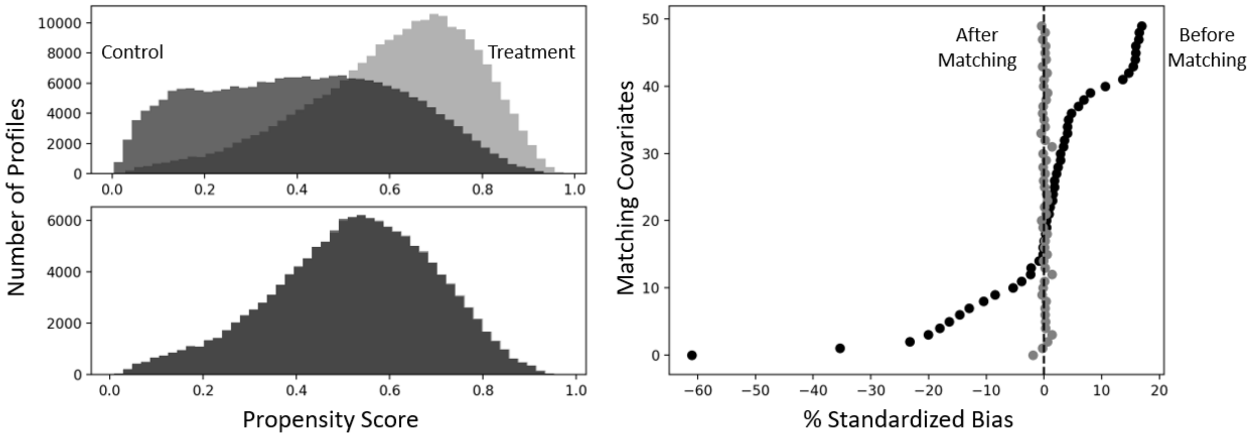


Figure 10: (Left) Histogram of propensity score for treatment and control groups before (Top Left) and after (Bottom Left) PSM. (Right) Percentage Standardized Bias for approximately 50 matching covariates before (dark) and after (light) PSM.

Table 9: Descriptive Statistics for numerical variables after PSM. UG Ranks between 1 to 200 are scaled from 0 to 1. MBA Ranks between 1 to 100 are scaled from 0 to 1.

Variable	Min	Max	Mean	Std. Dev.
Attractiveness ($b_{i,t}$)	0	1	.50	.50
Rank ($r_{i,t}$)	-2.44	3.49	.01	1.0
UG Rank	0	1	0.67	0.43
MBA Rank	0	1	0.38	0.40
Large Location	0	1	.50	.5
Large Employer	0	1	.50	.5

We estimate Table 10 Model 1 and 2 using a static persistent attractiveness premium β_2 throughout the 15 year career period. The estimates for $\beta_2 = 0.011$ are positive and significant at the 1% level. Since ranks ($r_{i,t}$) are standard normal, this means that attractive individuals gain a 0.011 standard deviation in a single year.

$$r_{i,t+\Delta t} = \beta_0 + \beta_1 D_i + \boldsymbol{\beta_2 b_{i,t}} + \beta_3 JD_{i,t} + \beta_4 E_i + u_{i,t} \tag{10}$$

Attractiveness Gap: Here, we define the attractiveness gap that will be used going forward. A gain of a 0.011 standard deviation in a single year means that the job ranks for the two groups are normally distributed as N(0,1) and N(0.011,1) at the end of the year. We can use these distributions to calculate the probability that an attractive individual's rank r_A exceeds the rank r_{nA} of their plain-looking counterpart. This probability comes out to 50.44%. We call this a single-year attractiveness gap of 0.44%. Going forward, we will use this definition of the attractiveness gap. Later in this section, we will look at the lifetime premium through the yearly accumulation of this premium.

$$beauty\ Gap = P(r_A > r_{nA}) = \frac{0.5}{\sqrt{2\pi}} \int |\exp(-(r - 0.011)^2) - \exp(-r^2)|\, dr \tag{11}$$

First, we modify the model to discern β_2^P and β_2^B, denoting the preference and belief biases respectively. The preference bias is modeled to play a persistent role throughout the career. The belief bias is modeled to vanish after the first half of the career, i.e., after 6 years. As discussed in section 2.3, we expect $T^{cutoff} = 6$ years to be more than sufficient duration for prior beliefs to be overcome by actual performance signals (Andreoni & Petrie 2003, Rezlescu et al. 2012). We expect that after 6 years, employers, managers or any other evaluators will have sufficiently long and informative performance history for the subject. In Appendix A.10.4, we report the robustness of our results with respect to three different cut off points for the end of early career, i.e., 4 years, 6 years and 8 years after MBA graduation.

$$r_{i,t+\Delta t} = \beta_0 + \beta_1 D_i + \mathbf{1}(t \leq T^{cutoff})\boldsymbol{\beta_2^B b_{i,t}} + \boldsymbol{\beta_2^P b_{i,t}} + \beta_3 JD_{i,t} + \beta_4 E_i + u_{i,t} \qquad (12)$$

Table 10: [Covariates are skipped here, full Table at end of appendix] (Model 1 and 2) The impact of attractiveness on career rank with and without gender interaction. (Model 3 and 4) The impact of attractiveness preference and attractiveness belief on career rank with and without gender interaction.

	(1) rank ($r_{i,t}$)	(2) rank ($r_{i,t}$)	(3) rank ($r_{i,t}$)	(4) rank ($r_{i,t}$)
Attractiveness	**0.011****	**0.012****		
	(0.003)	(0.004)		
Attractiveness * Female		-0.003		
		(0.008)		
Belief Bias			-0.005	-0.008
			(0.007)	(0.008)
Preference Bias			**0.013****	**0.015****
			(0.005)	(0.005)
Belief Bias * Female				0.013
				(0.015)
Preference Bias * Female				-0.009
				(0.010)
Gender(Male)	0.164**	0.162**	0.163**	0.178**
	(0.004)	(0.006)	(0.004)	(0.007)
Constant	0.495**	0.496**	0.500**	0.490**
	(0.014)	(0.014)	(0.014)	(0.015)
Obs.	275858	275858	275858	275858
R-squared	0.169	0.169	0.169	0.169
Adj R-squared	0.169	0.169	0.169	0.169
F	3111.048	2947.309	2801.290	2437.457
RMSE	0.914	0.914	0.914	0.914

Standard errors are in parenthesis
** $p<0.01$, * $p<0.05$.

We had set out to answer the following three research questions: (i) Is preference bias absent, positive or negative? (ii) Is belief bias absent, positive or negative? (iii) What is the combined effect when preference and belief bias coexist early in professional careers? In response to the first question, the preference bias is positive and significant (Table 10 Model 3 and 4). The preference bias is associated with a single-year attractiveness gap of 0.52% (coefficient of 0.013). This is equivalent to a 50.52% probability that an attractive individual exceeds a plain-looking counterpart, both individuals having started the year in a similar position. Regarding the second question, we find that belief bias is insignificant at a 5% significance level in professional careers after MBA graduation. This does not necessarily mean that belief biases do not exist at all. Prior experimental literature provides abundant evidence for the existence of belief bias in specific employer employee interaction. Our findings suggest that any belief bias in such interactions may not add up to a significant premium to attractive MBA graduates in the post-MBA professional jobs (white-collar jobs). Because the belief bias estimate is insignificant, we refrain making any strong claims about the belief bias being zero or negative. There is some weak indication that penalty for plain looks persist unabated and may even grow slightly even after the individual has spent years showing their skills in the job market. Finally in response to the third question, we find that there is a positive combined belief and preference bias when they coexist early in a career, and this is largely driven by preference.

6.2 Lifetime Bias

Prior empirical research (Hamermesh and Biddle 1994, Zebrowitz and Donald 1991) shows an attractiveness wage gap of 2-8% (Hamermesh and Biddle 1994, Biddle and Hamermesh 1998, Doorley and Siesminska 2015). Our primary model shows a 0.44%-0.52% advantage for attractive individuals in a single year. Similar to previous attractiveness gap studies[13], we also want to develop a model to measure lifetime premium for an attractive individual over the entire 15 year period.

We start with 19,405 individual profiles where we observe at least 15 years of career. As before, we use PSM to narrow down to matched sample of 12,096 profiles. The Table 11 reports that average imbalance across all confounders is reduced from 8.88% to 0.58% after PSM. Figure 11 also shows the propensity to be assigned to treatment and control groups before and after

[13] Wage gap calculations typically combine the effects of different jobs attained by two groups and different wages for the same job. We do not measure wages; instead, we focus only on the differential in the attainment of desirable jobs. The readers can reference the wage gap numbers but should be careful that it's not a like-for-like comparison.

matching. Table 12 provides the descriptive statistics key variables after propensity score matching. In this model, we replace the time-variant characteristics of the job domain $JD_{i,t}$ with time-invariant characteristics JD_i. For example, if an individual switches industries over his or her career, we only use the single industry where the individual spends the maximum duration. Similarly, we use a static measure of individuals' attractiveness at 15 years after MBA graduation.

$$r_{i,15} = \beta_0 + \beta_1 D_i + \boldsymbol{\beta_2 b_i} + \beta_3 JD_i + \beta_4 E_i + u_i \tag{13}$$

We find that attractiveness is associated with a 15 year attractiveness gap of 2.4% (coefficient of 0.06). We had earlier found a yearly attractiveness gap of 0.44-0.52%. It is important to highlight a key distinction using these results. If a successful senior job (say CEO) is dominated by attractive individuals, it may not necessarily be explained by bias when the board picks the CEO. It may be an outcome of gradual bias throughout the career, with early career gap (e.g., MBA school assignment or 1st post MBA-job) playing a significant role. A lifetime bias alone does not clarify when the bias occurs.

Table 11: Treatment and Control statistics before and after PSM

Statistic	Before Matching	After Matching
Mean % Std. Bias	8.88%	1.12%
Rubin's R	0.88	1.01
Rubin's B	93.66%	0.57%

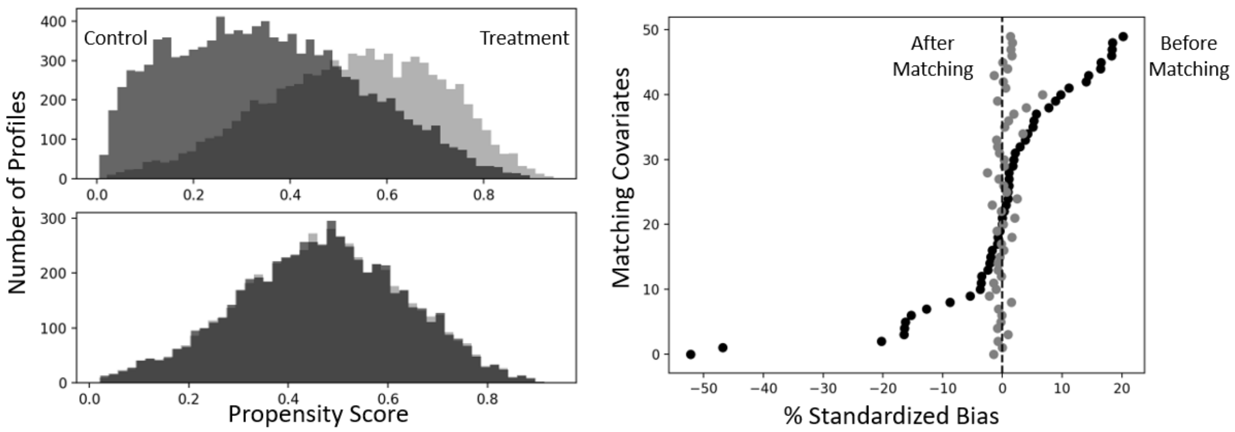


Figure 11: (Left) Histogram of propensity score for treatment and control groups before (Top Left) and after (Bottom Left) PSM. (Right) Percentage Standardized Bias for 50 matching covariates before (dark) and after (light) PSM.

Table 12: Descriptive Statistics after PSM. UG Ranks between 1 to 200 are scaled from 0 to 1. MBA Ranks between 1 to 100 are scaled from 0 to 1.

Variable	Min	Max	Mean	Std. Dev.
Attractiveness ($b_{i,0}$)	0	1	.50	0.5
Rank ($r_{i,t}$)	-2.44	3.29	.03	1.01
UG Rank	0	1	0.69	0.42
MBA Rank	0	1	0.41	0.41
Large Location	0	1	.40	0.49
Large Employer	0	1	.46	0.5

Table 13: [Covariates are skipped here, full Table at end of appendix] *(Model 1) The impact of attractiveness on career rank at the end of 15 years. (Model 2) The impact of attractiveness with gender interaction. (Model 3) The impact of attractiveness with ethnicity interaction. (Model 4) The impact of attractiveness for individual of European ethnicity alone*

	(1) rank ($r_{i,15}$)	(2) rank ($r_{i,15}$)	(3) rank ($r_{i,15}$)	(4) rank ($r_{i,15}$)
Attractiveness	**0.060****	**0.055****	**0.048***	**0.081****
	(0.017)	(0.019)	(0.022)	(0.018)
Attractiveness * Female		0.021		
		(0.040)		
Attractiveness * Ethnicity (Others)			-0.037	
			(0.168)	
Attractiveness * Ethnicity (European)			**0.033***	
			(0.015)	
Gender (Male)	0.203**	0.213**	0.203**	0.209**
	(0.021)	(0.029)	(0.021)	(0.022)
Ethnicity (European)	0.001	0.001	-0.016	
	(0.043)	(0.043)	(0.062)	
Ethnicity (Others)	0.109*	0.109*	0.175*	
	(0.050)	(0.050)	(0.070)	
Constant	0.376**	0.368**	0.385**	0.424**
	(0.069)	(0.070)	(0.081)	(0.058)
Obs.	12100	12100	12100	10176
R-squared	0.151	0.151	0.152	0.153
Adj R-squared	0.150	0.150	0.150	0.151
F	119.372	113.097	108.010	114.503
RMSE	0.932	0.932	0.931	0.926

Standard errors are in parenthesis
** $p<0.01$, * $p<0.05$.

We highlight some role of some familiar factors in career success. Among male and female peers, the male individual exceeds his counterpart with probability 57.9% at the end of 15 year career, i.e., gender gap of 7.9%. This is in line with gender bias studies. An undergraduate university ranked 40 places higher[14] adds a gap of 3.6%. An MBA program ranked 20 places higher[15] adds a gap of 8.3%. A “technical” job type in the “IT” industry hurts long term career, while “management” or “consultant” job type in the “management” industry assists career success.

6.3 Bias Interactions

Gender and Ethnicity

While a majority of attractiveness bias literature does not highlight any differences between men and women, a few experimental studies (Agthe et al. 2010, Ruffle and Shtudiner 2014) point to penalty for attractive young women in interaction with same sex evaluators. It is suggested that a penalty occurs when envy or perceived lack of career seriousness is heightened in an interaction. An attractive women may go through a mix of interactions where an evaluator’s beliefs and preferences aid or penalize them. The combined attractiveness bias is unclear.

Table 10 Model 1, 3 and Table 13 Model 2, all show statistically insignificant gender interaction terms for the dynamic and lifetime models. Our results suggest that all factors considered simultaneously, penalizing interactions faced by young attractive women may not have a dominant role overall. Given our focus on top 100 ranked MBA programs, misperception of lack of career seriousness for young attractive women is likely minimal. It is critical to note that while the preference bias may similarly favor attractive male and female employees, it may be more likely to take the form of sexual harassment for female employees. Premium from preference bias is likely driven by inclination for social interactions with both attractive male and female employees. This preference may go further toward pursuit of (often unwelcomed) romantic interactions with attractive female employees by predominantly male superiors.

[14] 40 places in undergraduate ranks is approximately equivalent to 1 std. dev. above average in standardized N(0,1) university ranks.

[15] 20 places in MBA ranks is approximately equivalent to 1 std. dev. above average in standardized N(0,1) MBA ranks.

We also discern attractiveness bias for ethnicities (Caucasian/European vs others). Table 13 Model 1 found an advantage of 2.4% for attractive individuals over their plain-looking counterparts. In comparison, Table 13 Model 3, 4 suggest that attractive Caucasian individuals have an advantage of 3.2% (coefficient 0.048 + 0.033) over their plain-looking Caucasian counterparts. We have limited data to differentiate among ethnicities with smaller representation in our sample e.g., African American, Asian, etc. We can speculate that a larger attractiveness bias for Caucasian individuals may stem from a large fraction of evaluators and managers also being Caucasian. Any preferences for social or romantic relationships with attractive individuals are likely heightened between employers and employees of the same ethnicity.

Skills and Job Domains

We expect preference-based bias in favor of attractive individuals to exist across the board, but it may be particularly large where an employee is expected to have significant social interaction with an employer and its clients. Table 14 Model 1 reports an attractiveness premium coefficient of 0.156 (0.065 + 0.091) for individuals with **undergraduate degrees in Arts** compared to premium of 0.065 or lower for all other undergraduate degrees. However, this coefficient is statistically insignificant (significant at 10% level). Model 2 reinforces a large premium for Arts undergraduates using a breakdown of Arts vs all other undergraduate degrees. We speculate that Arts undergraduate degree holders, more than their Business, Science or Engineering peers, are likely considered for roles with social interaction instead of isolated desk jobs. We would expect similar social interactions for consultants and management professionals. Model 3 reports a large attractiveness premium of 0.118 for individuals in the **Management industry category**. This industry category interaction effect is statistically significant at 5% level. Model 4 reinforces this result with a breakdown of Management vs all other industry categories. Model 5 reports a large attractiveness premium coefficient for **Consultants**, but this is statistically insignificant.

Thus far, we have discussed the role of attractiveness in career progression within a job domain. We have not discussed the role of attractiveness in assignment to these job domain in the first

Table 14: [Covariates are skipped here, full Table at end of appendix] *The impact of attractiveness on career rank at the end of 15 years. (Model 1) Interaction with Undergraduate degree, (Model 2) Interaction with undergraduate degree classified as Arts vs all others (Model 3) Interaction with Industry Category, (Model 4) Interaction with Industry Category classified as Management vs all others. (Model 5) Interaction with Job Type. (Model 6) Interaction with MBA Rank.*

	(1)	(2)	(3)	(4)	(5)	(6)
	rank	rank	rank	rank	rank	rank
	$(r_{i,15})$	$(r_{i,15})$	$(r_{i,15})$	$(r_{i,15})$	$(r_{i,15})$	$(r_{i,15})$
Attractiveness	**0.065***	**0.045***	**0.053***	**0.042***	**0.093**	**0.092****
	(0.027)	(0.018)	(0.028)	(0.019)	(0.089)	(0.024)
Attractiveness *	**0.091**					
UG (Arts)	(0.053)					
Attractiveness *	0.004					
UG (Business)	(0.082)					
Attractiveness *	-0.042					
UG (Science)	(0.038)					
Attractiveness *		**0.112***				
UG (Arts)		(0.049)				
Attractiveness *			-0.043			
Ind. Cat. (Others)			(0.083)			
Attractiveness *			-0.002			
Ind. Cat. (IT)			(0.066)			
Attractiveness *			**0.075***			
Ind. Cat. (Management)			(0.031)			
Unattractiveness *				**0.086***		
Ind. Cat. (Management)				(0.042)		
Attractiveness *					**0.060**	
Job Type (Consultant)					(0.360)	
Attractiveness *					-0.133	
Job Type (Technical)					(0.187)	
Attractiveness *					0.019	
Job Type (Others)					(0.091)	
Attractiveness *						**-0.076**
MBA Rank						(0.041)
Gender(Male)	0.203**	0.203**	0.203**	0.203**	0.203**	0.202**
	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)
Constant	0.330**	0.441**	0.373**	0.385**	0.351**	0.359**
	(0.072)	(0.074)	(0.071)	(0.069)	(0.076)	(0.069)
Obs.	12100	12100	12100	12100	12100	12100
R-squared	0.151	0.151	0.151	0.151	0.151	0.151
Adj R-squared	0.150	0.150	0.150	0.150	0.150	0.150
F	102.656	113.396	102.602	113.340	102.422	113.296
RMSE	0.931	0.931	0.931	0.931	0.932	0.931

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

place. In fact, Propensity Score Matching eliminates any imbalance in assignment across job domains. The Table 15 reports proportion of attractive individuals in different undergraduate degrees, industries, and job types prior to PSM. Science undergraduate, IT industry and Technical jobs have the fewest attractive individuals. If individuals were strategic, we would expect attractive individuals to steer away from domains where they expect to get the least premium for their looks. In line with this argument, Science or Engineering degrees and Technical jobs in IT industry typically require the least social interactions. We are not aware of any prior research which has looked at attractiveness bias differences across job domains, and related strategic education or career choices by individuals themselves. We have provided some model free breakdown and conjectures backed by (weak) correlations; this merely constitutes a starting point for future research.

Table 15: Number of attractive individuals as a percentage of all individuals in the domains – UG Degree, Job Type and Industry Category.

UG Degree	%Attractive
Science	**45.9**
Arts	48.3
Business	51.2
Others	55.3

Industry Category	%Attractive
IT	**48.4**
Others	49.3
Management	49.4
Finance	54.2

Job Type	%Attractive
Technical	**41.8**
Consultant	50.0
Others	50.5
Management	50.5

Table 14 Model 6 reports an attractiveness premium of 0.092 for attractive individuals and an interaction coefficient of -0.072 with MBA Rank. The interaction term is statistically weak, i.e., significant only at 10% level. The weak evidence would suggest that attractive individuals from top ranked MBA program has an advantage of 3.6% (coefficient of 0.092) over their plain-looking counterparts. In comparison, attractive individuals from a 100 ranked MBA program would have an advantage of merely 0.8% (coefficient of 0.022 = 0.092 – 1.0*0.072). In Appendix A.10.6 we also report results using profiles from top 20 MBA programs alone. We find a preference bias attractiveness gap of 1.1% per year (coefficient of 0.021) for top 20 MBA programs, instead of attractiveness gap of 0.44% per year for all MBA programs. Once again, we are not aware of prior literature which suggest more attractive premium in upper echelons of education and job market. It is likely that top MBA graduates have a realistic shot at extremely rare desirable career

outcomes such as C-level executives. Role of all secondary factors (gender, ethnicity, appearance) may be heightened in such settings.

7. Robustness Checks

Our analysis covers only a small fraction of all MBA graduates. A systematically biased selection of profiles can interfere with our results. We discuss the following selection concerns: (i) individuals' choice to not post a profile picture, (ii) individuals' choice to not create a profile or to drop out of the job market entirely because of lack of success, better outside options and phases of economy wide unemployment crisis.

One could argue that plain-looking individuals believe that by not posting their picture, an uncertain observer assigns a higher-than-actual face attractiveness guess. This means that individuals with observed pictures are under-sampled from the plain-looking subgroup. This in itself is not a problem as long as this strategy holds true across both successful and unsuccessful plain-looking individuals. However, one could extend the argument that serious individuals are strategic about the effect of their picture. Therefore, serious plain-looking individuals choose not to post their pictures. Non-serious plain-looking individuals do not pay attention to this strategic choice and post their true pictures. Thus, our dataset will suffer from under sampling of serious (or professional, strategic or ambitious) plain-looking individuals who are more likely to be successful. This is a problem if successful unattractive individuals are increasingly likely to remove their pictures. The remaining profiles (with posted pictures) would appear to have increasingly fewer successful unattractive individuals, thus giving a misperception of attractiveness bias. If true, this would mean that profiles without pictures are systematically more successful in job markets than profiles with pictures. In Appendix A.11.2, we test and reject this hypothesis.

One could argue that unsuccessful individuals (small $r_{i,t-1}$) may be more likely to drop out of the job market if they have an outside option. Consider profiles collected in 2017 for two graduating classes from 1997 and 2012. We observe average career outcome 5 years after MBA for both these graduating classes: in year 2002 for class of 1997, and in year 2017 for class of 2012. If unsuccessful individuals systematically drop out, then unsuccessful individuals in the class of 1997 would have dropped out between 2002 and 2017, thus they are left out of our sample. As a result, the 5 year performance for class of 1997 would appear to be higher compared to the 5 year performance for the class of 2012. In Appendix A.11.1, we test and reject this hypothesis formally using all MBA classes.

A second argument could be that attractive individuals are more (or less) likely to drop out. There may be better (or worse) outside options for attractive individuals (e.g., financially secure marriage prospects). If attractive individuals drop out systematically, we would expect the average attractiveness of the observed sample of class of 2005 to be smaller than that of class of 2007. We can match these two classes on the undergraduate graduation year as a proxy for matching the current age. Any difference between the average attractiveness would now point to systematic drop out by attractive individuals. In Appendix A.11.1, we test and reject this hypothesis formally using all MBA classes.

In addition to success and attractiveness, dropouts may also systematically happen based on gender and graduation era. This may be because of the following: (i) The outside option is better (or worse) for one gender than the other (e.g., homemaker or blue collar work). (ii) Certain periods may have unusually low employment, e.g., the 2008-2009 crisis, where unemployment temporarily went from 4% to 10%. We have access to a secondary dataset that provides visibility into the gender and graduating class of individuals whose profiles are selected out of our sample. In Appendix A.11.3, we estimate a Heckman's self-selection model on this dataset to match our original findings. This secondary dataset uses a completely different profile search procedure. However, it matches our original attractiveness bias results. This provides additional confidence that using the black box platform search functionality to collect profiles is not biasing our results.

7.1 Other Limitations

We highlight two key limitations for our research methodology. First, the majority of profiles in our analysis come from top 100 MBA schools. Therefore, our results are directly applicable to the higher echelon of white-collar professional careers in the United States. Practitioners should be cognizant before extrapolating our conclusion to wider professional work settings. Second, we only observe a single photograph for every person. This invites potential criticism that we do not know what the individual looked like when they were hired or promoted through their career. To allay such concerns, we apply a best-in-class deep learning based image-morphing technique. We also check the robustness of our findings with respect to different calculations for attractiveness. Nevertheless, we acknowledge that deep learning based image morphing techniques are nascent and must be taken with a grain of salt.

8. Conclusion

Our research finds an attractiveness gap of 2.4% over a 15 year career after MBA graduation. This 15 year gap is an accumulation of a yearly 0.44-0.52% gap driven by preference bias. This means that the penalty for plain looks persist unabated, even after an individual has spent years showing their skills in the job market. A gap of 2.4% may appear small in the context of the ongoing debate on perverse taste-based harassment at workplace. It must be noted that we only focus on permanent characteristics of face structure, and consciously eliminate numerous other superficial characteristics such as hairstyle, fashion, clothing and makeup. While we do not find any significant evidence of belief bias, it does not imply that there is no belief bias in the job market. Consider work settings, such as hiring an undergraduate with no work experience and thus no differentiating signal of skills, accepting a salespersons pitch in absence of objective signal of his reliability, voting for a candidate without looking at her record, etc. These are all regular instances where evaluators' beliefs take over, particularly when they do not have any expectation of spending time with the employee, salesperson or election candidate respectively. Our results suggest that such settings do not have a dominant role overall in the post-MBA professional job market. Further, while we capture average career differential over tens of thousands of professionals, particular individuals may face more or fewer biases.

Following from our research, future experimental researchers could identify what level of future interaction expectations creates preference bias during hiring and promotions. We also set the stage for future research on attractiveness bias for different employee skills, job type and industry. We also develop quantitative measures of evolving attractiveness ratings and job rankings that should be more generally usable for future research in bias, personal marketing, behavioral perception and labor market studies.

This direction of research has significant implications for practitioners. Gender and race biases have received much attention in recent times. Google, like many other companies, provides their employees with unconscious-bias training (Feloni 2016). Moreover, hiring procedures are increasingly emphasizing the importance of performance-related information rather than demographic information. Such training, awareness and emphasis on performance are targeted at the elimination of stereotypes and belief biases. The role of beliefs in creating bias may be avoided under the right settings, but preferences that engender bias are much harder to remove. The media has highlighted sexual harassment across industries. Such pure taste-based, perverse favoritism cannot simply be eliminated by small policy changes. It rather requires a social and cultural change.

References

[1] Agthe, M., Spörrle, M., & Maner, J. K. (2010). Don't hate me because I'm beautiful: Anti-attractiveness bias in organizational evaluation and decision making. *Journal of Experimental Social Psychology*, *46*(6), 1151-1154.

[2] Altonji, J. G., Smith Jr, A. A., & Vidangos, I. (2013). Modeling earnings dynamics. Econometrica, 81(4), 1395-1454.

[3] Amos, Brandon, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. OpenFace: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[4] Andreoni, James, and Ragan Petrie. ”Beauty, gender and stereotypes: Evidence from laboratory experiments.” Journal of Economic Psychology 29.1 (2008): 73-93.

[5] Aral, S., Brynjolfsson, E., & Van Alstyne, M. (2012). Information, technology, and information worker productivity. Information Systems Research, 23(3-part-2), 849-867.

[6] Baker, George, Michael Gibbs, and Bengt Holmstrom. 1994a. The internal economics of the firm: Evidence from personnel data. Quarterly Journal of Economics 109, no. 4:881919.

[7] Becker, H. S., & Geer, B. (1957). Participant observation and interviewing: A comparison. *Human organization*, *16*(3), 28-32.

[8] Berggren, N., Jordahl, H., & Poutvaara, P. (2010). The looks of a winner: Beauty and electoral success. *Journal of Public Economics*, *94*(1-2), 8-15.

[9] Biddle, J. E., & Hamermesh, D. S. (1998). Beauty, productivity, and discrimination: Lawyers' looks and lucre. *Journal of Labor Economics*, *16*(1), 172-201.

[10] Bloch, Peter H., Frederic F. Brunel, and Todd J. Arnold. ”Individual differences in the centrality of visual product aesthetics: Concept and measurement.” Journal of consumer research 29.4 (2003): 551-565.

[11] Bohren, J. A., Imas, A., & Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. *American economic review*, *109*(10), 3395-3436.

[12] Bradley J. Ruffle and Zeev Shtudiner. 2014. Are Good-Looking People More Employable?. Management Science. 1760 1776.

[13] Browning, M., & Ejrnæs, M. (2013). Heterogeneity in the dynamics of labor earnings. Annu. Rev. Econ., 5(1), 219-245.

[14] Bryant, Adam (1999), ”Fashion’s Frat Boy,” Newsweek, (September 13), 40.

[15] Buss, D. M., M. Haselton. 2005. The evolution of jealousy. Trends in Cognitive Sciences 9(11) 506507.

[16] Cao, Z., & Sorescu, A. (2013). Wedded bliss or tainted love? Stock market reactions to the introduction of cobranded products. *Marketing Science*, *32*(6), 939-959.

[17] Chen, B. C., Chen, C. S., & Hsu, W. H. (2014, September). Cross-age reference coding for age-invariant face recognition and retrieval. In *European conference on computer vision* (pp. 768-783). Springer, Cham.

[18] Christopher Y. Olivola and Alexander Todorov. 2010. Elected in 100 milliseconds: Appearance-Based Trait Inferences and Voting. Journal of non verbal behavior. 34:83110.

[19] Chung, D. J. (2015). How much is a win worth? An application to intercollegiate athletics. Management Science, 63(2), 548-565.

[20] Cryder, C., Botti, S., & Simonyan, Y. (2017). The Charity Beauty Premium: Satisfying Donors'"Want" Versus "Should" Desires. *Journal of Marketing Research*, *54*(4), 605-618.

[21] Daniel S Hamermesh. 2011. Beauty Pays: Why attractive people are more succesful. Princeton University Press.

[22] Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of personality and social psychology*, *24*(3), 285.

[23] Donors Want Versus Should Desires. Journal of Marketing Research: August 2017, Vol. 54, No. 4, pp.605-618.

[24] Eckles, D., & Bakshy, E. (2017). Bias and high-dimensional adjustment in observational studies of peer effects. *arXiv preprint arXiv:1706.04692*.

[25] Edwards, Jim (2003), ”Saving Face,” Brandweek, (October 6), 16.

[26] Edwards, Jim (2003), ”Whitewash?” Adweek, (October 6), 14.

[27] Erdem, T., & Keane, M. P. (1996). Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing science*, *15*(1), 1-20

[28] Fang, H., & Moro, A. (2011). Theories of statistical discrimination and affirmative action: A survey. In *Handbook of social economics* (Vol. 1, pp. 133-200). North-Holland.

[29] Fisher, Robert J., and Yu Ma. ”The price of being beautiful: Negative effects of attractiveness on empathy for children in need.” Journal of Consumer Research 41.2 (2014): 436-450.

[30] Fruhen, L. S., Watkins, C. D., & Jones, B. C. (2015). Perceptions of facial dominance, trustworthiness and attractiveness predict managerial pay awards in experimental tasks. *The Leadership Quarterly*, *26*(6), 1005-1016.

[31] Fu, Y., Hospedales, T. M., Xiang, T., Gong, S., & Yao, Y. (2014, September). Interestingness prediction by robust learning to rank. In *European conference on computer vision* (pp. 488-503). Springer, Cham.

[32] Gan, J., Li, L., Zhai, Y., & Liu, Y. (2014). Deep self-taught learning for facial beauty prediction. *Neurocomputing*, *144*, 295-303.

[33] Gao, L., Li, W., Huang, Z., Huang, D., & Wang, Y. (2018, August). Automatic facial attractiveness prediction by deep multi-task learning. In *2018 24th International Conference on Pattern Recognition (ICPR)* (pp. 3592-3597). IEEE.

[34] George-Levi Gayle, Limor Golan and Robert A. Miller. 2012. Gender Differences in Executive Compensation and Job Mobility. Journal of Labor Economic. Vol. 30, No. 4 (October 2012), pp. 829-872.

[35] Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, *38*(2), 193-225.

[36] Hamermesh D. S., J. E. Biddle. 1994. Beauty and the labor market. American Economic Review 84(5) 11741194.

[37] Heilman, M. E., & Saruwatari, L. R. (1979). When beauty is beastly: The effects of appearance and sex on evaluations of job applicants for managerial and nonmanagerial jobs. *Organizational Behavior and Human Performance*, *23*(3), 360-372.

[38] HendersonKing, Eaaron, and Donna HendersonKing. ”Media effects on women’s body esteem: Social and individual difference factors.” Journal of Applied Social Psychology 27.5 (1997): 399-417.

[39] Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, *50*(2), 373-419.

[40] Jennifer J. Argo, Darren W. Dahl, Andrea C. Morales (2008) Positive Consumer Contagion: Responses to Attractive Others in a Retail Context. Journal of Marketing Research: December 2008, Vol. 45, No. 6, pp. 690-701

[41] Kulka, R. A., & Kessler, J. B. (1978). Is Justice Really Blind?–The Influence of Litigant Physical Attractiveness on Juridical Judgment 1. *Journal of Applied Social Psychology*, *8*(4), 366-381.

[42] Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological bulletin*, *126*(3), 390.

[43] Liang, L., Xie, D., Jin, L., Xu, J., Li, M., & Lin, L. (2017, September). Region-aware scattering convolution networks for facial beauty prediction. In *2017 IEEE International Conference on Image Processing (ICIP)* (pp. 2861-2865). IEEE.

[44] Luxen, M. F., & Van De Vijver, F. J. (2006). Facial attractiveness, sexual selection, and personnel selection: When evolved preferences matter. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *27*(2), 241-255.

[45] Ma, Correll, & Wittenbrink (2015). The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data. *Behavior Research Methods, 47*, 1122-1135.

[46] Martin, Mary C., and James W. Gentry. ”Stuck in the model trap: The effects of beautiful models in ads on female pre-adolescents and adolescents.” Journal of advertising 26.2 (1997): 19-33.

[47] Mobius, Markus M. and Tanya S. Rosenblat. 2006. Why beauty matters. American Economic Review 96, no. 1: 222-235.

[48] Morales, Andrea C., and Gavan J. Fitzsimons. ”Product contagion: Changing consumer evaluations through physical contact with disgusting products.” Journal of Marketing Research 44.2 (2007): 272-283.

[49] Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of nonverbal behavior*, *34*(2), 83-110.

[50] Olivola, C. Y., & Todorov, A. (2017). The biasing effects of appearances go beyond physical attractiveness and mating motives. Behavioral Brain and Sciences, 40, 33-34.

[51] Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. Trends in Cognitive Sciences, 18(11), 566-570.

[52] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web*. Stanford InfoLab.

[53] Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. ”Running experiments on amazon mechanical turk.” (2010).

[54] Petty, Richard E., John T. Cacioppo, and David Schumann. ”Central and peripheral routes to advertising effectiveness: The moderating role of involvement.” Journal of consumer research 10.2 (1983): 135-146.

[55] Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review*, *62*(4), 659-661.

[56] Reingen, P. H., & Kernan, J. B. (1993). Social perception and interpersonal influence: Some consequences of the physical attractiveness stereotype in a personal selling setting. *Journal of Consumer Psychology*, *2*(1), 25-38.

[57] Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable Facial Configurations Affect Strategic Choices in Trust Games with or without Information about Past Behavior. PLoS ONE, 7(3), e34293.

[58] Richard Feloni. 2016. Google’s unconscious bias training presentation. Business Insider.

[59] Richins, Marsha L. ”Social comparison and the idealized images of advertising.” Journal of consumer research 18.1 (1991): 71-83.

[60] Rubinstein, Y., & Weiss, Y. (2006). Post schooling wage growth: Investment, search and learning. Handbook of the Economics of Education, 1, 1-67.

[61] Ruffle, B. J., & Shtudiner, Z. E. (2014). Are good-looking people more employable?. *Management Science*, *61*(8), 1760-1776.

[62] Sagarin, B., D. V. Becker, R. E. Guadagno, L. D. Nicastle, A. Millevoi. 2003. Sex differences (and similarities) in jealousy: The moderating influence of infidelity experience and sexual orientation of the infidelity. Evolution and Human Behavior 24 1723.

[63] Santos, J. Reynaldo A. ”Cronbachs alpha: A tool for assessing the reliability of scales.” Journal of extension 37.2 (1999): 1-5.

[64] Schmid, Kendra, David Marx, and Ashok Samal. ”Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios.” Pattern Recognition 41.8 (2008): 2710-2717.

[65] Small, Deborah A., and Nicole M. Verrochi. ”The face of need: Facial emotion expression on charity advertisements.” Journal of Marketing Research 46.6 (2009): 777-787.

[66] Solnick, S. J., & Schweitzer, M. E. (1999). The influence of physical attractiveness and gender on ultimatum game decisions. *Organizational behavior and human decision processes*, *79*(3), 199-215.

[67] Stanley, T. D., & Jarrell, S. B. (1998). Gender wage discrimination bias? A meta-regression analysis. *Journal of Human Resources*, 947-973.

[68] Tanzina Vega. 2015. Working while brown: What discrimination looks like now. CNN Money.

[69] Todorov, A., & Oosterhoof, N. N. (2011) Modeling social perception of faces. *Signal Processing Magazine, IEEE, 28,* 117-122.

[70] Todorov, A., Dotsch, R., Porter, J., Oosterhof, N., & Falvello, V. (2013).Validation of Data-Driven Computational Models of Social Perception of Faces.*Emotion, 13*(4), 724-738.

[71] Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. Annual Review of Psychology, 66.

[72] Townsend, Claudia, and Sanjay Sood. ”Self-affirmation through the choice of highly aesthetic products.” Journal of Consumer Research 39.2 (2012): 415-428.

[73] Townsend, Claudia, and Suzanne B. Shu. ”When and how aesthetics influences financial decisions.” Journal of Consumer Psychology 20.4 (2010): 452-458.

[74] Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, *3*(1), 71-86.

[75] Xie, D., Liang, L., Jin, L., Xu, J., & Li, M. (2015, October). Scut-fbp: A benchmark dataset for facial beauty perception. In *2015 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 1821-1826). IEEE.

[76] Xu, J., Jin, L., Liang, L., Feng, Z., Xie, D., & Mao, H. (2017, March). Facial attractiveness prediction using psychologically inspired convolutional neural network (PI-CNN). In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1657-1661). IEEE.

[77] Yoon, T. J. (2019). Quality Information Disclosure and Patient Reallocation in the Healthcare Industry: Evidence from Cardiac Surgery Report Cards. *Marketing Science*.

[78] Zhang, Y., & Zhou, Q. (2017). Estimation for time-invariant effects in dynamic panel data models with application to income dynamics. Econometrics and Statistics.

[79] Zhang, Z., Song, Y., & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5810-5818).

Appendix

A.1 Data Description

Table 16: Self-reported MBA school names with MBA ranks and number of profiles. Some profiles self-report either a less frequent pseudonym or a less frequent business school. We use numerous text matching methods to accumulate various pseudonyms.

MBA Rank	MBA School Name (as self reported)	Profiles
0	Harvard Business School	1837
	Harvard University	304
1	Stanford University	603
	Stanford University Graduate School of Business	287
2	The University of Chicago - Booth School of Business	1012
	The University of Chicago Booth School of Business	971
3	University of California, Berkeley - Walter A. Haas School of Business	619
	University of California, Berkeley, Haas School of Business	409
	University of California, Berkeley	278
4	Massachusetts Institute of Technology - Sloan School of Management	1052
	MIT Sloan School of Management	260
	Massachusetts Institute of Technology	187
5	Northwestern University - Kellogg School of Management	1710
	Northwestern University	250
	Kellogg School of Management	159
8	Carnegie Mellon University - Tepper School of Business	1565
	Duke University - The Fuqua School of Business	748
	Duke University	159
9	Cornell University - S.C. Johnson Graduate School of Management	832
	Cornell University - Johnson Graduate School of Management	821
	Cornell University	161
12	University of Michigan - Stephen M. Ross School of Business	1737
	University of Michigan	263
14	UCLA Anderson School of Management	1428
	University of California, Los Angeles - The Anderson School of Management	1045
	University of California, Los Angeles	302
15	University of North Carolina at Chapel Hill - Kenan-Flagler Business School	1956
	UNC Kenan-Flagler Business School	138
	University of North Carolina at Chapel Hill	228
16	The University of Texas at Austin - The Red McCombs School of Business	903
	The University of Texas at Austin	229
	The University of Texas at Austin - Red McCombs School of Business	765
19	New York University - Leonard N. Stern School of Business	245
20	Georgetown University - The McDonough School of Business	427
28	University of Washington, Michael G. Foster School of Business	354
	University of Washington	102
60	The University of Connecticut School of Business	625
97	The University of New Mexico - Robert O. Anderson School of Management	547
	University of Maryland - Robert H. Smith School of Business, Georgia State University - J. Mack Robinson College of Business, University of Washington, Michael G. Foster School of Business, University of Washington, Michael G. Foster School of Business, Indiana University - Kelley School of Business, Purdue University - Krannert School of Management etc.	18015

Our primary dataset is collected using "MBA" as a search string along with filters on location = "United States" and school names. The MBA school names are derived from top 100 ranked MBA programs on US News business school rankings (as per rankings in 2016). Even though we search for exact business school names, the search functionality often returns individuals who may have attended a similarly sounding MBA program. Table 16 shows a breakdown of number of profiles per school.

We also collect a secondary dataset to study sample selection issues in our primary dataset. We access the school directory of a top 10 MBA program in U.S. This gives us access to names and graduating year for 15,640 students between 1990 and 2016. We utilize Bing Search API to search for professional profiles for these graduates. Specifically, we rely on search terms such as "Name" + "School" + "MBA" + "professional platform web domain name" as well as numerous permutations of the same. We access all relevant professional profiles returned by such searches. In situations where Bing searches result in more than one closely relevant profiles (e.g. students with common names) we are able to reject profiles where the self-reported graduation year does not match the information from the school name. This secondary data does not rely on the platform's search functionality. Further it provides visibility into exact graduates that have been selected out during the data collection. The Figure 12 represents the size of graduating class from the school directory and the respective number of public profiles we are able to search via the procedure described above (3399 in total). Out of these 3,399 less than 1500 profiles have complete profile fields necessary for analysis.

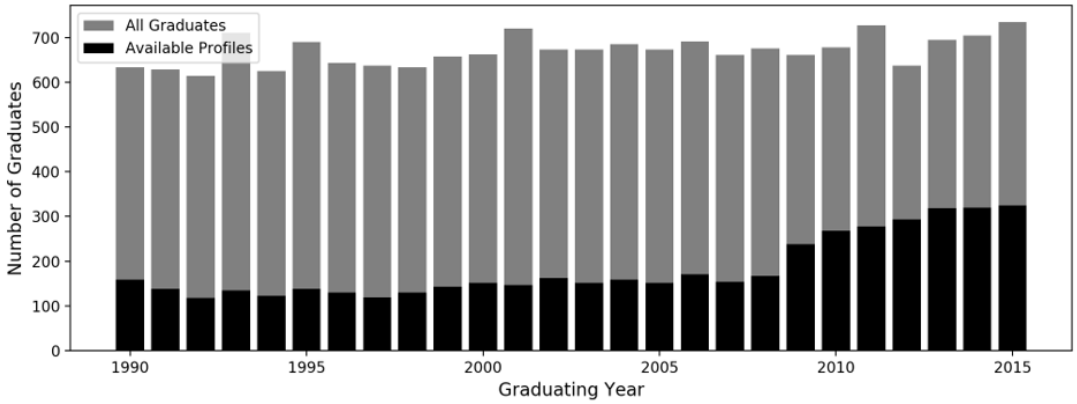


Figure 12: Size of graduating classes (from school directory) between 1990-2015 and the number of profiles collected (from professional social network) for each graduating class.

A.2 Attractiveness Prediction Model

A.2.1 Validation on held-out sample

We evaluate a variety of models as well as a host of hyper parameter settings on each model. We follow rudimentary practices in Machine Learning literature to iterate over some of standard available algorithms. We present below only a small glimpse of all these evaluations. *Table 17* shows tuning of penalty hyper parameter on a support vector regression model. A penalty parameter of 0.5 allows the model to attain a root mean squared error of 0.81 over a 1 to 7 scale on held out sample validation. Similarly, we fine tune hyper parameters for a variety of regression and classification models. Once optimal hyper parameters are identified for each model, we make a comparison among the models. *Table 18* shows comparison between classification models – Support Vector Classification (C = 1.2), Random Forest (depth 20, minimum features=3, minimum leaf size = 5), Logistic, K-Nearest Neighbor (K = 3), Adaboost, Gaussian Naïve Bayes and Quadratic Discriminant Analysis. We use the support vector regression model, which turns out to be the best model based on cross-validation results. Nevertheless we also identify support vector classification as the best binary classification model. The need for the classification model will become clear when we explain the external testing sample in A.2.2.

Table 17: Presence evaluation on tuning penalty hyper parameter C on Support Vector Regression. Results from five fold cross validation are presented as mean and s.d. over the five folds. As C increases beyond 0.5 the training error reduces but validation error increases i.e. overfitting. Therefore C = 0.5 is the optimal choice.

Hyper parameter	Adj. R-squared	Pearson Correlation	Training RMSE	Cross-Validation RMSE
C = 0.1	0.32	0.61	0.91	1.04 ± 0.12
C = 0.5	0.38	0.62	0.82	$\mathbf{0.81 \pm 0.07}$
C = 1.0	0.39	0.62	0.81	0.90 ± 0.04
C = 10.0	0.39	0.63	0.81	0.91 ± 0.03

We follow CS literature on facial attractiveness to decide on our evaluation measures (correlation, mse, roc-auc). Below are some of the key papers in the literature and their performances,

- Gan et al (Neurocomputing, 2014): Deep Self Taught Learning. Correlation 0.774. MSE 1.093.
- Xu et al (IEEE, 2017): PI-CNN. Pearson Correlation 0.85
- Liang et al (IEEE, 2017): Region-Aware CNN. Pearson Correlation 0.83
- Gao et al (ICPR, 2018): Multi Task CNN. Correlation 0.92.

Table 18: Results from alternative classification ML models. A cross validation accuracy of 0.69 for Support Vector Classification model means that 69% of labels are predicted correctly while 31% are predicted incorrectly. We also show F1 and Area under the RO curve.

Model	Training			Cross Validation		
	Accuracy	F1	ROC-AUC	Accuracy	F1	ROC-AUC
SVC	0.76	0.72	0.72	$\mathbf{0.71 \pm 0.03}$	$\mathbf{0.67 \pm 0.03}$	$\mathbf{0.76 \pm 0.02}$
Random Forest	0.93	0.92	0.92	0.62 ± 0.04	0.59 ± 0.05	0.68 ± 0.02
Logistic	0.72	0.71	0.71	0.67 ± 0.02	0.65 ± 0.02	0.71 ± 0.02
K-NN	0.77	0.76	0.76	0.59 ± 0.02	0.59 ± 0.02	0.63 ± 0.04
Ada Boost	0.89	0.89	0.89	0.64 ± 0.03	0.63 ± 0.03	0.71 ± 0.02
Gaussian NB	0.69	0.66	0.66	0.64 ± 0.02	0.62 ± 0.03	0.73 ± 0.02
Quadratic Discriminant Analysis	1.00	1.00	1.00	0.62 ± 0.04	0.55 ± 0.05	0.69 ± 0.04

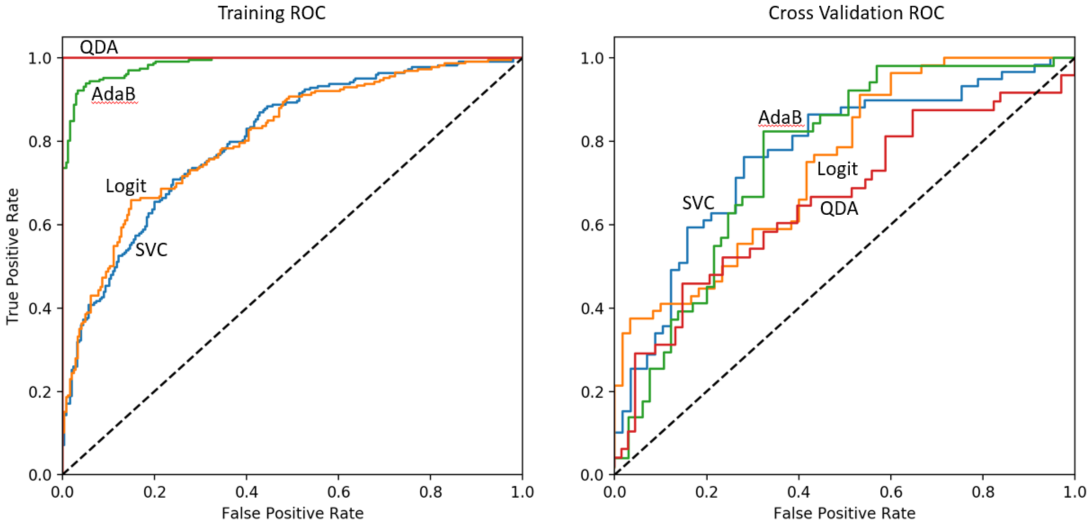


Figure 13: Training and Cross validation (1 of 5 folds) ROC curves for Support Vector Classification (SVC), Logistic Regression (Logit), Adaboost (AdaB) and Quadratic Discriminant Analysis (QDA).

Note that a direct one to one comparison between performance metrics is a little tricky since every paper uses a slightly different scale of attractive to plain looking ratings. Nevertheless admittedly our predictive algorithm lags behind some of these contemporary papers. This is largely explained by our corpus of subject pictures taken in "wild" i.e. picture not taken for prime purpose of attractiveness prediction modeling. In the section A.2.2, we report how well our predictive model does on other picture datasets. Our classification accuracy on lot more standardized dataset such as SCUT-FBP and Todorov Lab dataset is as high as 78% and 100% respectively.

A.2.2. Testing on external dataset

We want to verify that our attractiveness prediction model has picked up causal drivers of attractive looks, rather than alterable correlated features chosen by individual on their self posted profile picture. An example of correlated feature would be photograph quality, if attractive people are more likely to post high quality picture than the model may be predicting attractiveness ratings based on photograph quality rather than face structure or symmetry. This concern is reasonable because our training and validation is done on self posted professional profile pictures from the same platform, where individuals may be expected to have such systematically biased picture choice behavior. Therefore we test on dataset from experimental settings where individuals have no control on their face images. We collect images and labels from three different prior experiments – (i) Chicago Face Dataset (CFD) (ii) SCUT-FBP Dataset (iii) Todorov Lab Computer Generated. We want to ensure that the predictive model can learn from any of these datasets and predict on another. This would ensure that the predictive model can learn and predict attractiveness for any subject pictures (not just professional social network profile pictures). If we fail to do so, it may mean that our prediction is picking up some picture characteristics peculiar to how individuals post pictures on the professional social network. In this case we would not be able to trust the predicted attractiveness to be informative of how these individuals are perceived in the real world.

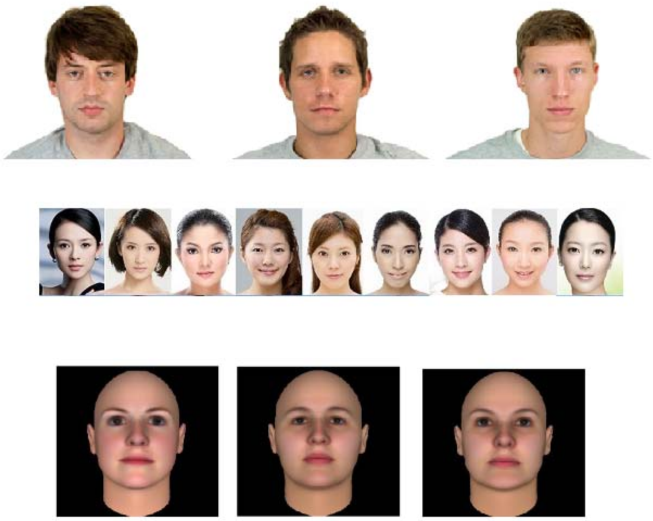


Figure 14: (Top) Chicago Face Dataset sample from Ma et al 2015, (Middle) SCUT-FBP sample from Xie et al 2015 and (Bottom) Todorov Lab computer generated face samples from Rojas et al 2011.

Figure 14 shows a small sample of face pictures from these three datasets. Note that the datasets capture face images in widely different settings. Chicago Face Dataset has high picture quality and standardized clothing and background. SCUT-FBP has a more variable setting but it predominantly captures Asian ethnicity faces. Todorov Lab faces are computer generated faces with fully standardized backgrounds, no hair, no skin blemishes or accessories. All of these independent datasets gather attractiveness ratings at different scales and in-fact very different distribution over the discrete rating options from human raters. We presume this is because of different definitions of discrete rating scale options. In order to standardize across datasets, we derive a simple attractive (1) and plain looking (0) binary classification for each dataset. Table 19 shows that model trained on our AMT dataset is able to correctly predict attractiveness, accuracy of 0.70,0.71,0.78 well above random 0.5, on all other datasets.

Table 19: Each cell is attractiveness prediction accuracy, with 0.5 being random prediction. The model is trained on the dataset in the rows and accuracy is tested on dataset in columns. A test accuracy of 0.78 means that the Support Vector Classification model predicted 78% of attractiveness labels on Todorov's dataset correctly and 22% incorrectly.

Train \ Test	AMT	CFD	SCUT-FBP	Todorov
AMT	**0.76**	**0.70**	**0.71**	**0.78**
CFD	0.76	0.73	0.66	0.97
SCUT-FBP	0.67	0.58	0.78	0.37
Todorov	0.62	0.57	0.46	1.00

A.3 CAAE Image Morphing

The CAAE produces an output X' (an entire image) for a profile picture X and a target age a. CAAE model learns the probability distribution $p(X'/X, a)$ i.e. age evolution path for face looks. For a new test face not seen during training, the model predicts an evolution path based on evolution seen for training faces that were similar to the new test face. It does so by creating a low-dimensional representation for an input face picture such that similar faces are close together in this low-dimensional space. Thus, the model has two building blocks - one used to find a low-dimensional latent representation z for an input image X i.e., $p(z/X)$ and the second used to create pictures at the target age a from this latent representation z i.e, $p(X'/z, a)$. This latent representation ensures that the image morphing is generalized to faces not seen during training. These two blocks are common in Computer Vision literature, i.e., Autoencoders (AE) and Generative Adversarial Networks (GAN, Goodfellow et al. 2014) respectively. We provide intuition behind these two building blocks and then describe how they are combined together in CAAE.

GAN: The GAN learns to "hallucinate" photo-realistic face pictures of a certain age a. The hallucinated images X' are generated by using the age a and white-noise random shock ϵ as inputs, i.e., $p(X'/\epsilon, a)$. A GAN has two components, namely, a generator (G) and a discriminator (D). The generator takes white noise as input. A single realization of white noise is a random shock ϵ. It outputs an image X'. The discriminator takes an image as an input and tries to distinguish whether the input is real (class $y = 1$) or fake ($y = 0$). Thus the generator hallucinates pictures; the discriminator tells if the image is "hallucinated" by the generator or its an actual image. The discriminator's job can be thought of as a person giving feedback on the quality of the hallucinated pictures. These models take millions of iterations: it is not feasible to use human feedback on millions of generated pictures.

Initially, both these components are terrible at their task. The generator parameters ϕ are initialized randomly and thus produce arbitrary outputs $X' = G_{\phi}(\epsilon)$ which look nothing like X. Even though generator creates pixel values that look nothing like a face, the discriminator cannot tell this garbage from actual face pictures. Over training iterations, the discriminator becomes better at discerning "hallucinated" pictures from actual images. The discriminator parameters θ are iteratively trained to classify real versus fake. The loss function pushes the discriminator to output $D_{\theta}(X) = 1$ for real and $D_{\theta}(X') = 0$ for fake inputs. Next, the generator parameters ϕ are trained to fool the discriminator into confusing fake samples for real. The generator loss function pushes $D_{\theta}(X')$ towards 1 for fake inputs. This discriminator and generator training steps are repeated numerous times. Thus the model's objective function pits the generator against the discriminator. Therefore, the generator also becomes truly proficient at producing near real images for any age. This method derives its power from the ability of the deep neural network layers of the generator and the discriminator to break down high-dimensional complex distributions into simpler building blocks. The shallow layers of the generator learn to create basic image pieces(e.g., edges, circles). The deeper layers combine together these basic pieces into more sophisticated forms (e.g., ears, hair, texture). We call pictures created by the GAN generator "hallucinations" because its random shock input ϵ means that pictures cannot be targeted to a specific individual. The arbitrary random shocks result in arbitrary (but real looking) face images.

$$J(\theta, \phi) = -\mathrm{E}_{x \sim p_{real}}[\log D_{\theta}(x)] \;\; + \mathrm{E}_{x \sim G_{\phi}}[\log 1 - D_{\theta}(x)] \qquad (14)$$

A conditional GAN used in CAAE allows the generator and discriminators to learn $G_{\phi}(\epsilon/a)$ and $D_{\theta}(X'/a)$ conditioned on age a respectively. While the images $X' \sim G_{\phi}(\epsilon/a)$ are now realistic for an age a, the random shock ϵ means that pictures can not be targeted to a specific individual. We need to replace these random shocks to be able to create more deterministic outputs.

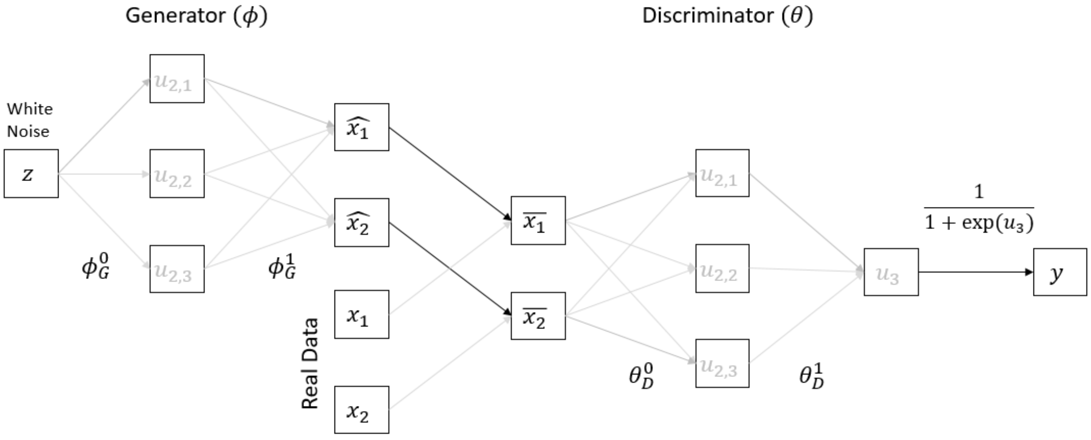


Figure 15: Generative Adversarial Network (GAN) architecture with a two layer generator and two layer discriminator.

AE: The AE block helps to create pictures targeted toward a specific individual. AE's are used to map a high dimensional (e.g. 640x480) image X onto a low dimensional embedding z (e.g 128-D) through a series of neuron layers called encoder ($X \to z$). Next this low dimensional embedding Z is mapped back to an output X' with the same dimensionality as the input through another series of convolutional filters called decoder ($z \to X'$). AE's do not need labels, instead they attempt to directly minimize the difference (or L2 loss) between their output X' and input X i.e., ($|X' - X| \to 0$). Intuitively, a low reconstruction loss means that the network has found encoder ($X \to z$) and decoder ($z \to X'$) parameters such that the low dimensional representation z has captured necessary information to reconstruct back the original image. This is similar to PCA which is used to find a small number of dimensions (principal components) that preserve the maximum variance in the original high dimensional space.

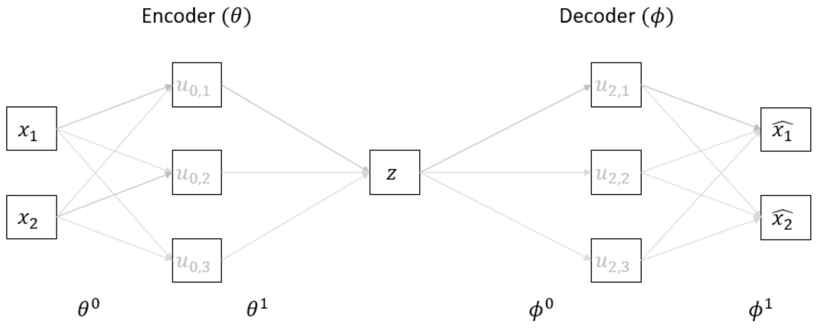


Figure 16: Autoencoders with a two layer encoder and a two layer decoder.

The CAAE model overall combines the AE and CGAN building blocks. More specifically it forces the decoder ($z \rightarrow X'$) on the AE module to also act as the generator $X' \sim G_{\phi}(\epsilon/a)$ for the GAN module. The latent representation z replaces the random shock ϵ. While the AE objective function forces the decoder output to be close to the input ($|X' - X| \rightarrow 0$), the output need not be realistic. The additional CGAN loss further forces the decoder to generate realistic ($D_{\theta}(X') = 1$) outputs. Once the entire CAAE network is trained it is able to produce realistic images for a new face image conditioned on any arbitrary age value.

A.4 Attractiveness Evolution Model

Our attractiveness evolution approach models an average decline of attractiveness over age ($\lambda_1 a$) and an additional heterogeneous component of decline among individuals ($\rho_i^1 a$). We first present some model free evidence to motivate this attractiveness evolution approach. Note that our original dataset only has access to a single face image per individual. We have access to two external dataset where individual face images are available at a sequence of age milestones. The CACD dataset [ref] contains 104,595 images for 1282 celebrities between age of 25 and 60. The FGNET dataset [ref] contains 217 images for 42 subjects between age of 21 and 34. The original FG NET dataset contains a large number of images for subjects at age below 21, but we exclude those. Figure 17 plots the average attractiveness of subjects at different age in three datasets. The first is our professional social network profile pictures, second is Cross Age Celebrity Dataset (CACD) [ref] and third is FG-NET dataset [ref]. As expected, the celebrity photographs in CACD dataset tend to be more attractive on an average. But all three datasets exhibit a decline in attractiveness over age which is central to our attractiveness evolution model.

$$beauty_{i,a} = \lambda_0 + (\lambda_1 + \rho_i^1)a + \rho_i^0 + e_{i,a} \tag{15}$$

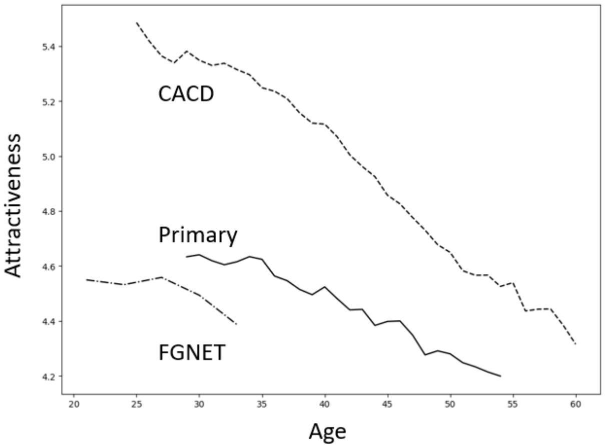


Figure 17: Average attractiveness of subjects at different age for three datasets – CACD, FGNET and our professional social network profile pictures.

Attractiveness of individual subjects may deviate from the average trend over age. This heterogeneity is not available to analyze in our professional social network profile pictures. Figure 18 explores this heterogeneity among four celebrities in the CACD dataset. For each of these celebrities the CACD dataset contains publicly available face images over 10 years. We average predicted attractiveness rating across all images of a celebrity at every available age. As an example we observe that Claire Danes has a higher attractiveness rating but declines more rapidly with age compared to Anthony Mackie. These examples also suggest that – (i) individuals attractiveness does not fluctuate wildly i.e. she is 1 std. dev above average at age 25, then 1 std. dev below average at age 26 and then average at age 27. (ii) individuals attractiveness may slowly evolve over 10 years from being above average to below average among their peers. (iii) while differences in picture characteristics (see Joan Cusack's picture samples) may seep into attractiveness prediction errors, these errors are still small relative to the std. dev. of attractiveness ratings.

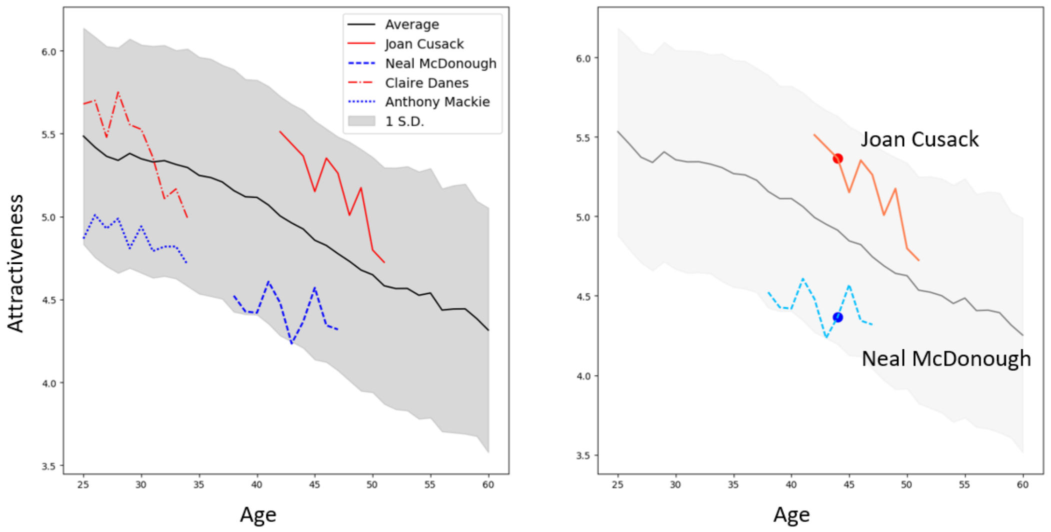


Figure 18: (Left) Attractiveness evolution of four celebrities compared against the average (± 1 std. dev.) attractiveness evolution. (Right) A single image is randomly chosen for every celebrity e.g. at age 44 years for Joan Cusack as the training sample. We will evaluate if the model can project the evolution across age 42 years to 52 years for Joan Cusack.

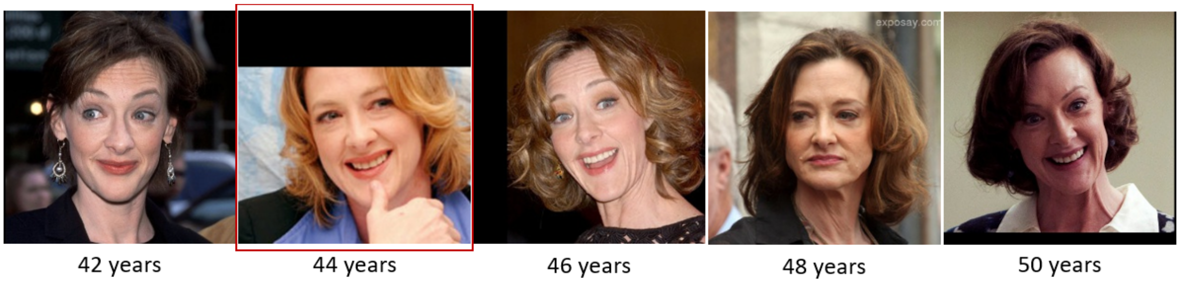


Figure 19: Joan Cusack's photograph at age 44 years randomly chosen into the training set, while others are left out for test performance evaluation.

We formally test our attractiveness evolution approach by first randomly sampling a single image for every celebrity e.g. snapshot at age 44 years for Joan Cusack. This acts as the training sample, which is akin to our professional social network profiles with one picture per individual. Next we predict the $\widehat{beauty}_{i,a}$ at all age milestones a using a single image for every celebrity i. Finally we compare the predicted $\widehat{beauty}_{i,a}$ with actual $beauty_{i,a'}$. The same steps are repeated for four candidate models. Table 20 describes the four candidates and *Figure 20* depicts how the first three are merely simplification of the final $beauty_{i,a}^{RHE}$ model in equation 3. *Figure 21* illustrates intuition on how these four candidate model differ in predicted attractiveness evolution.

$$beauty_{i,a} = \lambda_0 + (\lambda_1 + \rho_i^1)a + \rho_i^0 + e_{i,a} \tag{16}$$

Table 20*: Four candidate models for predictive individuals attractiveness at a sequence of age milestones.*

$beauty_{i,a}^{static}$	Attractiveness remain static over age.
$beauty_{i,a}^{Hom}$	Attractiveness evolves homogeneously over age. Homogeneous model ($\rho_i^1 = 0$) estimated using single picture per individual.
$beauty_{i,a}^{CAAE}$	Attractiveness predicted from morphed images.
$beauty_{i,a}^{RHE}$	Attractiveness evolves heterogeneously over age. Regularized Heterogeneous model estimated using multiple morphed picture per individual.

We report five performance metrics for each of the candidate models. Root mean square error (RMSE) and Mean Absolute Error (MAE) are standard for any prediction task. Note that $beauty_{i,a}^{CAAE}$ has a similar MAE to $beauty_{i,a}^{RHE}$ but much inferior RMSE. This is consistent with intuition illustrated in *Figure 21* that morphs and therefore direct attractiveness predictions on the morphs can have large noise at some age milestones. The regularized RHE model essentially removes such unjustifiable outliers. The remaining two measures inform us if the prediction have systematic bias. This is again consistent with the intuition

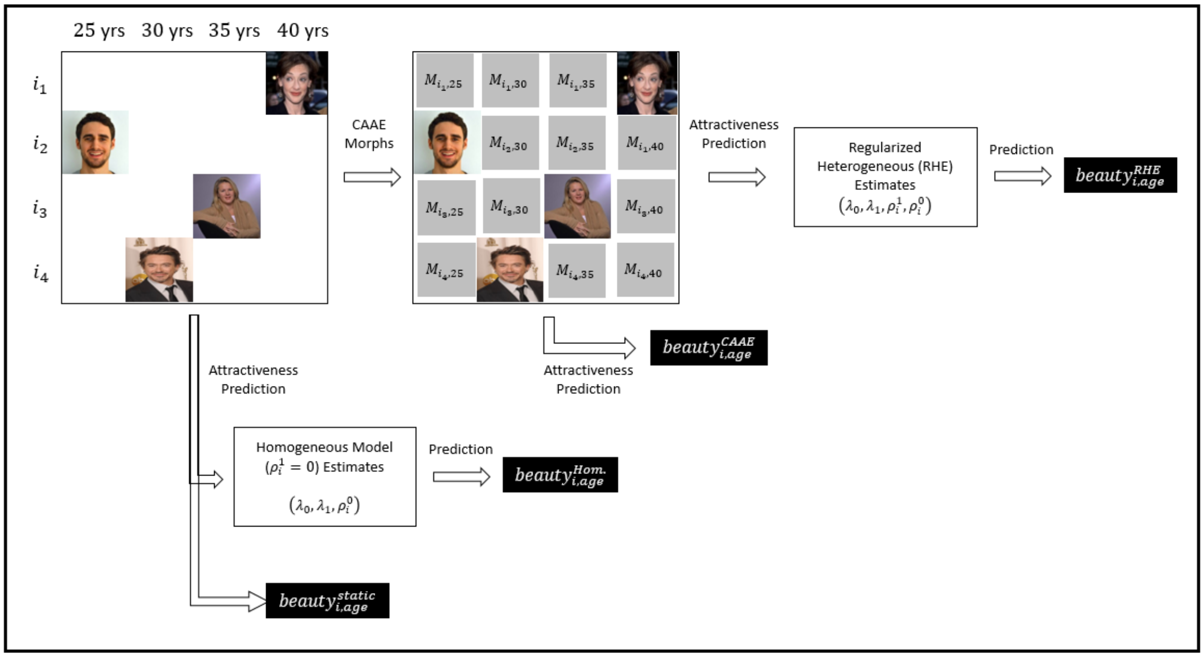


Figure 20: The $beauty_{i,a}^{RHE}$ model is our primary model. We first create morphed picture $M_{i,age}$ at all age milestones where actual picture is not available for individual i. Then RHE model is estimated using predicted attractiveness of morhped pictures. The other three candidates are simplification of this primary model. $beauty_{i,a}^{CAAE}$ doesn't use the RHE estimation. $beauty_{i,a}^{Hom}$ doesn't use image morphing at all. $beauty_{i,a}^{static}$ does not apply any attractiveness evolution at all.

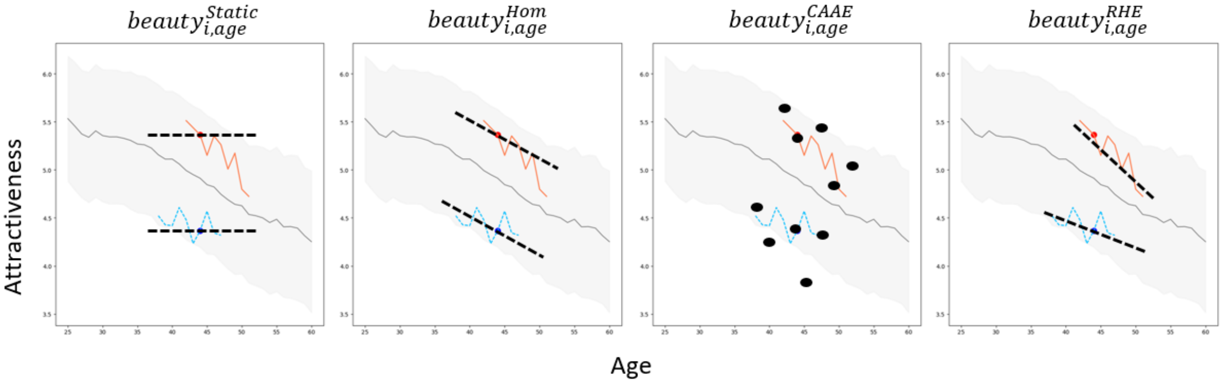


Figure 21: Comparison of actual attractiveness evolution of two individuals (Joan Cusack and Neal McDonough) against candidate model outcomes. $beauty_{i,a}^{static}$ is constant over age. $beauty_{i,a}^{Hom}$ has same slope over age for all individuals. $beauty_{i,a}^{CAAE}$ evolution is heterogeneous but morphs and therefore attractiveness predictions at some age milestones can have large noise. $beauty_{i,a}^{CAAE}$ has heterogeneous slope over age for all individuals. Note that these model predictions are hand picked for illustration only.

illustrated in *Figure 21* that static model will over predicted at higher age milestones and under predict at lower age milestones. Note that the fifth model is same as RHE, but using actual images instead of morphed images at sequence of age milestones. We can not use this model in practice because unlike the Celebrity CACD dataset we do not have access to actual images at sequence of age milestones. This is meant to be a reference point to evaluate how well the four viable candidate models perform. The RMSE for $beauty_{i,a}^{RHE}$ is 0.245 on a 1 to 7 scale.

$$rmse = \sqrt{E\left[\left(\widehat{beauty}_{i,a} - beauty_{i,a}\right)^2\right]} \quad ; \quad mae = E\left[\left|\widehat{beauty}_{i,a} - beauty_{i,a}\right|\right]$$

$$biasRight = E\left[\widehat{beauty}_{i,a} - beauty_{i,a} / a \geq 43\right] \quad ; \quad biasLeft = E\left[\widehat{beauty}_{i,a} - beauty_{i,a} / a \leq 42\right]$$

Table 21: *Performance metrics of four candidate models for predictive individuals attractiveness at a sequence of age milestones. The fifth model is same as RHE, but using actual images instead of morphed images at sequence of age milestones.*

Model	**RMSE**	**MAE**	**Bias Right**	**Bias Left**
$beauty_{i,a}^{static}$	0.307	0.206	0.040	-0.029
$beauty_{i,a}^{Hom}$	0.342	0.238	0.008	-0.003
$beauty_{i,a}^{CAAE}$	0.299	0.191	-0.003	0.002
$\boldsymbol{beauty_{i,a}^{RHE}}$	**0.245**	**0.183**	**-0.001**	**0.000**
$beauty_{i,a}^{RHE'}$ (Without morphing)	0.196	0.140	-0.000	0.000

A.5 Job Rank Model

In Section 4.3 we describe a three step procedure to assign job ranks. First, we define jobs as "title + employer size + industry". Second, we assign page ranks to N = 1000 most frequent jobs. Third, we extrapolate ranks to remaining infrequent jobs from ranked frequent jobs in the neighborhood. The neighborhood is defined as k=10 nearest jobs where every job is represented by tf-idf vocabulary size of v = 2000. This procedure makes very specific choices for hyper parameters (N = 1000, k = 10, v = 2000), job definition, ranking procedure and extrapolation procedure. We arrive at this specification by comparing performance of alternative specification on held out validation and test sets.

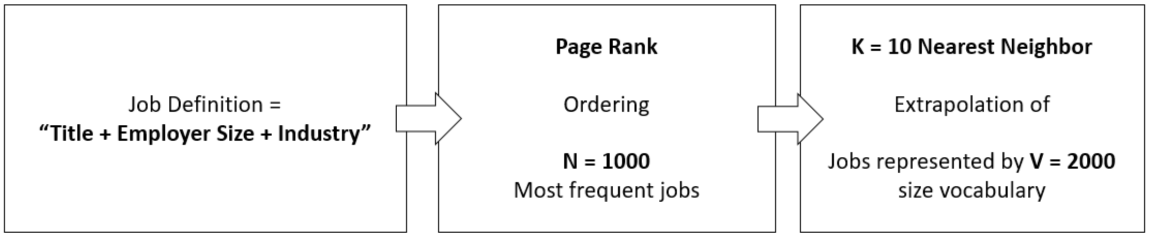


Figure 22: The final job desirability model. This is discovered after ruling out alternative choices for k, v, N, job definition, ranking procedure and extrapolation procedure.

Validation: We perform a traditional five-fold cross validation with 80% of job switches as training and remaining 20% as validation in every fold. We calculate the ratio of job switches in the validation set that go from lower-ranked job to a higher-ranked job. Our validation set ratio is 0.69 i.e. 69% of job switches go from lower-ranked job to a higher-ranked job. If the job ranks were uninformative (or random), this ratio would be 0.5. This validation is inherently limited because we do not know the ground truth on ratio of job switches that move from low desirability to high desirability jobs. While our job rank models ratio of 0.69 is better than an entirely random 0.5, the ground truth ratio is not known.

Test1: We collect dataset on job salaries from salarylist.com. This website reports large number of salaries for jobs across different companies, location and time periods. As an example it reported 901 salaries for the job title "Account Manager" across companies like Mu Sigma, IBM, Microsoft etc. Note that the website suggest that its data is self reported by companies and US Department of Labor, but we have limited transparency in confirming the data collection process. We pick out for each job title the median job salary. We discard job titles that are not available in our professional social network dataset or have less than 10 reported salaries on salarylist.com. We rank the remaining 817 job titles by median salary. We test if rank ordering by salary matches rank ordering by job ranks. The rationale being that a higher salary job is typically more desirable and therefore should have higher rank.

Test2: We collect dataset on job hierarchies from heirarchystructure.com. This website maintains a description of job hierarchies across a large number of domains such as – Financial Services, Management, Investment Banking, Corporate Finance, Advertising etc. The Investment Banking hierarchy consists of Managing Director, Director, Vice President, Associate and Analyst. We assign ranks to job title in each of these hierarchies. We discard hierarchies that do not have significant overlap with our professional social network dataset. We are left with 13 rank ordered job lists with every list consisting of 61 distinct jobs on an average. We test if rank ordering by hierarchy matches rank ordering by job ranks. The rationale being that a higher seniority job is typically more desirable and therefore should have higher rank. For the salary list and the 13 hierarchy lists combined, we get a spearman rank correlation coefficient of 0.44 and Kendall rank correlation coefficient of 0.36[16].

[16] The job hierarchies from heirarchystructure.com are manually crafted and limited to a single career path within an industry. Thus it is not very apt for cross industry switches observed in our dataset. Similarly, order by salary captures only one facet of a job's desirability. In-fact empirically observed switches over 100,000 MBA graduate profiles is perhaps a better approximation of average desirability then an industry specific hierarchy or salary data. We expect our desirability measure to be loosely correlated

Table 22: Model in the first row represents the final choice of Page Rank with k = 10, v = 2000, N = 1000 and job = "Title + Employer Size + Industry". The subsequent rows show validation/test evaluation with alternative choices for k, v, N, job definition, ranking procedure and extrapolation procedure.

Models	Held-out Validation	13 Hierarchy Lists + 1 Salary List	
		Spearman Correlation	Kendall Correlation
Page Rank	**0.69**	**0.44**	**0.36**
Page Rank (k=2)	0.64	0.34	0.31
Page Rank (k=50)	0.59	0.38	0.27
Page Rank (v = 1000)	0.68	0.41	0.24
Page Rank (v = 5000)	0.63	0.46	0.23
Page Rank (N = 100)	0.66	0.34	0.27
Page Rank (N = 5000)	0.71	0.33	0.37
Page Rank (Job = "Title")	0.63	0.31	0.30
Page Rank (Job = "Title + Employer Size")	0.64	0.33	0.25
Average Duration (instead of page rank)	0.58	0.24	0.22
Levenshtein Distance (instead of tf-idf)	0.54	0.25	0.22

We also report in Table 22 performance for alternative specifications –

(i) Alternative hyper parameters for the job rank calculations - k=2 or 50 instead of 10 for k-NN, N=100 or 5000 instead of 1000 for the top N frequent jobs, and tf-idf vocabulary size = 1000 or 5000 instead of 2000.

(ii) Alternative definitions of job - "title" or "title + employer size" instead of "title + employer size + industry".

(iii) Alternatives for ranking 1000 most frequent jobs - average duration to reach job instead of the page rank procedure.

(iv) Alternative for extrapolating ranking of 1000 most frequent jobs to remaining infrequent jobs - Levenshtein distance instead of tf-idf.

The search for good hyper parameters and model design is largely informal based on heuristics and subjective judgement.

with industry hierarchies or salary data but we do not expect a high degree of match. Therefore we report the correlation purely as a sense check.

A.6 Feature Calculations

Section 4 in the main text provided a list of features, their source (or calculation) and resulting values. Below are the details on how these features are specifically calculated.

A.6.1 Demographics

Gender: We use Microsoft Azure Cognitive Face API to identify individual's gender using their profile picture. While this service does not reveal exact algorithms used, our best guess would be a deep CNN trained on gender labels collected from human taggers. We manually verify at least 100 gender tags provided by this service. We find it to be accurate in 100 out of 100 cases.

Ethnicity: We use ethnicity classifier built by Ambekar et al (2009) which uses names to identify ethnicity. The possible ethnicities provided by the classifier are – GreaterEuropean, EastAsian, SouthAsian, NorthAfrican or WestAsian, GreaterAfrican, Others. A vast majority of profiles in our dataset belong to the first three classes. We re-categorize these outputs into three possible values – Greater European, East Asian and Others. Ambekar et al (2009) report validation F-scores of greater than 0.8 for easily discernible coarse classes such as GreaterEuropean and Asian. We also use Face Plus Plus API (https://www.faceplusplus.com/) to detect ethnicity from face images. While this service does not reveal exact algorithms used, once again our best guess would be a deep CNN trained on ethnicity labels collected from human taggers. We retrieve three possible values – Caucasian, Asian, Black and Others. Table 23 below provides the degree of match between name based and face based classification. We do not necessarily expect name based and face based ethnicities to match up since some ethnicities have more diverse looks and skin tone, while global migration has led to a degree of standardization of first names. While the match seems reasonable, we are unable to explain a large number of "East Asian" names corresponding to "Caucasian" face appearance. We have two reasons to collect these ethnicity tags – (i) attractiveness ratings may be systematically higher or lower for individuals whose face is closer to certain ethnicities, (ii) individual may face bias in their career because their name or face are perceived to belong to certain ethnicities. These goals are achieved by randomizing attractive and plain looking groups on both ethnicity tags.

Table 23: The percentage of profiles with every combination of name based and face based ethnicity tags.

	Face Based			
Name Based	White	Asian	Others	Black
Greater European	**73.5**	**14.3**	2.8	**9.3**

East Asian	**87.4**	4.7	2.7	5.0
Other	34.3	5.6	**28.5**	**31.3**

Age: We use year of undergraduate completion to calculate individual's age. We assume that everyone completes their undergraduate at the age of 22. Thus we find age at MBA graduation and all subsequent career milestones. We have no plausible reason nor empirical evidence to believe that this measure is systematically biased across attractive and plain looking individuals. We also use Microsoft Azure Cognitive Face API which predicts age for a face picture. This Face API has a large mean absolute error of 7.62 years (Dehgan et al 2017). Nevertheless a comparison between the two calculations serves as a sense check. The two methods, Face API and self reported UG graduation year, differ on a case by case basis by 0-10 years (less than 3 years more often than not). We calculate the number of profiles for each undergraduate year using the self-reported graduation year. We also calculate the same numbers using the current age from the Face API service on pictures. If the current age in 2017 is 40, we presume the individual completed undergraduate studies in 1999 = (2017 – 40 + *"22"*). We compare these two distributions. A visual analysis of these distributions (Figure 23) suggests a good match, with the face image based age prediction being more clustered in the middle, while UG graduation based measure more spread out. This is not surprising since the former, a deep learning based prediction which likely minimizes some form of squared error, tends to predict closer to the mean/median when uncertain.

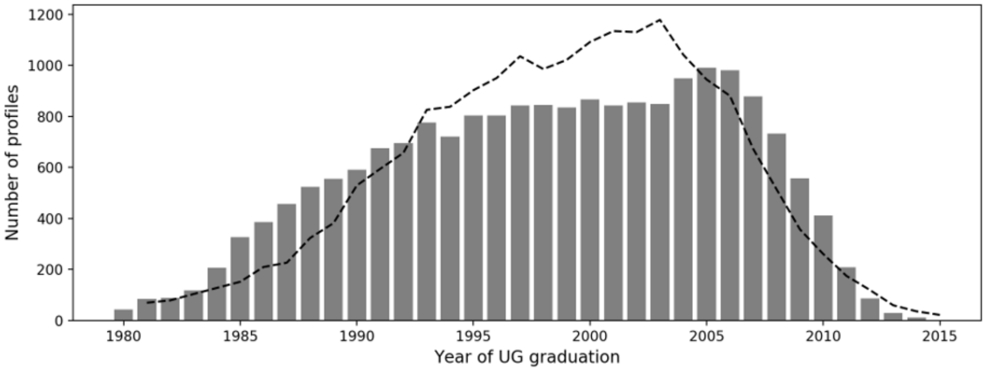


Figure 23: Year of undergraduate completion by (bar plot) self reported UG graduation and (dashed line) likely UG graduation based on profile picture age calculated as 2017 – "age prediction on current picture" + 22.0

We have two reasons to collect age tags – (i) attractiveness ratings are systematically lower for individuals with older facial looks, (ii) individual may face ageism positive/negative bias in their career compared to peers with same quality and type of experience and training. Randomizing attractive and plain looking groups on both age tags caters to both the issues. Now we will effectively be comparing performance of attractive and plain looking counterparts who appear to be of the same age by looks and who have the same UG graduation year.

A.6.2 Education

University Ranking: We use US News (www.usnews.com) to retrieve rankings for undergraduate and MBA education. We use National University ranking (up to ~1000) for assigning rank to an individuals undergraduate education. We use Business Program rankings for assigning rank to an individuals MBA. As an example, New York University is currently ranked 30th among National Universities while its MBA program at Stern school is ranked 19th. These rankings vary over time, so we collect university rankings for all years between 2008 and 2019 and MBA rankings for all years between 2001 and 2012. We show robustness of our results using alternative rankings years.

Table 24*: University undergraduate rankings for a sample of 10 universities from 2008 to 2019.*

University	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Harvard University	2	1	1	1	1	1	2	2	2	2	2	2
Stanford University	4	4	4	5	5	6	5	4	4	5	5	7
Princeton University	1	2	1	2	1	1	1	1	1	1	1	1
Yale University	3	3	3	3	3	3	3	3	3	3	3	3
MIT	7	4	4	7	5	6	7	7	7	7	5	3
Columbia University	9	8	8	4	4	4	4	4	4	5	5	3
University of Pennsylvania	5	6	4	5	5	8	7	8	9	8	8	8
Duke University	8	8	10	9	10	8	7	8	8	8	9	8
Dartmouth College	11	11	11	9	11	10	10	11	12	11	11	12
University of Notre Dame	19	18	20	19	19	17	18	16	18	15	18	18

Undergraduate Degree: We first identify education milestone corresponding to the MBA degree. This is done by searching for one of the terms "MBA", "M.B.A" or "masters of business administration" in the individuals self-reported degree. Next we classify education milestone prior to their MBA into undergraduate Science, Arts and Business. The Table 25 provides the search terms used to make this classification. Every self-reported education degree is searched for these terms. These search term lists are manually created to ensure we can classify more than 90% of undergraduate degrees into one of these categories.

Table 25: Education degree and corresponding search terms used for classification.

Degree	**Search Terms**
MBA	MBA, M.B.A, masters of business administration
UG Science	b.tech, btech, b. tech, b.s, b.e, BS, SB, s.b, BE, b. e

UG Arts	BA, b.a, AB, a.b
UG Business	bba, bsba, b.com, business

University Fees: We collect university undergraduate tuition fees and MBA program tuition fees from https://www.usnews.com. For public universities we average in-state and out-of-state tuition fees. MBA tuition fees have a smaller variation over and above the rankings, compared to UG tuition fees. The fees for most of the top 20 MBA programs that make up more than 50% of profiles in our sample are clustered above $50,000.

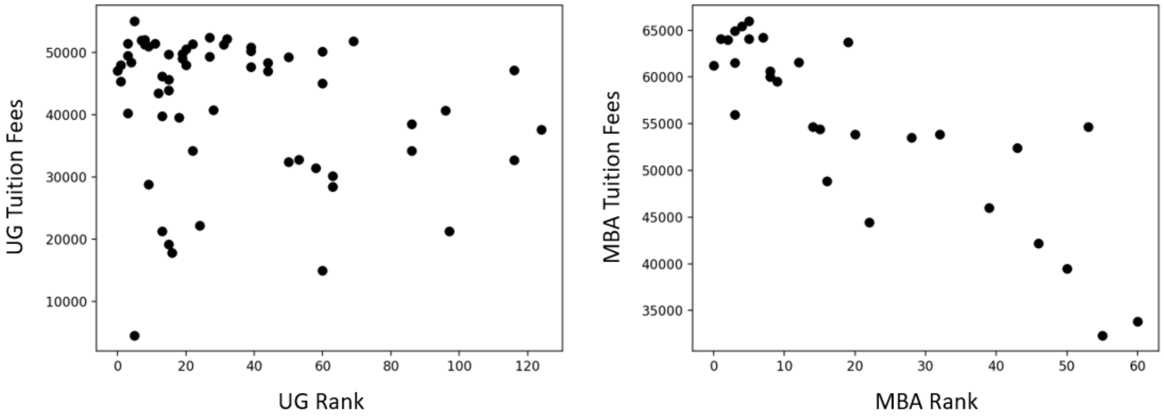


Figure 24: *(Left) A scatter of university fees against university ranking. (Right) A scatter of MBA fees against MBA ranking.*

A.6.3 Job Domain

Location Size: We determine location size endogenously. We first find a complete list of locations self-reported in profile career milestones. For each location we calculate the number of profiles where the location is mentioned at least once. A large location (e.g. New York, San Francisco) is mentioned by more profiles compared to a small location (e.g. Cleveland). The binary tags - large Location (1) and small Location (0) equally split all available profiles.

Employer Size: We determine employer size endogenously. We first find a complete list of employers self-reported in profile career milestones. For each employer we calculate the number of profiles where the employer is mentioned at least once. A large employer (e.g. Google, IBM, McKinsey, Citibank) is mentioned by more profiles compared to a small employer. The binary tags - large Employer (1) and small Employer (0) equally split all available profiles

We use this employer size in two ways. First, our main analysis uses a binary classification large Employer (1) and small Employer (0) as a control. Second, our job rank calculations using page rank use employer size. Remember that we defined job as combination of title, employer size and employer industry e.g. "CEO at very small internet". Here the employer size is split into five categories – "Very Small", "Small", "Medium", "Large", "Very Large". A fine grained employer size (say 10 categories) is more informative but it makes page rank calculations difficult because of sparse job switch observations. A coarse split (say 2 categories) is less informative but the resulting jobs have a large number of observed switches. The 5-way split was determined as an optimal hyper parameter setting. This was set to maximize the performance at held out validation set.

Industry Category: The professional social network uses 100s of industry names. An individual can self-report their job to belong to one of these available industry names. We manually[17] create 3 industry categories – IT, Finance and Management. The Table 26 below shows the mapping from industry to industry categories. These 3 industry categories cover more than 80% of all profile career milestones. All remaining industry names are mapped to Others. We use these industry categories in two ways – (i) In our main attractiveness bias analysis to control for career growth idiosyncrasies in specific industries (ii) In definition of a "job" during calculation of job ranks. We do this because same title "analyst" may mean something very different in Finance vs IT vs Management Consulting. As with employer size, a very spare or very coarse categorization creates a problem. This 4-way split works relatively well. This was chosen to maximize the performance of job ranks at held out validation set.

Table 26: Mapping table from industry category to industry

Industry Category	**Industry**
IT/Computers	Computer Software, Information Technology and Services, Internet, Telecommunications, Online Media, Computer Networking, Information Services, Computer & Network Security, Wireless, Computer Games, Computer Hardware, Consumer Electronics, Semiconductors, Electrical/Electronic Manufacturing, Market Research
Finance	Financial Services, Venture Capital & Private Equity,Accounting,Investment Management, Real Estate, Banking, Investment Banking,Insurance,Capital Markets, Commercial Real Estate, International Trade and Development
Management	Management Consulting, Marketing and Advertising, Nonprofit Organization Management, Education Management, Human Resources, Executive Office, Staffing and Recruiting,Hospitality,Government Administration, Public Relations and Communications, Broadcast Media, Program Development, Philanthropy, Outsourcing/Offshoring, Public Policy, International Affairs, Government Relations, Think Tanks, Religious Institutions, Civic & Social Organization,Fund-Raising,Import and Export, Political Organization

[17] This manual mapping is informed by a hierarchical clustering. This ML algorithm clusters together different industries with a large number of within cluster job transitions and a small number of across cluster job transitions.

Others	Hospital & Health Care, Consumer Goods,Retail,Pharmaceuticals,Higher Education, Entertainment, Medical Devices, Defense & Space,Research,Logistics and Supply Chain, Law Practice,Utilities,Health, Wellness and Fitness, Food & Beverages, Professional Training & Coaching etc.

JobType: We manually classify self-reported titles into one of four categories – "Technical", "Consultant", "Management" and "Others". Note that our research does not attempt to answer if attractiveness bias is more or less in certain jobs. We only use these to capture idiosyncrasies with respect to different career progression in different job types.

Table 27: Strings used to search self-reported titles for three job types.

Job Type	**Search String**
Technical	software, developer, engineer, scientist
Consultant	consultant, account
Management	manager, director, president, partner, ceo, founder, owner, chief

A.6.4 Picture Characteristics

We use the Microsoft Azure Face API and Face++ API (https://www.faceplusplus.com/) to extract Picture Quality (PQ) features, Face Characteristics (VC, IC) and health characteristics (H). These features largely cover features provided by major face image analysis tools including those offered by Microsoft, Amazon and Google. Collecting these wide range of features comes with the caveat that these face image analysis tools don't reveal their exact methodology. We believe that most of them are deep learning based predictive models trained on human labels. For example, human raters being asked to rate "stains on skin" or tag earrings. Its plausible that some of these features like Picture Quality may be algorithmically tagged without human labels. We re-scale these features to restrict the min and max values to 0 and 1 respectively (Table 13).

We would expect serious individuals to pay attention to desirable picture characteristics on their posted profile pictures and therefore these features are likely correlated among each other. In Table 29, we report presence of some of these correlations. It is also plausible that some of these characteristics seep into attractiveness ratings assigned by human raters and mimicked by our attractiveness prediction module. Therefore we also want to know which alterable features (PQ,VC, IC, H) are correlated with attractiveness ratings. We see that low picture blur, low noise and high exposure are associated with higher attractiveness ratings. Facial hair is associated with higher attractiveness ratings, likely indicating role of popular fashionable choices. Good face health and minimal face stains are associated with higher attractiveness ratings. Some correlations are not trivial to interpret. For example high face quality and high picture resolution are associated with lower attractiveness. We also see formal clothing associated with higher attractiveness for men but lower for women. Similarly, formal background associated with higher attractiveness for men but lower for women. We also see correlation between racy clothing and attractiveness ratings. Background and clothing are characteristics are not visible to the attractiveness raters or our attractiveness prediction. We believe that these correlation are likely spurious. For example individual wearing racy clothing, may also post higher quality picture with eye and lips makeup.

Table 28: Descriptive statistics for Picture Characteristics, all scaled between [0,1].

	Feature	% pictures where feature detected	mean	std. dev.	median
Photograph Quality PQ	resolution	100	0.6	0.32	0.58
	blur	100	0.16	0.19	0.1
	exposure	100	0.65	0.11	0.66
	noise	100	0.16	0.22	0.06
	roll	100	0.11	0.11	0.08
	pitch	100	0.11	0.09	0.09
	yaw	100	0.08	0.07	0.06
	face quality	100	0.76	0.31	0.91
Visible Characteristics VC	beard	100	0.12	0.14	0.1
	moustache	100	0.12	0.14	0.1
	sideburns	100	0.09	0.09	0.1
	eye makeup	100	0.3	0.46	0
	lip makeup	100	0.29	0.45	0
	smile	100	0.89	0.28	1
	glasses	100	0.85	0.36	1
Invisible Characteristics IC	necklace	9.7	0.71	0.12	0.21
	fashion accessory	13	0.74	0.14	0.24
	hat	0.3	0.8	0.15	0.32
	earrings	0.1	0.6	0.09	0.08
	bald	100	0.17	0.25	0.07
	racy clothing	100	0.01	0.02	0.01
Health H	dark circles	100	0.1	0.18	0.04
	skin health	100	0.15	0.2	0.06
	skin stain	100	0.22	0.26	0.11

Table 29*: Correlation among four category of picture characteristics (PQ, VC, IC, H) and attractiveness rating. We report and highlight a few strong and intuitively explainable correlations.*

	attractiveness	resolution	blur	exposure	noise	face quality
attractiveness	1					
resolution	-0.01	1				

blur	**-0.05**	0.11	1			
exposure	0.03	0.01	0.01	1		
noise	-0.02	-0.13	**0.33**	0.05	1	
face quality	-0.03	0.2	-0.14	0.04	-0.12	1

	attractiveness	beard	moustache	sideburns	eyeMakeup	lipMakeup
attractiveness	1					
beard	**0.13**	1				
moustache	**0.13**	**0.88**	1			
sideburns	**0.13**	0.71	0.73	1		
eyeMakeup	0.02	-0.37	-0.37	**-0.44**	1	
lipMakeup	0.02	-0.32	-0.31	-0.39	**0.56**	1

	attractiveness	dark circle	skin health	skin stain
attractiveness	1			
dark circle	0	1		
skin health	**0.1**	-0.2	1	
skin stain	**-0.1**	0.17	**-0.37**	1

	attractiveness	bald	necklace	hat	earrings	racy clothing
attractiveness	1					
bald	0.05	1				
necklace	-0.03	0.04	1			
hat	0	-0.01	0	1		
earrings	**-0.2**	0.03	**0.57**	0	1	
racy clothing	0.01	0.06	0.03	-0.11	-0.04	1

There is a reasonable concern that serious individuals who pay attention to desirable picture characteristics are also likely to be more successful in job market. Thus association of attractiveness pictures with job market success is purely coincidental. In order to resolve this concern, we match attractive and plain looking groups on all picture characteristics. We would be comparing attractive and their plain looking peers who have posted pictures with similar characteristics and thus no systematic difference in attention to profile pictures. This lends to a natural follow up question - How much explanatory power (r-squared) do all these picture characteristics (PQ,VC, IC, H) combined have for attractiveness ratings? If these picture characteristics largely explain the attractiveness ratings, then the profile pictures don't really contain a useful signal of individuals un-alterable attractiveness.

The Table 30 reports explanatory power (r-squared) of picture characteristics using an Ordinary Least Square (OLS) and Support Vector Regression (SVR) models. We also report r-squared on a five fold held out cross validation. We use different set features –picture characteristics visible to attractiveness ratings (PQ, VC, H), and all picture characteristics (PQ, VC, H, IC). The addition of cropped out characteristics (IC) invisible to our raters or attractiveness prediction model does not add much to explanatory power. Explanatory power of less than 15% suggest that the pictures do contain significant useful signal of individuals un-alterable attractiveness that are not trivially altered by adjusting picture characteristics alone.

Table 30*: Training and validation R-squared of OLS and Support Vector Regression to predict attractiveness ratings from picture characteristics.*

	OLS	SVR Training	SVR 5-fold cross validation
$beauty = f(PQ, VC, H, IC)$	13.0%	14.7%	$\mathbf{13.2} \pm 2.6\%$
$beauty = f(PQ, VC, H)$	12.7%	14.5%	$\mathbf{13.2} \pm 2.5\%$

A.6.5 Other Text Cleaning

We deal with numerous self-reported free text fields – school name, degree name, employer name etc. This creates two problems – (i) mis-spellings because of the free text nature, (ii) usage of different terms with otherwise synonymous meaning. For example MBA school names – “University of Pennsylvania”, “Wharton School” and “Whrton School” should all be treated the same. We use numerous small tools for data cleaning, a couple of examples include - Levenshtein distances to match employer and school name strings, and autocorrect software packages on degree and title strings.

A.7 Propensity Score Matching for Model Free Evidence

We use propensity score matching for model free comparison between career progression of attractive and plain looking groups. We perform PSM to match the two groups on pre MBA graduation characteristics – Undergraduate area of study, Undergraduate university ranking, Year of undergraduate completion, Length of Pre MBA experience, MBA School and MBA rank. Individuals are classified into the two groups using their attractiveness at the time of graduation ($t = 0$). $pScore_{i,t}(\phi)$ parameterized by ϕ denotes individual i’s propensity to be assigned in attractive group instead of the plain looking group. We find a plain looking ($b_{j,t} = 0$) counterpart for every attractive individual ($b_{i,t} = 1$) with closest propensity score $pScore_{i,t}(\phi) \simeq pScore_{j,t}(\phi)$. We use nearest neighbor matching with a caliper width at 0.01, matching ratio at 1 and without replacement for matching. The dataset contained 23,638 individual profiles initially, that were reduced to 19,075 after matching. The Table 31 provides the distribution of characteristics $\overrightarrow{E_i}$ in the two groups before and after matching.

$$pScore_{i,t}(\phi) = \frac{1}{1 + \exp(-\vec{E}_i * \boldsymbol{\phi})} \tag{17}$$

$$E_i = (ugRank_i, ugArea_i, preMbaExperience_i, mbaRank_i, mbaSchool_i, graduationYear_i)$$

Table 31: Propensity Score Matching on pre MBA characteristics for Male Profiles

	Before Matching		**After Matching**	
	Attractive	**Plain looking**	**Attractive**	**Plain looking**
Number of Profiles	8511	8461	7069	7089
Undergraduate Rank	389	435.6	413.5	396.8
MBA Program Rank	14.8	15.3	14.7	14.5
Year of Graduation	2003.4	2003.1	2003	2003.6
Undergraduate Degree (Arts,Business,Science)	18.8%,5.6%, 75.5%	14.5%,5.3%, 80.1%	16.53%,5.6%, 77.8%	16.04%,5.45%, 78.4%

Table 32: Propensity Score Matching on pre MBA characteristics for Female Profiles

	Before Matching		**After Matching**	
	Attractive	**Plain looking**	**Attractive**	**Plain looking**
Number of Profiles	3401	3265	2472	2445
Undergraduate Rank	342.2	391.7	361.8	352.7
MBA Program Rank	14.4	14.9	13.7	13.8
Year of Graduation	2004.5	2003.7	2004.0	2004.0
Undergraduate Degree (Arts,Business,Science)	34.1%,9.3%, 56.5%	33.2%,8.1%, 58.6%	33.4%,8.9%, 57.6%	33.4%,8.5%, 58.0%

A.8 Propensity Score Matching for Main Econometric Models

We create propensity score matched (PSM) sample of attractive and plain looking groups both for our dynamic bias model that discerns preference and beliefs biases as well as lifetime models that studies attractiveness premium over 15 years. Table 33 report the percentage standardized bias for all matching covariates $(D_i, E_i, JD_{it}, PQ_i, VC_i, H_i)$ before and after matching. The largest percentage standardized bias is 1.35% and 3.94% for the dynamic and lifetime models respectively. Note that for categorical variables (e.g., job type) we create $(L-1)$ binary variables corresponding to L categories for PS model and %Std bias calculations. Categories which accounts for less than say 1% observations are typically combined into an “Other” category. We reports at most three largest categories in the Table.

Table 33: Percentage Standardized Bias for all matching covariates before and after matching. Dynamic Bias Model is used to discern preference versus belief bias. Lifetime Bias Model is used to measure attractiveness bias at 15 years career outcome.

	Dynamic Bias Model		**Lifetime Bias Model**	
Covariates	**Before Matching**	**After Matching**	**Before Matching**	**After Matching**
Rank Previous	2.63	-0.06	n.a.	n.a.
Large Location	1.42	0.59	0.95	2.45
Large Employer	2.13	0.12	-1.75	0.17
Job Type (Management)	-2.39	1.35	4.22	3.38
Job Type (Others)	4.06	-0.52	-0.46	-0.88
Job Type (Consultant)	2.85	-0.18	-2.42	-0.74
Industry Category (IT)	-8.47	-0.41	-5.39	-2.17
Industry Category (Management)	4.73	-0.29	-0.84	-0.19
Industry Category (Others)	-0.23	0.52	7.78	3.94
Ethnicity (Eu)	16.9	-0.45	14.02	0.09
Ethnicity (Other)	-23.23	0.62	-20.28	-0.72
Ethnicity (As)	5.93	-0.22	5.29	1
Face Ethnicity (Asian)	-61.01	-1.95	-52.18	-1.45
Face Ethnicity (Black)	0.23	0.52	1.89	0.26
Face Ethnicity (Other)	-20.05	1.38	-16.48	0.94
MBA Graduation Year	7.98	0.57	11.14	0.49
UG Graduation Year	16.51	0.19	20.17	1.3
UG Graduation Year (Unknown)	-12.94	0.14	-12.71	-0.7
Face Age	-35.38	-0.29	-46.82	-0.06
UG Degree (Science)	-18.08	0.26	-15.29	-0.13
UG Degree (Others)	16.42	-0.27	16.4	0.85
UG Degree (Arts)	1.75	0.09	-2.16	-0.78
UG Rank	-0.21	0.03	0.96	0.69
UG Rank Unknown	-16.42	0.27	-16.4	-0.85
UG School (Others)	1.05	0.06	-0.75	1.54
UG School (Stanford)	0.4	-0.46	0.59	-1.75
UG School (Cornell)	0.88	0.29	1.69	0.27
MBA Rank	0.39	-0.23	0	2.06
MBA School (Others)	4.21	-0.06	5.68	1.85
MBA School (Kenan-Flagler)	-2.25	0.16	1.11	-2.55
MBA School (Harvard)	2.41	-0.18	-3.69	-1.1
Blur	-10.45	0.31	-8.78	1.52
Exposure	3.48	0.13	3.85	-1.01
Noise	-3.88	0.1	-3.47	-0.17
Roll	2.82	0.29	2.1	-0.51
Pitch	15.91	0.23	16.46	0.05
Yaw	6.93	0.35	5.12	0.36
Face Quality	-5.37	-0.21	-3.57	-1.52
Resolution	-0.86	-0.1	-1.94	-0.4
Lip Makeup	10.61	-0.13	0.31	-0.19
Eye Makeup	13.63	-0.01	1.11	-0.6
Bald	4.06	0.19	8.83	-0.83

Beard	15.76	0.31	18.23	1.56
Moustache	15.87	0.13	18.34	1.4
Side Burns	14.63	0.46	18.39	1.61
Dark Circles	1.51	0.4	-0.01	0.16
Face Health	15.48	-0.25	14.43	-1.37
Face Stains	-14.64	0.11	-16.23	-0.29
UG Fees	-0.03	0.14	1.06	0.19
MBA Fees	1.78	-0.14	2.94	-0.82

A.9 Alternative PSM Design

We discuss PSM model specification choices in Section 5.2. We show here alternative candidates for the propensity score model and the matching. In our main analysis we use PSM for the dynamic preference vs belief bias model and lifetime bias model. We only report below performance of candidate models on matching for the lifetime bias model. The relative performance of candidate models are largely similar for the dynamic preference vs belief bias model as well.

A.9.1 Propensity Score Models

We use a logistic propensity score model $pScore_{i,t}(\phi)$ with linear covariate effects (equation 5). Instead of the linear effects we could potentially include squared terms as well as interaction effects (equation 6) to attain greater fit. Such a specification is likely to attain greater fit (high training accuracy) at the cost of lower generalizability (low validation accuracy). We also evaluate alternative choices for $pScore_{i,t}(\phi)$ namely – Logistic with small regularization penalty (C = 100), Logistic with large regularization penalty (C = 0.01), Support Vector classification and Gradient Boosting classifier. The Table 34 reports the evaluation results. We evaluate the alternatives using five-fold cross validation on a held out sample.

$$pScore_{i,t}(\phi) = \frac{1}{1 + \exp(-[D_i, E_i, JD_i, PQ_i, VC_i, H_i] * \boldsymbol{\phi})} \tag{18}$$

$$pScore_{i,t}(\phi) = \frac{1}{1 + \exp(-[D_i, E_i, JD_i, PQ_i, VC_i, H_i] * [D_i, E_i, JD_i, PQ_i, VC_i, H_i] * \boldsymbol{\phi})} \tag{19}$$

Table 34: Training and validation accuracies of candidate propensity score models. The Logistic Model with low regularization (C = 100) and linear covariates produces the highest validation accuracy.

$e_{i,t}(\phi)$	$l_{i,t}(\phi)$	Training Accuracy	Cross-Validation Accuracy
Only Linear Effects	Logistic (C = 0.01)	67.47	68.74
All Interaction Effects	Logistic (C = 0.001)	67.62	66.95
Only Linear Effects	**Logistic (C = 100.0)**	**69.32**	**69.01**

All Interaction Effects	Logistic (C = 1000.0)	69.84	68.27
Only Linear Effects	SVC (C = 5)	69.84	68.30
Only Linear Effects	Gradient Boosting (GBC)	71.19	68.20

A.9.2 Matching

We use nearest neighbor matching of treated units with un-treated ones without replacement using a caliper width of 0.01. We could potentially match with replacement. We could use alternative caliper widths. We could use exact or optimal match instead of nearest neighbor match. These alternatives matching choices need to be evaluated based on the following tradeoff – minimal bias in matched samples while keeping the matched sample large enough for successful estimation. We report in Table 35 mean percentage standardized bias, Rubin's R, Rubin's B and size of matched sample as a percentage of total sample.

Table 35: Mean percentage standardized bias, Rubin's R, Rubin's B and matched sample size before and after matching. A Rubin's R between 0.5-2.0 and a Rubin's B less than 25% indicate a reasonable match.

Before Matching	8.88%	0.88	93.65%	100%
	Mean % Std. Bias	Rubin's R	Rubin's B	Matched Samples
Without replacement Caliper Width = 0.01	**1.12%**	**1.01**	**0.57%**	**62.3%**
Caliper Width = 0.1	1.37%	1.05	7.55%	65.7%
Caliper Width = 1.0	6.54%	1.08	65.45%	85.3%
With replacement	3.44%	1.03	32.58%	66.8%

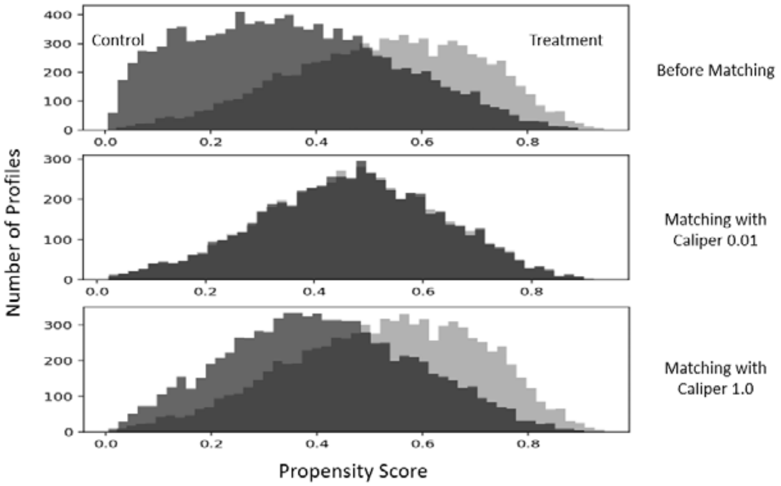


Figure 25: Histogram of treatment and control group propensity scores before and after matching. Rubin'B captures the difference between the means of the treatment and control propensity score distributions. Rubin'R captures the ratio of variance of the treatment and control propensity score distributions.

A.10 Alternative Regression Specification

A.10.1 Stratified treatment effects

In our main analysis, we estimate a single attractiveness bias (or single preference and belief bias) across the entire matched sample. Alternatively, we can get an attractiveness bias estimate for different propensity score stratas (Imbens 2015, Eckles and Bakshy 2017). We divide the propensity scores (*Figure 25* Top Plot) into four contiguous stratas (pScore [0-0.25], [0.25-0.45], [0.45-0.6], [0.6-1])[18]. Then we match attractive and plain looking individuals in each strata. The motivation is to study if the attractiveness bias differs significantly across the covariate space. The first strata (pScore 0 to 0.25) compared to fourth strata (pScore 0.6 to 1) represents individuals who have low propensity of being assigned the attractiveness treatment. Stratified results will show if attractive individuals in this strata have significantly different premium. This is important because the support of propensity scores for attractive and plain looking group is quite different before matching. We want to be certain that attractiveness bias results do not increase or decrease with higher propensity of being attractive. *Table 36* Model 1 shows interaction effect of attractiveness with strata to be statistically insignificant.

[Insert Table Here – Refer End of Appendix]

Table 36: (Model 1) The impact of attractiveness bias across four strata with strata indicator interaction effect. (Model 1 to 4) The impact of attractiveness bias in one propensity score strata at a time..

A.10.2 Attractiveness Measure

In our main analysis, we use binary classification for attractiveness ($b_{i,t}$). We report robustness of our results for alternative measures of attractiveness - continuous rating scale instead of binary classes ($beauty_{i,t}$), static measure as on MBA graduation ($beauty_{i,t=0}$), and static measure as at data collection in 2017 ($beauty_{i,t=2017-mbaGradYear}$). We find the impact of attractiveness on lifetime career success to be positive and significant.

[Insert Table Here – Refer End of Appendix]

Table 37: The impact of attractiveness bias using different measures for attractiveness - (Model 1) binary classification ($b_{i,t}$), (Model 2) continuous measure ($beauty_{i,t}$), (Model 3) static measure as on MBA graduation ($beauty_{i,t=0}$), (Model 4) static measure as at data collection in 2017.

[18] The strata boundaries are chosen to attain reasonable sample sizes.

A.10.3 Job Rank Measure

In our main analysis, we standardize the job ranks for all individuals i whose job transition is observed between $t \rightarrow t + \Delta t$ to have zero mean and unit standard deviation. We report robustness of our results for alternative unstandardised job ranks ($rank_{i,t}$).

[Insert Table Here – Refer End of Appendix]

Table 38: The impact of attractiveness preference and belief bias using (Model 1) standardized job ranks ($r_{i,t}$), (Model 2) unstandardized job ranks ($rank_{i,t}$). The impact of attractiveness on career outcome at the end of 15 years using (Model 3) standardised job ranks ($r_{i,t}$), (Model 4) unstandardised job ranks ($rank_{i,t}$).

We also construct an alternative model that does not rely on job rank calculations. Our profile data captures jobs that are self-reported by individuals in free text form. This creates an extremely large and sparse list of jobs. While a handful of job titles (e.g. Software Developer, Associate Consultant) are reported frequently, most job titles are somewhat customized by individuals (e.g. Principal consultant in Energy and Oil). One approach could be to compare individuals on time taken to achieve the same job. Attractiveness bias can be established if attractive individuals systematically take a shorter duration. θ_j represents job specific effects i.e. average time to reach the specific job.

$$\Delta t_{i,j} = \theta_0 + \theta_j + \theta_2 b_i + \overrightarrow{\theta_4} Z_i + u_{i,j} \tag{20}$$

Table 39: A small sample of very frequent jobs in Management Consulting

Job	Number of individual
Associate at McKinsey and Company	235
Case Team Leader at Bain and Company	145
Consultant at The Boston Consulting Group	315
Manager at Deloitte	108

We observe a total of 59,854 distinct jobs in our dataset. The number of jobs are extremely large, with a vast majority of jobs not providing sufficient variation across individual characteristics (beauty, gender etc.). We can focus only on frequently observed jobs, say jobs self-reported by at least 20 individuals. We are left with 209 jobs and 23,443 career milestones. This model finds attractiveness to have a significant and negative impact on time taken. A coefficient of -0.268 means that attractive individual attains the same job approximately quarter year faster than plain looking individuals. Alternatively we could also model θ_j as an unobserved random effects. As expected, we find (model 2) beauty to have a significant and negative impact on time taken. This establishes that our attractiveness bias finding is robust with and without the use of page rank procedure.

We can theoretically extend this method to identify belief and preference bias. We can model dependent variable $\Delta t_{i,j1,j2}$ to represent time taken by individual i to move from job $j1$ to job $j2$. We can split β_2 to (β_2^B, β_2^P). Here the belief bias β_2^B plays a role if the transition $(j_1 \rightarrow j_2)$ happens during the first 6 years. The $\beta_{j1,j2}$ captures fixed effects specific to the $j1$ to $j2$ transition. This would allow us to compare time taken by different individuals for the same job transition.

$$\Delta t_{i,j1,j2} = \beta_0 + \beta_{j1,j2} + \beta_2 b_i + \overrightarrow{\beta_4} \mathrm{Z}_i + u_{i,j1,j2} \tag{21}$$

$$\Delta t_{i,j1,j2} = \beta_0 + \beta_{j1,j2} + \mathbf{1}(t \leq 6)\beta_2^B b_i + \beta_2^P b_i + \overrightarrow{\beta_4} \mathrm{Z}_i + u_{i,j1,j2} \tag{22}$$

Table 40: A small sample of very frequent jobs transitions in Management Consulting

Job From	Job To	Num of individual
Associate at McKinsey and Co.	Engagement Manager at McKinsey and Co.	94
Consultant at Bain and Company	Case Team Leader at Bain and Company	61
Consultant at BCG	Project Leader at BCG	94
Manager at Deloitte	Senior Manager at Deloitte	12

Unfortunately we observe 75,669 unique job transitions in our dataset. Further, restricting to job transitions reported by at least 20 individuals, we are left with 36 transitions and 1082 career milestones. Not surprisingly, all estimates are insignificant. This model fails to utilize more than 90% of all job transitions. The lack of generalization results in large standard errors for key parameters $(\beta_2, \beta_2^B, \beta_2^P)$ we are interested in. We also unsuccessfully attempt a random effects model for $\beta_{j1,j2}$. This limitation is why we use page rank procedure to calculate a create a ranked list of otherwise dissimilar jobs.

[Insert Table Here – Refer End of Appendix]

Table 41: (Model 1) Time Taken to attain job j with θ_j fixed effects for all jobs reported by at least 20 individuals, (Model 2) Time Taken to attain job j with $\beta_j$random effect with all jobs reported by at least 2 individuals, (Model 3) Time Taken in job transition $j_1 \rightarrow j_2$ with $\beta_{j1,j2}$ random effects with all jobs reported by at least 2 individuals.

A.10.4 Career Stage Cut-off

In our main analysis, we use 6 years as an early career cut off. The underlying assumption being that a period of 6 years is more than sufficient time duration for prior beliefs to be overcome by actual performance signals. Belief-based attractiveness rewards should vanish after this point. We report robustness of our results with alternative cut off points – 4 years and 8 years.

[Insert Table Here – Refer End of Appendix]

Table 42*: The impact of attractiveness preference and belief bias with different early career cutoff for belief bias - (Model 1) original cut off at 6 years, (Model 2) cut off at 4 years (Model 3) cut off at 8 years.*

A.10.5 University Rankings

In our main analysis, we use university rankings and MBA rankings from 2017. We report robustness of our results with alternatives – undergraduate university rankings from 2010, MBA rankings from 2001 and MBA rankings from 2012.

[Insert Table Here – Refer End of Appendix]

Table 43*: The impact of attractiveness preference and belief bias using different university rankings - (Model 1) original UG and MBA ranks (Model 2) UG rank from 2010, (Model 3) MBA rank from 2001, and (Model 4) MBA rank from 2012.*

A.10.6 Alternative Scope

In our main analysis, we collect MBA graduates from 1990 and 2015. This provides confidence that the preference and belief bias findings are perpetual over last few decades. We report robustness of our results with alternative sets of – recent graduates between 1998 and 2015 and older graduates between 1990 and 2008.

[Insert Table Here – Refer End of Appendix]

Table 44*: The impact of attractiveness preference and belief bias using (Model 1) all MBA graduates, (Model 2) MBA graduates after 1998, (Model 3) MBA graduates prior to 2008, (Model 4) MBA Rank less than 20, and (Model 5) MBA Rank less than 100.*

In our data collection, we target top 100 MBA programs. But due to opaque search functionality on the professional social network we retrieve profiles outside of top 100 MBA programs. We want to be certain that lower tier MBA programs where individuals may have very different career goals and prospects are not driving the results. We report robustness of our results with alternative sets of profiles – graduates between from top 20 MBA programs and graduates from top 100 MBA programs.

A.10.7 Addition Controls

In our main analysis, we match attractive and plain looking groups on various potential confounders. A large set of these are derived from the profile picture. After the groups are matched on profile picture characteristics, we do not use these as controls in the regression model. This is motivated by significant degree of correlation among these variables. We report below the variance inflation factors (VIF) if all possible variables where included. Different researchers have suggested a VIF greater than 2 or greater than 5 to be a matter of concern. With the risk of correlations in mind, we report robustness of our results with all controls thrown in together.

Table 45*: Variance Inflation Factors when largest set of controls are added in. Attractiveness is almost uncorrelated because of propensity score matching on most of the remaining features.*

Features	Attractiveness	Gender	Ethnicity (Eu)	Ethnicity (Other)	Face Ethnicity (Asian)	Face Ethnicity (Black)	Face Ethnicity (Other)
VIF	1.02	**5.14**	**4.07**	**4.33**	1.14	1.25	1.24

Features	Job Type (Management)	Job Type (Other)	Job Type (Technical)	Industry (IT)	Industry (Management)	Industry (Other)	Large Employer	Large Location
VIF	3.04	2.92	1.47	1.81	1.71	1.63	1.05	1.05

Features	MBA Rank	UG Rank	MBA Grad Year	UG Grad Year	UG Fees	MBA Fees	UG Degree (Business)	UG Degree (Science)	UG Degree (Other)
VIF	1.18	1.62	**4.42**	**4.57**	1.11	1.03	1.4	2.51	3.04

Features	blur	exposure	noise	roll	pitch	yaw	face Quality	resolution
VIF	1.29	1.1	1.18	1.11	1.03	1.45	1.60	1.18

Features	lip Makeup	eye Makeup	bald	beard	moustache	sideburns	dark circle	health	stain	Eye Glasses	Lipstick
VIF	1.88	2.74	1.15	**5.12**	**5.15**	2.65	1.11	1.47	1.25	1.18	1.24

Features	Smile	Racy Clothing	Formal Background	Informal Background	Formal Clothing	Informal Clothing	Necklace	Hat	Earrings
VIF	1.1	1.07	1.45	1.05	1.88	1.01	1.6	1.02	1.03

[Insert Table Here – Refer End of Appendix]

Table 46*: The impact of attractiveness preference and belief bias with additional controls - (Model 1) Original set of controls (Model 2) Picture quality controls (PQ_i), (Model 3) Picture quality, visible face characteristic (AVC_i) and face health controls (H_i). (Model 4) Picture quality, facial hair, face health controls and picture characteristics from outside of cropped area (AIC_i).*

A.11 Robustness Checks

Our primary dataset (43,533 profiles) suffers from sample selection issue wherein MBA graduates may be systematically missing from our data collection. We collect profiles in 2017 for graduates between 1990 and 2015. Every individual in the original graduating class lies in one of four mutually exhaustive sets - (a) she drops out of the job market prior to data collection in 2017 i.e. they either never create a profile or delete their profile prior to the data collection period. (b) she has an active profile but our search criterion does not successfully find their profiles. (c) she has an active profile but do not have a picture. (d) she has an active profile with a profile picture. The first three sets (a∪b∪c) are together censored profiles. Note that while set (b) is an idiosyncratic result of our data collection methodology, set (a) and (c) are based on individuals choice. This drop out choice for individuals can be modeled as follows.

$$d_{i,t}^{*} = \delta_0 + \delta_1 r_{i,t-1} + \delta_2 b_{i,t-1} + \delta_3 gender_i + \delta_4 gradYear_i + \psi_{i,t} \qquad (23)$$

An individual makes a choice whether to drop out of the job market if $d_{i,t}^{*} < 0$. We assume that all graduates choose to stay in the market at the year of MBA graduation. Subsequently, if they exit the job market at year t because ($d_{i,t}^{*} < 0$), then they stay out of the job market permanently ($d_{i,t'}^{*} < 0 \; \forall \; t' \geq t$). Individual can make the choice to exit based on various factors – (i) An unsuccessful individual (small $r_{i,t-1}$) may be more likely to drop out of the job market if they have an outside option (ii) the outside option may be better (or worse) for attractive individuals (e.g. via marriage prospects), (iii) the outside option may be better (or worse) for one gender than the other (e.g. homemaker or blue collar work). (iv) certain periods may have unusually low employment e.g. 2008-09 crisis where unemployment temporarily went from 4% to 10%. The drop out model above captures all these factors.

Since our dataset captures MBA classes from 1990 to 2015, comparisons between graduating classes is one way to ascertain whether or not drop out occurs. In Appendix A.11.1, we show lack of evidence for systematic drop out based on job rank ($\delta_1 r_{i,t-1}$) or attractiveness ($\delta_2 b_{i,t-1}$). In Appendix A.11.2, we reinforce lack of evidence for systematic drop out based on job rank ($\delta_1 r_{i,t-1}$). In Appendix A.11.3, we show evidence for systematic drop out based on gender ($\delta_3 gender_i$) and year of graduation ($\delta_4 gradYear_i$). We also present a robustness check to ensure attractiveness bias estimates remain unaffected.

A.11.1 Job Rank and Attractiveness Dropout

First we describe a test for job rank based dropout. Then we describe a test for attractiveness based dropout. Finally we highlight a statistical, but not very plausible, situation where our attractiveness bias results may be effected even though we reject both job rank based and attractiveness based drop outs individually.

Let us consider that an individual makes a choice whether to drop out of the job market based on the job rank $r_{i,t}$. The average rank of individuals remaining in the job market at year 1 ($E'\left[r_{i,1}\right]$) would be different from an uncensored average ($E\left[r_{i,1}\right]$) as shown in equation 12. Similarly, the average rank of individuals in the job market at year 2 ($E'\left[r_{i,2}\right]$) is further censored by the dropouts in year 1 or year 2.

$$d_{i,t}^{*} = \delta_0 + \delta_1 r_{i,t-1} + \psi_{i,t} \tag{24}$$

$$E'\left[r_{i,1}\right] = E[r_{i,1}/\psi_{i,o} > -(\delta_0 + \delta_1 r_{i,0})] \neq E\left[r_{i,1}\right] \tag{25}$$
$$E'\left[r_{i,2}\right] = E[r_{i,2}/\psi_{i,o} > -(\delta_0 + \delta_1 r_{i,0}), \psi_{i,1} > -(\delta_0 + \delta_1 r_{i,1})] \neq E\left[r_{i,2}\right]$$

Consider profiles collected in 2017 for graduates from 1997. The average observed job rank $E'\left[r_{i,5}\right]$ for these individuals in 2002 will be different than the actual $E\left[r_{i,5}\right]$. This is because some of those graduates may have dropped out prior to 2002 ($d_{i,t}^{*} < 0\ for\ t \in [0,5]$) or between 2002 and 2017 ($d_{i,t}^{*} < 0\ for\ t \in [5,20]$). If individuals drop out because of lower job ranks, we would expect that remaining available profiles are upwardly biased in terms of career success ($E'\left[r_{i,5}\right]$ > $E\left[r_{i,5}\right]$).

$$E'_{grad=1997}\left[r_{i,5}\right] = E\left[r_{i,5}/d_{i,t}^{*} > 0\ for\ t \in [0,5] \cup [5,20]\right] \tag{26}$$
$$E'_{grad=2012}\left[r_{i,5}\right] = E\left[r_{i,5}/d_{i,t}^{*} > 0\ for\ t \in [0,5]\right]$$

Now consider a different graduating class from 2012. Once again, the observed average ($E'\left[r_{i,5}\right]$ > $E\left[r_{i,5}\right]$) is upwardly biased. However, this upward bias is only the result of dropouts prior to 2017 ($d_{i,t}^{*} < 0\ for\ t \in [0,5]$). We can compare the average job rank for 1997 graduates in 2002 and 2012 graduates in 2017. The second sample also looks at job ranks 5 years from graduation but has a smaller dropouts. If unsuccessful individuals systematically drop out ($\delta_1 > 0$), then $E'_{grad=2012}\left[r_{i,5}\right] < E'_{grad=1997}\left[r_{i,5}\right]$. The Figure 26 provides this comparison and suggests no significant impact of job rank on drop out. We also find that the job rank ($E'_{g}\left[r_{i,t}\right]$) and graduation year ($g$) correlation to be statistically insignificant. This means that dropout is not a systematic result of a lower (or higher) job rank.

Similar as above, a second argument could be that attractive individuals are more (or less) likely to drop out ($\delta_2 \neq 0$). We can compare the average attractiveness for MBA class of 2005 and 2007 in 2017 matched on age in 2017. If attractive individuals systematically drop out ($\delta_2 > 0$), then $E'\big[b_{i,2017}/ age, gender, gradYear = 2005\big] < E'\big[b_{i,2017}/age, gender, gradYear = 2007\big]$. We construct a simple regression (equation 14) to check if MBA Graduation Year has an impact on observed attractiveness among individuals of same age in 2017 and similar demographics (gender, ethnicity). We do not find statistically significant relationship between MBA Graduation Year and attractiveness. Thus we do not have any evidence to support systematic drop our based on attractiveness.

$$b_{i,2017} = \alpha_0 + \alpha_1 age + \alpha_2 D_i + \ \alpha_3 mbaGradYear + \epsilon_i \qquad (27)$$

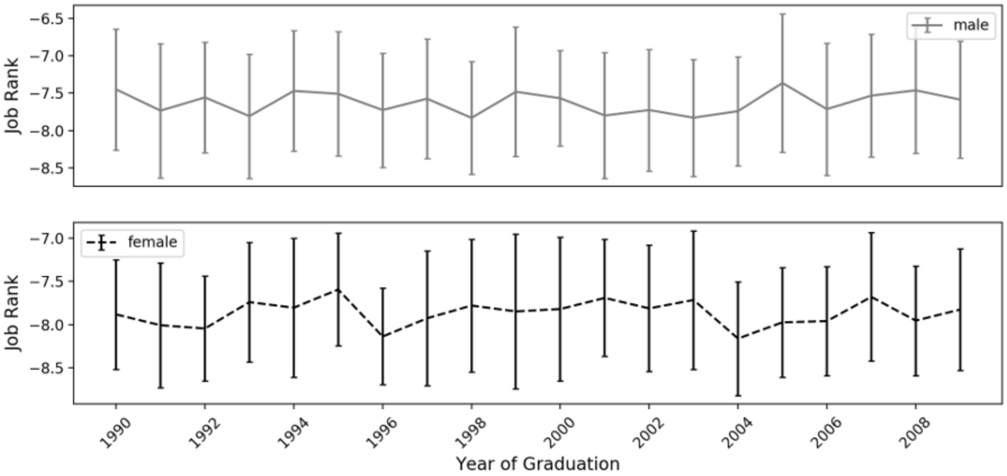


Figure 26: Average job page ranks (and standard deviation) attained in 5 years for different graduating classes between 1990 and 2010. If lower-ranked individuals drop out, the earlier graduation year should appear to have higher job page ranks. No evidence of dropouts of lower-ranked individuals was found.

Even though we reject drop out based on rank and attractiveness individually, consider a setting where both of the following are true: (i) Attractive unsuccessful individuals are more likely to drop out. Our data will only observe relatively successful attractive individuals, thus creating a false perception that attractive individuals have a premium. (ii) Plain-looking, successful individuals are more likely to drop out. Our data will only observe relatively unsuccessful, plain-looking individuals, thus creating a false perception that plain-looking individuals incur a penalty. Note that (i) and (ii) must hold simultaneously; otherwise, we would have seen a systematically higher drop out based on job rank or attractiveness alone. We admit that (i) is plausible since attractive individuals may be more likely to have marriage prospects with financially well-off individuals. We do not see a plausible explanation for (ii). Nevertheless we admit that we do not have evidence to reject this.

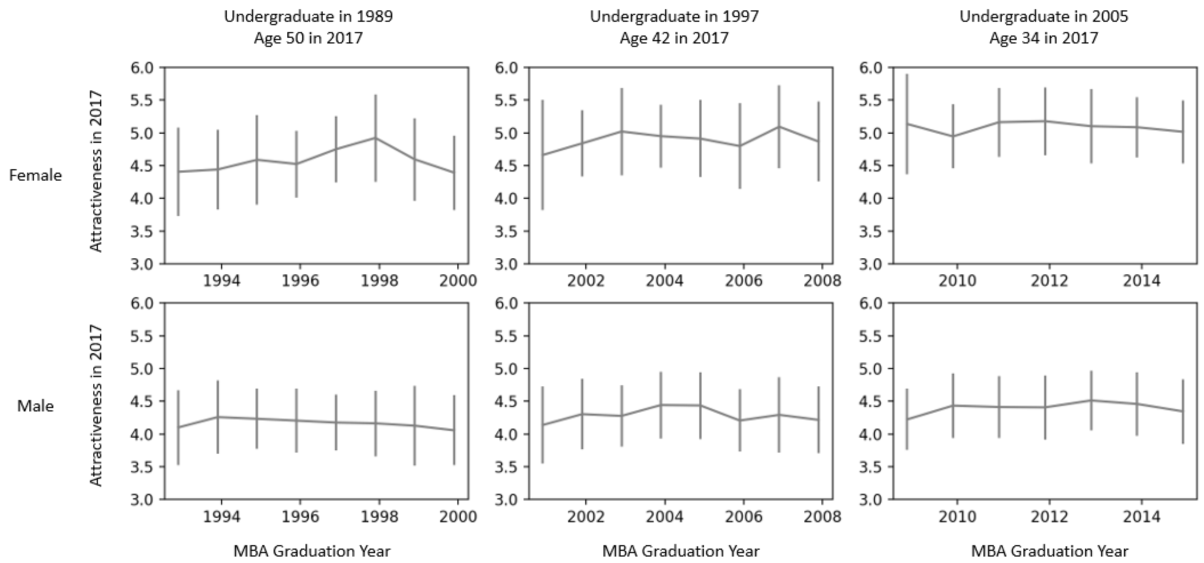


Figure 27*: Average observed attractiveness (and standard deviation) for different MBA graduating classes among individuals of same gender and age in 2017. No correlation or evidence of systematic attractiveness dropout.*

A.11.2 No picture drop out

We further reinforce the argument that job rank ($\delta_1 r_{i,t-1}$) does not play a role in sample drop outs by specifically comparing profiles with and without pictures. Its plausible that plain looking ambitious individuals strategically decide to not post pictures in order to hide their plain looks. If so, profiles without pictures would be systematically more successful in job market than profiles with pictures. We run a simple regression (equation 15) to check if job rank (r_i) is associated with individuals decision to post picture. Unfortunately, our original data gathering did not fetch full profiles for individuals who do not post profile pictures. Nevertheless we do have their "headline" – Name, Current Job Title and Industry. We can use the name to infer the gender and ethnicity. We can apply the page rank module to rank desirability of "Title" + "Industry Category". Note that this is a little different from the definition of Job = "Title" + "Employer Size" + "Industry Category" used in our main analysis. We do not find statistically significant relationship between decision to post picture and the job rank (r_i). Thus we do not have any evidence to support strategic decision to not post profile pictures by plain looking successful individuals.

$$picture_i = \alpha_0 + \alpha_1 genderName_i + \alpha_2 ethnicityName_i + \alpha_4 industry_i + \alpha_3 r_i + \epsilon_i \qquad (28)$$

[Insert Table Here – Refer End of Appendix]

Table 47*: Relationship of job rank and likelihood of a clearly posted profile picture.*

A.11.3 Gender and Rank Dropout

We use a secondary dataset (2,506 profiles) where we can observe gender and year of graduation based drop outs. This dataset captures entire MBA directory for one school, thus sheds light on exact number of individuals by gender and graduation year that are selected out of the professional profile collection. Figure 28 show percentage of profiles missing from each graduation year. Missing profiles for year 2000 graduates can be interpreted as all dropout between 1-17 years of MBA. Similar missing profiles for 2010 graduates can be interpreted as all dropout between 1-7 years of MBA. Naturally, drop out is much greater for earlier graduation years than recent graduation years. This is reasonable as there may simply be retirements as well as individuals who never got onboard professional social networking websites. We can see that dropout rate for women is much higher. This may be explained by switch to homemaking or childcare for women.

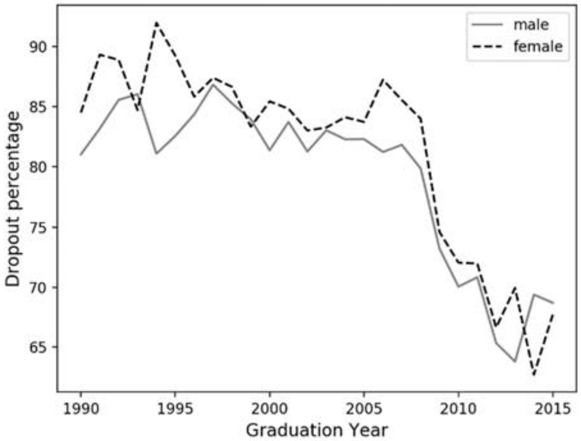


Figure 28: Percentage of profiles missing from each graduation year for men and women.

There is no obvious argument why gender or graduation year based dropout would impact our attractiveness bias results. Nevertheless, we create an alternate specification of our attractiveness bias analysis following Heckman's sample selection model. We only observe $(r_{i,t} = r_{i,t}^*)$ for individuals who choose to remain in the job market ($d_i^* > 0$). The drop out decision depends on gender (δ_3) and the graduation year (δ_4). Model free evidence above suggests that both these factors play a big role in dropouts. Note that we model a single drop out decision for an individual d_i^* instead of a time varying decision $d_{i,t}^*$. This is because we have already dropped time varying drop out factors ($r_{i,t-1}, b_{i,t-1}$) and we directly observe accumulation of all drop outs between graduation and 2017.

$$d_i^* = \delta_0 + \delta_3 gender_i + \delta_4 gradYear_i + \psi_i \tag{29}$$

$$r_{i,2017}^* = \beta_1 + \beta_2 b_i + \overrightarrow{\beta_4} Z_i + u_i \tag{30}$$

$$r_{i,2017} = \begin{cases} r_{i,2017}^* & if\ d_i^* > 0 \\ Not\ observed & Else \end{cases}$$

[Insert Table Here – Refer End of Appendix]

Table 48: (Model1) Impact of attractiveness without selection model (Model2) Impact of attractiveness with Heckman selection model (Model3) Impact of belief and preference bias with heckman selection model.

Table 48 reports Heckman estimation results. As expected we find dropout decision to be significant for both gender and graduation year. Our primary results – persistent preference bias, insignificant belief bias and lifetime premium remain the same. As expected, the standard errors are much larger on this dataset given a 8 times smaller dataset (2,506 in secondary vs 43,533 in primary dataset). The data size further reduces below 1000 as we try to replicate all the analysis done with the primary dataset, which requires dropping profiles with missing control covariates (Education, Job Domain etc.) and using matched samples only. After controlling for or randomizing over 50 covariates most of the standard errors become too large. This small robustness check should only be seen as indicative. We dropped any attempts to increase the sample size, because it would requires us to go after full MBA alumni directories for more schools.

A.11.4 Reverse Causality

We study if attractiveness leads to career success. It could be argued that successful individuals are financially well off and may have more resources to spend on their appearance. These resources may be spent directly on the posted profile picture (e.g., professional photoshoots of executives) or indirectly spent over a long term on healthy lifestyle or cosmetic products. The latter may lead to fewer episodes of illnesses and wholesome skin. Therefore, success leads to attractiveness. This concern is alleviated in our paper because of our attractiveness calculations which crops, re-orients, standardizes photograph quality and devalues temporary characteristics. Further we randomize on alterable photograph quality, facial hair, makeup and even health related features such as skin health, dark circles and skin stains. All of these are meant to ensure that subjects choice of profile picture or long term investment into their appearance does not effect our results. Keeping all of these mitigants discussed above aside, let us assume that long term lifestyle and beautification investment by successful people leads to a permanent (not just a superficial) change in their attractiveness. These factors presumably take significant time (say 3-5 years) to have permanent impact. This means that an MBA graduate from 2005 who had a successful career has better looks in 2017 than a counterpart who was unsuccessful. However, an MBA graduate in 2015 who has a successful first two years would hardly look different in 2017 than a counterpart who was unsuccessful. Two years between 2015 and 2017 is too little time for the successful individual to improve their looks via better health, lower stress etc.

$$b_{i,2017} = \alpha_0 + \alpha_1 D_i + \alpha_2 JD_i \; + \; \alpha_3 E_i + \alpha_4 r_i \; + \alpha_5 mbaGradYear + \alpha_6 r_i * mbaGradYear \; + \epsilon_i \tag{31}$$

We construct a simple regression (equation 18) check if a higher job rank is associated with greater attractiveness in 2017. In particular we would expect the relationship to be stronger for older MBA graduates compared to more recent graduates. We do not find this interaction effect (α_6) to be statistically significant. Thus we do not have any evidence to support the hypothesis that attractiveness is driven by investment of greater resources by successful individuals.

[Insert Table Here – Refer End of Appendix]

Table 49*: Relationship of job rank and individuals attractiveness with interaction with MBA graduation year.*

Main Text Regression Tables

Table 10: (Model 1 and 2) The impact of attractiveness on career rank with and without gender interaction. (Model 3 and 4) The impact of attractiveness preference and attractiveness belief on career rank with and without gender interaction.

	(1)	(2)	(3)	(4)
	rank ($r_{i,t}$)	rank ($r_{i,t}$)	rank ($r_{i,t}$)	rank ($r_{i,t}$)
Attractiveness	0.011**	0.012**		
	(0.003)	(0.004)		
Attractiveness * Female		-0.003		
		(0.008)		
Belief Bias			-0.005	-0.008
			(0.007)	(0.008)
Preference Bias			0.013**	0.015**
			(0.005)	(0.005)
Belief Bias * Female				0.013
				(0.015)
Preference Bias * Female				-0.009
				(0.010)
Gender(Male)	0.164**	0.162**	0.163**	0.178**
	(0.004)	(0.006)	(0.004)	(0.007)
Ethnicity (European)	-0.046**	-0.046**	-0.046**	-0.046**
	(0.009)	(0.009)	(0.009)	(0.009)
Ethnicity (Others)	0.024*	0.024*	0.025*	0.025*
	(0.010)	(0.010)	(0.010)	(0.010)
UG (Business)	-0.147**	-0.147**	-0.147**	-0.147**
	(0.009)	(0.009)	(0.009)	(0.009)
UG (Science)	-0.019**	-0.019**	-0.018**	-0.018**
	(0.006)	(0.006)	(0.006)	(0.006)
UG (Others)	-0.072**	-0.072**	-0.073**	-0.073**
	(0.006)	(0.006)	(0.006)	(0.006)
MBA Rank	-0.199**	-0.199**	-0.200**	-0.200**
	(0.005)	(0.005)	(0.005)	(0.005)
UG Rank	-0.094**	-0.094**	-0.094**	-0.094**
	(0.005)	(0.005)	(0.005)	(0.005)
Job Type (Management)	-0.186**	-0.186**	-0.188**	-0.189**
	(0.006)	(0.006)	(0.006)	(0.006)
Job Type (Others)	-0.594**	-0.594**	-0.595**	-0.595**
	(0.006)	(0.006)	(0.006)	(0.006)
Job Type (Technical)	-0.685**	-0.685**	-0.685**	-0.686**
	(0.010)	(0.010)	(0.010)	(0.010)
Industry Category (IT)	-0.118**	-0.118**	-0.118**	-0.118**
	(0.005)	(0.005)	(0.005)	(0.005)
Industry Category (Management)	0.406**	0.406**	0.407**	0.407**
	(0.005)	(0.005)	(0.005)	(0.005)
Industry Category (Others)	-0.397**	-0.397**	-0.398**	-0.398**
	(0.005)	(0.005)	(0.005)	(0.005)
Large Employer	-0.208**	-0.208**	-0.206**	-0.206**
	(0.004)	(0.004)	(0.004)	(0.004)
Large Location	0.058**	0.058**	0.058**	0.058**
	(0.003)	(0.003)	(0.003)	(0.003)
Early Career			-0.015**	-0.025**
			(0.005)	(0.006)

Early Career * Female				0.036** (0.011)
pScore	0.075** (0.011)	0.075** (0.011)	0.081** (0.011)	0.080** (0.011)
Constant	0.495** (0.014)	0.496** (0.014)	0.500** (0.014)	0.490** (0.015)
Obs.	275858	275858	275858	275858
R-squared	0.169	0.169	0.169	0.169
Adj R-squared	0.169	0.169	0.169	0.169
F	3111.048	2947.309	2801.290	2437.457
RMSE	0.914	0.914	0.914	0.914

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

Table 13: (Model 1) The impact of attractiveness on career rank at the end of 15 years. (Model 2) The impact of attractiveness with gender interaction. (Model 3) The impact of attractiveness with ethnicity interaction. (Model 4) The impact of attractiveness for individual of European ethnicity alone.

	(1) rank ($r_{i,15}$)	(2) rank ($r_{i,15}$)	(3) rank ($r_{i,15}$)	(4) rank ($r_{i,15}$)
Attractiveness	**0.060**** (0.017)	0.055** (0.019)	**0.048*** (0.022)	0.081** (0.018)
Attractiveness * Female		0.021 (0.040)		
Attractiveness * Ethnicity (Others)			-0.037 (0.168)	
Attractiveness * Ethnicity (European)			**0.033*** (0.015)	
Gender (Male)	0.203** (0.021)	0.213** (0.029)	0.203** (0.021)	0.209** (0.022)
Ethnicity (European)	0.001 (0.043)	0.001 (0.043)	-0.016 (0.062)	
Ethnicity (Others)	0.109* (0.050)	0.109* (0.050)	0.175* (0.070)	
UG (Business)	-0.123** (0.045)	-0.123** (0.045)	-0.122** (0.045)	-0.114* (0.047)
UG (Science)	-0.083** (0.027)	-0.084** (0.027)	-0.083** (0.027)	-0.100** (0.028)
UG (Others)	-0.035 (0.031)	-0.035 (0.031)	-0.033 (0.031)	-0.023 (0.033)
MBA Rank	-0.198** (0.022)	-0.198** (0.022)	-0.198** (0.022)	-0.206** (0.024)
UG Rank	-0.094** (0.025)	-0.094** (0.025)	-0.096** (0.025)	-0.099** (0.027)
Job Type (Management)	0.100** (0.035)	0.101** (0.035)	0.099** (0.035)	0.094* (0.038)
Job Type (Others)	-0.219** (0.037)	-0.219** (0.037)	-0.220** (0.037)	-0.221** (0.039)

Job Type (Technical)	-0.426**	-0.426**	-0.429**	-0.415**
	(0.055)	(0.055)	(0.055)	(0.060)
Industry Category (IT)	-0.306**	-0.307**	-0.307**	-0.333**
	(0.024)	(0.024)	(0.024)	(0.026)
Industry Category (Management)	0.119**	0.119**	0.120**	0.089**
	(0.026)	(0.026)	(0.026)	(0.028)
Industry Category (Others)	-0.617**	-0.617**	-0.617**	-0.624**
	(0.025)	(0.025)	(0.025)	(0.027)
Large Employer	-0.346**	-0.346**	-0.347**	-0.352**
	(0.018)	(0.018)	(0.018)	(0.019)
Large Location	0.077**	0.077**	0.076**	0.065**
	(0.017)	(0.017)	(0.017)	(0.019)
pScore	0.043	0.043	0.040	-0.011
	(0.054)	(0.054)	(0.054)	(0.058)
Constant	0.376**	0.368**	0.385**	0.424**
	(0.069)	(0.070)	(0.081)	(0.058)
Obs.	12100	12100	12100	10176
R-squared	0.151	0.151	0.152	0.153
Adj R-squared	0.150	0.150	0.150	0.151
F	119.372	113.097	108.010	114.503
RMSE	0.932	0.932	0.931	0.926

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

Table 14: The impact of attractiveness on career rank at the end of 15 years. (Model 1) Interaction with Undergraduate degree, (Model 2) Interaction with undergraduate degree classified as Arts versus all others (Model 3) Interaction with Industry Category, (Model 4) Interaction with Industry Category classfied as Management versus all others. (Model 5) Interaction with Job Type. (Model 6) Interaction with MBA Rank.

	(1) rank $(r_{i,15})$	(2) rank $(r_{i,15})$	(3) rank $(r_{i,15})$	(4) rank $(r_{i,15})$	(5) rank $(r_{i,15})$	(6) rank $(r_{i,15})$
Attractiveness	**0.065***	**0.045***	**0.053***	**0.042***	**0.093**	**0.092****
	(0.027)	(0.018)	(0.028)	(0.019)	(0.089)	(0.024)
Attractiveness * UG (Arts)	**0.091**					
	(0.053)					
Attractiveness * UG (Business)	0.004					
	(0.082)					
Attractiveness * UG (Science)	-0.042					
	(0.038)					
Attractiveness * UG (Arts)		**0.112***				
		(0.049)				
Attractiveness * Ind. Cat. (Others)			-0.043			
			(0.083)			
Attractiveness * Ind. Cat. (IT)			-0.002			
			(0.066)			
Attractiveness *			**0.075***			

Ind. Cat. (Management)			(0.031)			
Unattractiveness * Ind. Cat. (Management)				**0.086***		
				(0.042)		
Attractiveness * Job Type (Consultant)					**0.060**	
					(0.360)	
Attractiveness * Job Type (Technical)					-0.133	
					(0.187)	
Attractiveness * Job Type (Others)					0.019	
					(0.091)	
Attractiveness * MBA Rank						**-0.076**
						(0.041)
Gender(Male)	0.203**	0.203**	0.203**	0.203**	0.203**	0.202**
	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)
Ethnicity (European)	0.001	0.001	0.002	0.001	0.001	0.003
	(0.043)	(0.043)	(0.043)	(0.043)	(0.043)	(0.043)
Ethnicity (Others)	0.108*	0.109*	0.109*	0.108*	0.107*	0.110*
	(0.050)	(0.050)	(0.050)	(0.050)	(0.050)	(0.050)
UG (Business)	-0.079	-0.179**	-0.124**	-0.123**	-0.123**	-0.122**
	(0.063)	(0.052)	(0.045)	(0.045)	(0.045)	(0.045)
UG (Science)	-0.017	-0.140**	-0.084**	-0.084**	-0.083**	-0.084**
	(0.038)	(0.037)	(0.027)	(0.027)	(0.027)	(0.027)
UG (Others)	0.010	-0.091*	-0.036	-0.036	-0.035	-0.034
	(0.041)	(0.040)	(0.031)	(0.031)	(0.031)	(0.031)
MBA Rank	-0.198**	-0.198**	-0.197**	-0.198**	-0.197**	-0.160**
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.030)
UG Rank	-0.094**	-0.094**	-0.093**	-0.093**	-0.094**	-0.095**
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
Job Type (Management)	0.100**	0.100**	0.100**	0.100**	0.130**	0.101**
	(0.035)	(0.035)	(0.035)	(0.035)	(0.049)	(0.035)
Job Type (Others)	-0.220**	-0.220**	-0.220**	-0.220**	-0.199**	-0.219**
	(0.037)	(0.037)	(0.037)	(0.037)	(0.051)	(0.037)
Job Type (Technical)	-0.427**	-0.427**	-0.427**	-0.427**	-0.356**	-0.424**
	(0.055)	(0.055)	(0.055)	(0.055)	(0.074)	(0.055)
Industry Category (IT)	-0.307**	-0.307**	-0.301**	-0.307**	-0.307**	-0.307**
	(0.024)	(0.024)	(0.034)	(0.024)	(0.024)	(0.024)
Industry Category (Management)	0.119**	0.119**	0.087*	0.162**	0.119**	0.119**
	(0.026)	(0.026)	(0.037)	(0.034)	(0.026)	(0.026)
Industry Category (Others)	-0.617**	-0.617**	-0.590**	-0.617**	-0.617**	-0.616**
	(0.025)	(0.025)	(0.036)	(0.025)	(0.025)	(0.025)
Large Employer	-0.346**	-0.346**	-0.346**	-0.346**	-0.346**	-0.345**
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Large Location	0.076**	0.076**	0.077**	0.077**	0.077**	0.076**
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
pScore	0.042	0.043	0.044	0.043	0.043	0.044
	(0.054)	(0.054)	(0.054)	(0.054)	(0.054)	(0.054)
Constant	0.330**	0.441**	0.373**	0.385**	0.351**	0.359**

	(0.072)	(0.074)	(0.071)	(0.069)	(0.076)	(0.069)
Obs.	12100	12100	12100	12100	12100	12100
R-squared	0.151	0.151	0.151	0.151	0.151	0.151
Adj R-squared	0.150	0.150	0.150	0.150	0.150	0.150
F	102.656	113.396	102.602	113.340	102.422	113.296
RMSE	0.931	0.931	0.931	0.931	0.932	0.931

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

Appendix Regression Tables

Table 21: (Model 1) The impact of attractiveness bias across four strata with strata indicator interaction effect. (Model 1 to 4) The impact of attractiveness bias in one propensity score strata at a time. The sample sizes are smaller and standard errors are larger in the first and fourth strata. Attractiveness bias is insignificant at 5% level (significant at 10% level) in these stratas.

	(1)	(2)	(3)	(4)	(5)
	All Strata	Strata 1	Strata 2	Strata 3	Strata 4
	rank ($r_{i,t}$)	rank ($r_{i,t}$)	rank ($r_{i,t}$)	rank ($r_{i,t}$)	rank ($r_{i,t}$)
Attractiveness	**0.107***	0.095	0.081*	0.081**	0.060
	(0.053)	(0.051)	(0.032)	(0.030)	(0.036)
Attractiveness * Strata [1-4]	**-0.011**				
	(0.018)				
Strata [1-4]	-0.001				
	(0.024)				
Gender (Male)	0.188**	0.157**	0.191**	0.228**	0.147**
	(0.022)	(0.059)	(0.039)	(0.040)	(0.053)
Ethnicity (European)	-0.044	-0.062	-0.034	-0.011	-0.098
	(0.045)	(0.164)	(0.097)	(0.077)	(0.077)
Ethnicity (Others)	0.056	-0.114	0.145	0.155	-0.083
	(0.052)	(0.172)	(0.104)	(0.091)	(0.118)
UG (Business)	-0.160**	-0.079	-0.235**	-0.040	-0.294**
	(0.047)	(0.134)	(0.085)	(0.080)	(0.101)
UG (Science)	-0.103**	-0.035	-0.118*	-0.081	-0.163**
	(0.028)	(0.082)	(0.051)	(0.047)	(0.063)
UG (Others)	-0.066*	-0.063	-0.081	0.005	-0.167*
	(0.033)	(0.097)	(0.059)	(0.054)	(0.074)
MBA Rank	-0.183**	-0.247**	-0.148**	-0.177**	-0.204**
	(0.023)	(0.066)	(0.042)	(0.039)	(0.047)
UG Rank	-0.102**	-0.162*	-0.080	-0.102*	-0.092
	(0.026)	(0.072)	(0.045)	(0.046)	(0.063)
Job Type (Management)	0.096**	-0.055	0.107	0.099	0.164*
	(0.037)	(0.106)	(0.065)	(0.063)	(0.077)
Job Type (Others)	-0.229**	-0.384**	-0.244**	-0.168*	-0.213**
	(0.038)	(0.110)	(0.068)	(0.066)	(0.080)
Job Type (Technical)	-0.474**	-0.477**	-0.582**	-0.495**	-0.380**
	(0.058)	(0.153)	(0.109)	(0.107)	(0.115)

Industry Category (IT)	-0.323**	-0.324**	-0.325**	-0.348**	-0.289**
	(0.025)	(0.075)	(0.047)	(0.042)	(0.050)
Industry Category (Management)	0.128**	0.137	0.181**	0.106*	0.091
	(0.028)	(0.081)	(0.051)	(0.047)	(0.056)
Industry Category (Others)	-0.613**	-0.563**	-0.558**	-0.657**	-0.648**
	(0.026)	(0.080)	(0.049)	(0.045)	(0.053)
Large Employer	-0.325**	-0.320**	-0.279**	-0.318**	-0.388**
	(0.018)	(0.054)	(0.034)	(0.032)	(0.038)
Large Location	0.054**	0.056	0.060	0.009	0.103**
	(0.018)	(0.052)	(0.032)	(0.031)	(0.037)
pScore	0.115	-0.014	0.352	-0.188	0.492*
	(0.123)	(0.364)	(0.264)	(0.209)	(0.236)
Constant	0.413**	0.652**	0.232	0.502**	0.263
	(0.076)	(0.223)	(0.157)	(0.167)	(0.204)
Obs.	11000	1348	3086	3802	2764
R-squared	0.153	0.157	0.164	0.150	0.162
Adj R-squared	0.151	0.146	0.159	0.146	0.157
F	99.091	13.784	33.326	36.958	29.488
RMSE	0.928	0.934	0.897	0.934	0.948

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

Table 22: The impact of attractiveness bias using different measures for attractiveness - (Model 1) binary classification ($b_{i,t}$), (Model 2) continuous measure ($beauty_{i,t}$), (Model 3) static measure as on MBA graduation ($beauty_{i,t=0}$), (Model 4) static measure as at data collection in 2017.

	(1) rank ($r_{i,t}$)	(2) rank ($r_{i,t}$)	(3) rank ($r_{i,t}$)	(4) rank ($r_{i,t}$)
Attractiveness Class ($b_{i,t}$)	**0.011***			
	(0.003)			
Attractiveness Rating ($beauty_{i,t}$)		**0.016***		
		(0.004)		
Attractiveness Rating at $t = 0$			**0.011***	
			(0.004)	
Attractiveness Rating in 2017				**0.017***
				(0.004)
Gender(Male)	0.164**	0.163**	0.164**	0.174**
	(0.004)	(0.004)	(0.004)	(0.004)
Ethnicity (European)	-0.046**	-0.046**	-0.046**	-0.046**
	(0.009)	(0.009)	(0.009)	(0.009)
Ethnicity (Others)	0.024*	0.023*	0.024*	0.023*
	(0.010)	(0.010)	(0.010)	(0.010)
UG (Business)	-0.147**	-0.148**	-0.147**	-0.148**
	(0.009)	(0.009)	(0.009)	(0.009)
UG (Science)	-0.019**	-0.019**	-0.019**	-0.019**

	(0.006)	(0.006)	(0.006)	(0.006)
UG (Others)	-0.072**	-0.072**	-0.072**	-0.072**
	(0.006)	(0.006)	(0.006)	(0.006)
MBA Rank	-0.199**	-0.199**	-0.199**	-0.199**
	(0.005)	(0.005)	(0.005)	(0.005)
UG Rank	-0.094**	-0.094**	-0.094**	-0.094**
	(0.005)	(0.005)	(0.005)	(0.005)
Job Type (Management)	-0.186**	-0.185**	-0.186**	-0.186**
	(0.006)	(0.006)	(0.006)	(0.006)
Job Type (Others)	-0.594**	-0.594**	-0.594**	-0.594**
	(0.006)	(0.006)	(0.006)	(0.006)
Job Type (Technical)	-0.685**	-0.685**	-0.685**	-0.685**
	(0.010)	(0.010)	(0.010)	(0.010)
Industry Category (IT)	-0.118**	-0.118**	-0.118**	-0.118**
	(0.005)	(0.005)	(0.005)	(0.005)
Industry Category (Management)	0.406**	0.406**	0.406**	0.406**
	(0.005)	(0.005)	(0.005)	(0.005)
Industry Category (Others)	-0.397**	-0.397**	-0.397**	-0.397**
	(0.005)	(0.005)	(0.005)	(0.005)
Large Employer	-0.208**	-0.208**	-0.208**	-0.208**
	(0.004)	(0.004)	(0.004)	(0.004)
Large Location	0.058**	0.058**	0.058**	0.059**
	(0.003)	(0.003)	(0.003)	(0.003)
pScore	0.075**	0.068**	0.071**	0.067**
	(0.011)	(0.011)	(0.011)	(0.011)
Constant	0.495**	0.506**	0.503**	0.422**
	(0.014)	(0.014)	(0.014)	(0.021)
Obs.	275858	275858	275858	275858
R-squared	0.169	0.169	0.169	0.169
Adj R-squared	0.169	0.169	0.169	0.169
F	3111.048	3111.735	3111.024	3111.891
RMSE	0.914	0.914	0.914	0.914

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

Table 23: The impact of attractiveness preference and belief bias using (Model 1) standardised job ranks ($r_{i,t}$), (Model 2) unstandardised job ranks ($rank_{i,t}$). The impact of attractiveness on career outcome at the end of 15 years using (Model 3) standardised job ranks ($r_{i,t}$), (Model 4) unstandardised job ranks ($rank_{i,t}$).

	(1) Standardized rank ($r_{i,t}$)	(2) Unstandardized rank ($rank_{i,t}$)	(3) Standardized rank ($r_{i,t}$)	(4) Unstandardized rank ($rank_{i,t}$)
Belief Bias	-0.005	-0.003		
	(0.007)	(0.005)		
Preference Bias	**0.013****	**0.008****		

	(0.005)	(0.003)		
Attractiveness Bias			**0.060****	**0.038****
			(0.017)	(0.011)
Gender(Male)	0.163**	0.108**	0.203**	0.128**
	(0.004)	(0.003)	(0.021)	(0.013)
Ethnicity (European)	-0.046**	-0.031**	0.001	0.001
	(0.009)	(0.006)	(0.043)	(0.028)
Ethnicity (Others)	0.025*	0.016*	0.109*	0.069*
	(0.010)	(0.007)	(0.050)	(0.032)
UG (Business)	-0.147**	-0.098**	-0.123**	-0.078**
	(0.009)	(0.006)	(0.045)	(0.029)
UG (Science)	-0.018**	-0.012**	-0.083**	-0.053**
	(0.006)	(0.004)	(0.027)	(0.017)
UG (Others)	-0.073**	-0.051**	-0.035	-0.022
	(0.006)	(0.004)	(0.031)	(0.020)
MBA Rank	-0.200**	-0.134**	-0.198**	-0.125**
	(0.005)	(0.003)	(0.022)	(0.014)
UG Rank	-0.094**	-0.062**	-0.094**	-0.060**
	(0.005)	(0.003)	(0.025)	(0.016)
Job Type (Management)	-0.188**	-0.139**	0.100**	0.064**
	(0.006)	(0.004)	(0.035)	(0.022)
Job Type (Others)	-0.595**	-0.409**	-0.219**	-0.139**
	(0.006)	(0.004)	(0.037)	(0.023)
Job Type (Technical)	-0.685**	-0.467**	-0.426**	-0.270**
	(0.010)	(0.007)	(0.055)	(0.035)
Industry Category (IT)	-0.118**	-0.074**	-0.306**	-0.194**
	(0.005)	(0.003)	(0.024)	(0.015)
Industry Category (Management)	0.407**	0.276**	0.119**	0.075**
	(0.005)	(0.003)	(0.026)	(0.017)
Industry Category (Others)	-0.398**	-0.258**	-0.617**	-0.391**
	(0.005)	(0.004)	(0.025)	(0.016)
Large Employer	-0.206**	-0.133**	-0.346**	-0.219**
	(0.004)	(0.002)	(0.018)	(0.011)
Large Location	0.058**	0.039**	0.077**	0.049**
	(0.003)	(0.002)	(0.017)	(0.011)
Years from graduation (t)		0.016**		
		(0.001)		
Early Career	-0.015**	-0.015**		
	(0.005)	(0.005)		
pScore	0.081**	0.053**	0.043	0.027
	(0.011)	(0.007)	(0.054)	(0.034)
Constant	0.500**	-7.600**	0.376**	-7.493**
	(0.014)	(0.011)	(0.069)	(0.044)
Obs.	275858	275858	12100	12100
R-squared	0.169	0.178	0.151	0.151
Adj R-squared	0.169	0.178	0.150	0.150
F	2801.290	2838.294	119.372	119.372
RMSE	0.914	0.607	0.932	0.591

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

Table 26: (Model 1) Time Taken to attain job j with θ_j fixed effects for all jobs reported by at least 20 individuals, (Model 2) Time Taken to attain job j with $\beta_j$random effect with all jobs reported by at least 2 individuals, (Model 3) Time Taken in job transition $j_1 \rightarrow j_2$ with $\beta_{j1,j2}$ random effects with all jobs reported by at least 2 individuals.

	(1)	(2)	(3)
Gender(Male)	-0.053*	-0.052	-0.106*
	(0.021)	(0.050)	(0.045)
Attractiveness	**-0.401****	**-0.268****	**-0.009**
	(0.018)	(0.041)	(0.038)
Constant	8.373**	10.451**	3.701**
	(0.100)	(0.584)	(0.256)
Job (var)	9.433**		
	(0.0746)		
Job Pair (var)			2.136**
			(0.129)
Residual (var)	5.710**		0.679**
	(0.002)		(0.016)
Obs.	209924	23443	4365
Groups	54,582	209	694
R-squared		0.436	
r2_a		0.431	
F		85.792	
rmse		3.012	
Log Likelihood	-526678.9		-6272.7

Standard errors are in parenthesis
** p<0.01, * p<0.05

Table 27: The impact of attractiveness preference and belief bias with different early career cutoff for belief bias - (Model 1) original cut off at 6 years, (Model 2) cut off at 4 years (Model 3) cut off at 8 years.

	(1)	(2)	(3)
	$T^{cutoff} = 6yrs$	$T^{cutoff} = 4yrs$	$T^{cutoff} = 8yrs$
	rank ($r_{i,t}$)	rank ($r_{i,t}$)	rank ($r_{i,t}$)
Belief Bias	-0.005	-0.008	-0.012
	(0.007)	(0.008)	(0.007)
Preference Bias	**0.013****	**0.013****	**0.018****
	(0.005)	(0.004)	(0.005)
Gender(Male)	0.163**	0.163**	0.163**
	(0.004)	(0.004)	(0.004)
Ethnicity (European)	-0.046**	-0.046**	-0.045**
	(0.009)	(0.009)	(0.009)

Ethnicity (Others)	0.025*	0.025*	0.025*
	(0.010)	(0.010)	(0.010)
UG (Business)	-0.147**	-0.147**	-0.147**
	(0.009)	(0.009)	(0.009)
UG (Science)	-0.018**	-0.019**	-0.018**
	(0.006)	(0.006)	(0.006)
UG (Others)	-0.073**	-0.073**	-0.073**
	(0.006)	(0.006)	(0.006)
MBA Rank	-0.200**	-0.200**	-0.200**
	(0.005)	(0.005)	(0.005)
UG Rank	-0.094**	-0.094**	-0.094**
	(0.005)	(0.005)	(0.005)
Job Type (Management)	-0.188**	-0.188**	-0.188**
	(0.006)	(0.006)	(0.006)
Job Type (Others)	-0.595**	-0.595**	-0.595**
	(0.006)	(0.006)	(0.006)
Job Type (Technical)	-0.685**	-0.686**	-0.685**
	(0.010)	(0.010)	(0.010)
Industry Category (IT)	-0.118**	-0.118**	-0.118**
	(0.005)	(0.005)	(0.005)
Industry Category (Management)	0.407**	0.407**	0.406**
	(0.005)	(0.005)	(0.005)
Industry Category (Others)	-0.398**	-0.398**	-0.398**
	(0.005)	(0.005)	(0.005)
Large Employer	-0.206**	-0.207**	-0.206**
	(0.004)	(0.004)	(0.004)
Large Location	0.058**	0.058**	0.058**
	(0.003)	(0.003)	(0.003)
Early Career	-0.015**	-0.012*	-0.013*
	(0.005)	(0.006)	(0.005)
pScore	0.081**	0.079**	0.082**
	(0.011)	(0.011)	(0.011)
Constant	0.500**	0.498**	0.500**
	(0.014)	(0.014)	(0.014)
Obs.	275858	275858	275858
R-squared	0.169	0.169	0.169
Adj R-squared	0.169	0.169	0.169
F	2801.290	2800.942	2801.685
RMSE	0.914	0.914	0.914

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

Table 28: The impact of attractiveness preference and belief bias using different university rankings - (Model 1) original UG and MBA ranks (Model 2) UG rank from 2010, (Model 3) MBA rank from 2001, and (Model 4) MBA rank from 2012.

	(1)	(2)	(3)	(4)
	rank ($r_{i,t}$)	rank ($r_{i,t}$)	rank ($r_{i,t}$)	rank ($r_{i,t}$)

Belief Bias	-0.005	-0.005	-0.004	-0.005
	(0.007)	(0.007)	(0.007)	(0.007)
Preference Bias	**0.013****	**0.013****	**0.011***	**0.011***
	(0.005)	(0.005)	(0.005)	(0.005)
Gender(Male)	0.163**	0.165**	0.162**	0.162**
	(0.004)	(0.004)	(0.004)	(0.004)
Ethnicity (European)	-0.046**	-0.046**	-0.042**	-0.042**
	(0.009)	(0.009)	(0.009)	(0.009)
Ethnicity (Others)	0.025*	0.026**	0.032**	0.030**
	(0.010)	(0.010)	(0.010)	(0.010)
UG (Business)	-0.147**	-0.133**	-0.138**	-0.141**
	(0.009)	(0.009)	(0.009)	(0.009)
UG (Science)	-0.018**	-0.012*	-0.015**	-0.016**
	(0.006)	(0.006)	(0.006)	(0.006)
UG (Others)	-0.073**	-0.072**	-0.031**	-0.038**
	(0.006)	(0.006)	(0.007)	(0.007)
MBA Rank	**-0.200****	-0.196**		
	(0.005)	(0.005)		
MBA Rank 2001			**-0.198****	
			(0.005)	
MBA Rank 2012				**-0.183****
				(0.005)
UG Rank	**-0.094****		-0.094**	-0.098**
	(0.005)		(0.005)	(0.005)
UG Rank 2010		**-0.113****		
		(0.005)		
Job Type (Management)	-0.188**	-0.189**	-0.194**	-0.192**
	(0.006)	(0.006)	(0.006)	(0.006)
Job Type (Others)	-0.595**	-0.594**	-0.594**	-0.594**
	(0.006)	(0.006)	(0.006)	(0.006)
Job Type (Technical)	-0.685**	-0.685**	-0.657**	-0.667**
	(0.010)	(0.010)	(0.010)	(0.010)
Industry Category (IT)	-0.118**	-0.119**	-0.117**	-0.116**
	(0.005)	(0.005)	(0.005)	(0.005)
Industry Category (Management)	0.407**	0.407**	0.407**	0.408**
	(0.005)	(0.005)	(0.005)	(0.005)
Industry Category (Others)	-0.398**	-0.397**	-0.401**	-0.400**
	(0.005)	(0.005)	(0.005)	(0.005)
Large Employer	-0.206**	-0.206**	-0.204**	-0.204**
	(0.004)	(0.004)	(0.004)	(0.004)
Large Location	0.058**	0.058**	0.060**	0.061**
	(0.003)	(0.003)	(0.003)	(0.003)
Early Career	-0.015**	-0.015**	-0.019**	-0.018**
	(0.005)	(0.005)	(0.005)	(0.005)
pScore	0.081**	0.081**	0.106**	0.104**
	(0.011)	(0.011)	(0.011)	(0.011)
Constant	0.500**	0.516**	0.500**	0.495**
	(0.014)	(0.014)	(0.014)	(0.014)
Obs.	275858	275858	275858	275858

R-squared	0.169	0.169	0.169	0.168
Adj R-squared	0.169	0.169	0.169	0.168
F	2801.290	2809.789	2803.706	2785.996
RMSE	0.914	0.914	0.914	0.915

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

Table 29: The impact of attractiveness preference and belief bias using (Model 1) all MBA graduates, (Model 2) MBA graduates after 1998, (Model 3) MBA graduates prior to 2008, (Model 4) MBA Rank less than 20, and (Model 5) MBA Rank less than 100.

	(1)	(2)	(3)	(4)	(5)
	All	MBA	MBA	MBA	MBA
		Grad Year	Grad Year	Rank	Rank
		> 1998	< 2008	<= 20	< 100
	rank ($r_{i,t}$)	rank_std	rank_std	rank_std	rank_std
Belief Bias	-0.005	-0.015	0.001	-0.002	-0.009
	(0.007)	(0.009)	(0.008)	(0.009)	(0.008)
Preference Bias	**0.013****	**0.014****	**0.011***	**0.021****	**0.017****
	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)
Gender(Male)	0.163**	0.175**	0.162**	0.180**	0.178**
	(0.004)	(0.005)	(0.004)	(0.005)	(0.005)
Ethnicity (European)	-0.046**	-0.044**	-0.053**	-0.051**	-0.043**
	(0.009)	(0.010)	(0.010)	(0.012)	(0.011)
Ethnicity (Others)	0.025*	0.025*	0.019	0.021	0.026*
	(0.010)	(0.012)	(0.011)	(0.013)	(0.012)
UG (Business)	-0.147**	-0.151**	-0.143**	-0.128**	-0.138**
	(0.009)	(0.011)	(0.010)	(0.012)	(0.010)
UG (Science)	-0.018**	-0.033**	-0.011	-0.030**	-0.018**
	(0.006)	(0.006)	(0.006)	(0.007)	(0.006)
UG (Others)	-0.073**	-0.060**	-0.070**	-0.082**	-0.085**
	(0.006)	(0.007)	(0.007)	(0.008)	(0.007)
MBA Rank	-0.200**	-0.184**	-0.211**	-2.113**	-0.366**
	(0.005)	(0.005)	(0.005)	(0.043)	(0.010)
UG Rank	-0.094**	-0.094**	-0.094**	-0.061**	-0.083**
	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Job Type (Management)	-0.188**	-0.151**	-0.219**	-0.181**	-0.166**
	(0.006)	(0.007)	(0.007)	(0.008)	(0.007)
Job Type (Others)	-0.595**	-0.552**	-0.612**	-0.544**	-0.558**
	(0.006)	(0.007)	(0.007)	(0.008)	(0.007)
Job Type (Technical)	-0.685**	-0.706**	-0.678**	-0.660**	-0.638**
	(0.010)	(0.012)	(0.011)	(0.015)	(0.012)
Industry Category (IT)	-0.118**	-0.177**	-0.083**	-0.178**	-0.152**
	(0.005)	(0.006)	(0.006)	(0.007)	(0.006)
Industry Category (Management)	0.407**	0.362**	0.417**	0.364**	0.380**
	(0.005)	(0.006)	(0.006)	(0.007)	(0.006)
Industry Category (Others)	-0.398**	-0.436**	-0.389**	-0.475**	-0.435**
	(0.005)	(0.006)	(0.006)	(0.007)	(0.006)

Large Employer	-0.206**	-0.212**	-0.228**	-0.235**	-0.229**
	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)
Large Location	0.058**	0.056**	0.066**	0.049**	0.059**
	(0.003)	(0.004)	(0.004)	(0.005)	(0.004)
Early Career	-0.015**	-0.024**	-0.020**	0.008	0.001
	(0.005)	(0.006)	(0.006)	(0.007)	(0.006)
pScore	0.081**	0.111**	0.063**	0.089**	0.085**
	(0.011)	(0.013)	(0.012)	(0.014)	(0.012)
Constant	0.500**	0.481**	0.535**	0.660**	0.504**
	(0.014)	(0.017)	(0.016)	(0.019)	(0.016)
Obs.	275858	202163	221801	161347	207000
R-squared	0.169	0.164	0.171	0.170	0.164
Adj R-squared	0.169	0.164	0.171	0.170	0.164
F	2801.290	1988.577	2288.985	1648.141	2024.070
RMSE	0.914	0.913	0.910	0.921	0.920

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

Table 31: The impact of attractiveness preference and belief bias with additional controls - (Model 1) original set of controls, (Model 2) Picture quality controls (Model 3) Picture quality, facial hair and face health controls (Model 4) Picture quality, facial hair, face health controls and picture characteristics from outside of cropped area.

	(1)	(2)	(3)	(4)
	rank ($r_{i,t}$)	rank ($r_{i,t}$)	rank ($r_{i,t}$)	rank ($r_{i,t}$)
Belief Bias	-0.005	-0.005	-0.005	-0.004
	(0.007)	(0.007)	(0.007)	(0.007)
Preference Bias	**0.013****	**0.013****	**0.013****	**0.013****
	(0.005)	(0.005)	(0.005)	(0.005)
MBA Graduation Year	0.288**	0.289**	0.293**	0.297**
	(0.015)	(0.015)	(0.015)	(0.015)
UG Graduation Year	-0.493**	-0.482**	-0.514**	-0.514**
	(0.030)	(0.030)	(0.030)	(0.030)
UG Fees	0.163**	0.165**	0.164**	0.162**
	(0.014)	(0.014)	(0.014)	(0.014)
MBA Fees	0.316**	0.314**	0.312**	0.307**
	(0.011)	(0.011)	(0.011)	(0.011)
Picture Blur		-0.049**	-0.047**	-0.020*
		(0.010)	(0.010)	(0.010)
Picture Exposure		0.097**	0.074**	0.075**
		(0.017)	(0.017)	(0.017)
Picture Noise		0.001	-0.001	0.005
		(0.008)	(0.008)	(0.008)
Picture Face Quality		0.031**	0.030**	0.018**
		(0.006)	(0.006)	(0.006)

Picture Resolution		-0.041**	-0.039**	-0.051**
		(0.006)	(0.006)	(0.006)
Picture Bald			-0.065**	-0.063**
			(0.008)	(0.008)
Picture Beard			-0.060*	-0.060*
			(0.029)	(0.029)
Picture Moustache			-0.062*	-0.057
			(0.029)	(0.029)
Picture Sideburns			0.078*	0.080*
			(0.032)	(0.032)
Picture Dark Circles			0.011	0.026**
			(0.010)	(0.010)
Picture Health			0.085**	0.076**
			(0.010)	(0.010)
Picture Skin Stains			0.045**	0.041**
			(0.007)	(0.008)
Picture Smile				-0.035**
				(0.007)
Picture Eye Glasses				0.032**
				(0.012)
Racy Clothing				-0.268**
				(0.078)
Informal Background				-0.007
				(0.013)
Formal Clothing				-0.076**
				(0.012)
Informal Clothing				0.008
				(0.033)
Picture Necklace				-0.001
				(0.010)
Picture Lipstick				-0.059**
				(0.016)
Picture Hat				0.026
				(0.038)
Picture Earrings				-0.113
				(0.075)
Industry Category (IT)	-0.123**	-0.121**	-0.120**	-0.119**
	(0.005)	(0.005)	(0.005)	(0.005)
Industry Category (Management)	0.400**	0.400**	0.401**	0.402**
	(0.005)	(0.005)	(0.005)	(0.005)
Industry Category (Others)	-0.400**	-0.399**	-0.397**	-0.395**
	(0.005)	(0.005)	(0.005)	(0.005)
Large Employer	-0.205**	-0.206**	-0.206**	-0.208**
	(0.004)	(0.004)	(0.004)	(0.004)
Large Location	0.063**	0.063**	0.063**	0.063**
	(0.004)	(0.004)	(0.004)	(0.004)
Early Career	-0.027**	-0.026**	-0.027**	-0.028**
	(0.005)	(0.005)	(0.005)	(0.005)
pScore	0.100**	0.093**	0.108**	0.112**

	(0.011)	(0.011)	(0.011)	(0.011)
Constant	0.238**	0.180**	0.174**	0.176**
	(0.022)	(0.025)	(0.025)	(0.028)
Obs.	275858	275858	275858	275858
R-squared	0.173	0.173	0.174	0.175
Adj R-squared	0.173	0.173	0.174	0.174
F	2398.652	1991.540	1611.199	1241.384
RMSE	0.912	0.912	0.912	0.911

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

Table 32: Relationship of job rank and likelihood of a clearly posted profile picture.

	(1) Profile Picture Posted
Rank	**-0.016**
	(0.015)
Gender (Female)	0.116**
	(0.032)
Gender (Male)	0.048
	(0.028)
Gender (Mostly Female)	0.037
	(0.054)
Gender (Mostly Male)	0.119**
	(0.045)
Gender (Unknown)	-0.054
	(0.032)
Ethnicity (Britsh-GreaterEuropean)	-0.062
	(0.051)
Ethnicity (EastEuropean-GreaterEuropean)	-0.030
	(0.067)
Ethnicity (GreaterEastAsian-Asian)	-0.449**
	(0.053)
Ethnicity (IndianSubContinent-Asian)	-0.121*
	(0.054)
Ethnicity (Jewish-GreaterEuropean)	-0.029
	(0.054)
Ethnicity (Muslim-GreaterAfrican)	-0.105
	(0.067)
Ethnicity (WestEuropean-GreaterEuropean)	-0.037
	(0.053)
Constant	1.749**
	(0.145)
Obs.	97771
R-squared	
Adj R-squared	

F
rmse

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

Table 33: (Model1) Impact of attractiveness without selection model (Model2) Impact of attractiveness with heckman selection model (Model3) Impact of belief and preference bias with heckman selection model.

	(1)	(2)	(3)
Gender (Male)	-0.004	-0.011	-0.004
	(0.047)	(0.034)	(0.039)
Attractiveness	0.085*	0.060*	
	(0.041)	(0.026)	
Preference Bias			0.065*
			(0.030)
Belief Bias			-0.010
			(0.027)
Graduation Year	0.003	0.004	0.005
	(0.004)	(0.004)	(0.006)
Select: Graduation Year		-0.020**	-0.027**
		(0.003)	(0.003)
Select: Gender (Male)		0.109**	0.107**
		(0.034)	(0.037)
Select: Constant		0.596**	0.514**
		(0.100)	(0.106)
rho		-.058	-0.054*
		(0.032)	(0.037)
sigma		0.995**	0.997**
		(0.012)	(0.014)
Obs.	1,198	3,082	3,082
Uncensored Obs.		1,198	1,198

Standard errors are in parenthesis
** p<0.01, * p<0.05

Table 34: Relationship of job rank and individuals attractiveness with interaction with MBA graduation year.

	(1)	(2)	(3)	(4)
	Attractiveness	Attractiveness	Attractiveness	Attractiveness
Rank	0.009	0.009	0.011	0.010
	(0.006)	(0.006)	(0.006)	(0.006)

MBA Graduation Year	0.076**		0.045**	
	(0.010)		(0.010)	
Rank * MBA Graduation Year	**0.000**	**0.001**	**-0.007**	**-0.006**
	(0.010)	(0.010)	(0.010)	(0.010)
Gender(Male)	0.014*	0.010		
	(0.006)	(0.006)		
Ethnicity (European)	-0.047**	-0.048**		
	(0.013)	(0.013)		
Ethnicity (Others)	-0.179**	-0.176**		
	(0.014)	(0.014)		
UG (Business)	0.009	0.014		
	(0.013)	(0.013)		
UG (Science)	-0.042**	-0.041**		
	(0.008)	(0.008)		
UG (Others)	0.084**	0.083**		
	(0.009)	(0.009)		
MBA Rank	0.002	-0.002		
	(0.007)	(0.007)		
UG Rank	-0.066**	-0.065**		
	(0.007)	(0.007)		
Large Employer	0.015**	0.017**		
	(0.005)	(0.005)		
Large Location	0.010*	0.002		
	(0.005)	(0.005)		
Constant	0.549**	0.598**	0.489**	0.514**
	(0.017)	(0.016)	(0.006)	(0.003)
Obs.	38129	38129	38129	38129
R-squared	0.023	0.022	0.001	0.000
Adj R-squared	0.023	0.021	0.001	0.000
F	69.039	70.272	9.290	3.882
RMSE	0.494	0.494	0.500	0.500

Standard errors are in parenthesis
** p<0.01, * p<0.05 .

Chapter 2:
Why Bitcoin will Fail to Scale?
Economics of Collusion on Blockchains

Bitcoin falls dramatically short of the scale provided by banks for payments. Its ledger grow by the addition of blocks of $\sim$ 2000 transaction every 10 minutes. Intuitively, one would expect that increasing the block capacity would solve this scaling problem. However, we show that increasing the block capacity would be futile. We analyze strategic interactions of miners, who are heterogeneous in their power over block addition, and users, who are heterogeneous in the value of their transactions, using a game-theoretic model. We show that a capacity increase can facilitate large miners to tacitly collude – artificially reversing back the capacity via strategically adding partially filled blocks in order to extract economic rents. This strategic partial filling crowds out low value payments. Collusion is sustained if the smallest colluding miner has a share of block addition power above a lower bound. We provide empirical evidence of such strategic partial filling of blocks by large miners of Bitcoin.

We show that a protocol design intervention can breach the lower bound and eliminate collusion. However, this also makes the system less secure. On the one hand, collusion crowds out low-value payments; on the other hand, if collusion is suppressed, security threatens high-value payments. Thus its untenable to include range of payments with vastly different outside options, willingness to bear security risk and delay onto a single chain. Thus, we show economic limits to the scalability of Bitcoin. Under these economic limits, collusive rent extraction acts an effective mechanism to invest in platform security and build responsiveness to demand shocks. These traits are otherwise hard to attain in dis-intermediated setting owing to the high cost of consensus. We discuss recommendations to public Blockchain designers.

1. Introduction

Bitcoin has been touted as a revolutionary technology that would disrupt the traditional payments industry (Popper 2017, Tasca 2018). Bitcoin's market capitalization stood at $100 billion in Oct 2018 (Zaitsev 2018), and it is being used for transactions worth $1 billion on a daily basis (BlockchainExplorer 2018). In June 2018, Bitcoin was used for a $300 million payment completed in 10 minutes at a fee of 4 cents (Blockchain.com 2018). In comparison, transactions through banks as intermediaries are costly, and cross-border transactions take several days to complete (TransferWise 2018). In recent times, these slow and costly banking intermediaries have also faltered at providing security against data hacks (McMillan 2018, Glazer and Farrell 2018) and frauds (Glazer 2018). As a result, Bitcoin has garnered enormous attention with promise of inexpensive, fast and *trustless* payments (Jordan and Kerr 2018).

Traditional banks maintain a ledger of transactions. Users make payments by instructing banks to add desired transactions onto this ledger. Banks guarantee integrity (no one can spend more than their balance) and security (outsiders cannot steal balances). Bitcoin maintains a similar ledger, but it is maintained by consensus among a peer-to-peer network. Bitcoin's current technology is capable of adding only up to 3 transactions per second (Eyal et al. 2016) to this ledger, compared to VISA's 5000 transactions per second (Yli-Huumo et al. 2016). In December 2017, this limited throughput fell well short of demand; a vast majority of users were turned away, and a large number of users were unable to spend any of their Bitcoins. On December 21, 2017, under severe congestion, Bitcoin witnessed users offering up to $54.90 to obtain preferential treatment on this limited throughput (Buntinx 2017).

Bitcoin's ledger consists of a chain of blocks. Blocks with a limited size of 1 MB (~ 2200 transactions) are added every 10 minutes. This low frequency of block addition is required to propagate a new block over a large network to keep all participants in sync. As a result, to increase the transaction throughput, Bitcoin developers have been aggressively debating technology upgrades for increasing the block capacity. Segwit2x was an attempted upgrade of Bitcoin to 2 MB blocks in Nov 2017 (Bitcoin Wiki 2017). Meanwhile, Bitcoin Cash was launched as a competing platform in Aug 2017 with an 8 MB block size (Wilmoth 2018). In this paper, we ask a fundamental question: Would a simple redesign (e.g., 2x block size or 0.5x block duration) allow Bitcoin to scale?

Counterintuitively, we show that increasing the capacity (e.g., doubled block size or halved block duration) would be futile in terms of scaling Bitcoin. Thousands of participants, called "miners", who validate and add transactions to the Bitcoin ledger collude to revert back increased Block capacity. These miners in large numbers are meant to provide ledger integrity and security. Miners spend computational resources to add transactions for which they collect a fixed block reward plus any transaction fees offered by the users. Although transaction fees are optional, miners can chose which transactions to process and prioritize those that pay high transaction fees. As a result, users tend to offer high transaction fees when the block capacity is well below the demand to obtain preferential treatment. However, when the demand competition decreases, users do not face the same risk of being left out. As a result, they offer near-zero transaction fees. We show that under such a scenario, a large number of miners could enter a tacit collusion to artificially lower the effective block size to raise transaction fees. They would strategically add partially filled blocks by including only those transactions that offer high transaction fees. We identify one such equilibrium condition where such tacit collusion is sustained. This equilibrium requires the computing power of the smallest colluding miner to be larger than a threshold. A large miner invests more in computational power and, as a result, wins blocks more frequently. These large miners are most suited to sustain collusion via a long-term punishment threat. In contrast, small miners win so infrequently that they are not deterred by future punishment and thus deviate from collusion.

There are two aspects of Bitcoin's mining process that are necessary for such a collusive equilibrium. First, the public ledger allows anyone to identify which miner added a particular block and their choice of full or

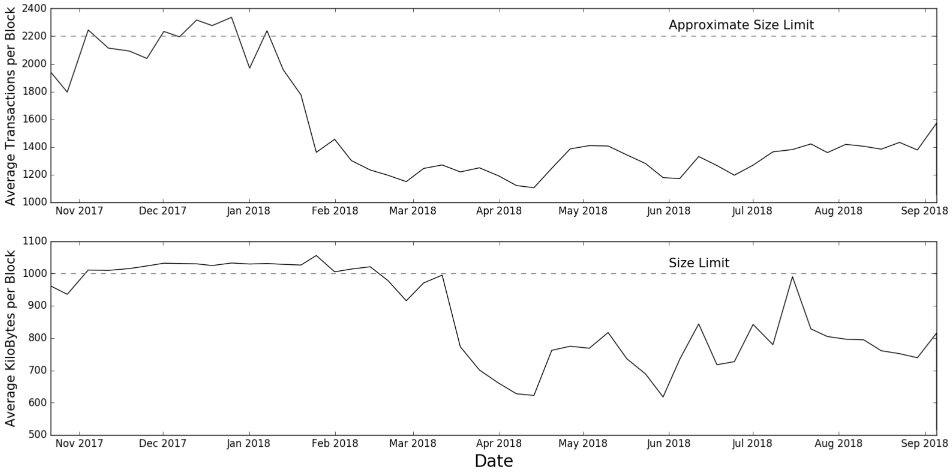


Figure 1 *The average block size has been significantly below full capacity since January 2018, shown in terms of the number of transactions (top) as well as the actual size in KB (bottom).*

partial block fill. This allows other miners to verify if a colluding miner has deviated. Second, the miners are in the game for perpetuity. The group can punish the deviating miner by getting into a transaction fee war in perpetuity, making all miners worse off, which ensures that no colluding miner deviates.

A closer examination of Bitcoin' ledger confirms the above intuition – miners indeed partially fill their blocks when demand competition is weak. Figure 1 shows block sizes pushing on the ~ 2200-transaction ceiling around December 2017. Since then, block sizes ($\sim 500 - 1500$ transactions) in 2018 have been well below their maximum limit. This observation seems to suggest insufficient interest to make payments via Bitcoin. In reality, user payments are still being turned away as empty blocks are added to the ledger. Numerous partially filled blocks have recently been observed regularly, even when tens of thousands of transactions were waiting to be processed. Figure 2 shows a very specific instance on September 31, 2018, when AntPool (which is a very large mining pool) added partially filled blocks. The first two blocks used up less than 10% of the block capacity, while more than 1000 transactions were in the queue. It is critical to note that very few transactions in waiting were offering a high fee ($>$ 15 satoshi [1]/Bytes) when these blocks were added. The third block utilized its full capacity when a significant number of high-fee-offering transactions were in the queue.

Figure 3 reports the criticism received by AntPool on Twitter for partially filling their blocks when thousands of transactions were waiting to be processed. Figure 4 provides the reply of an AntPool representative

[1] 1 Bitcoin $= 1 \times 10^8$ satoshi

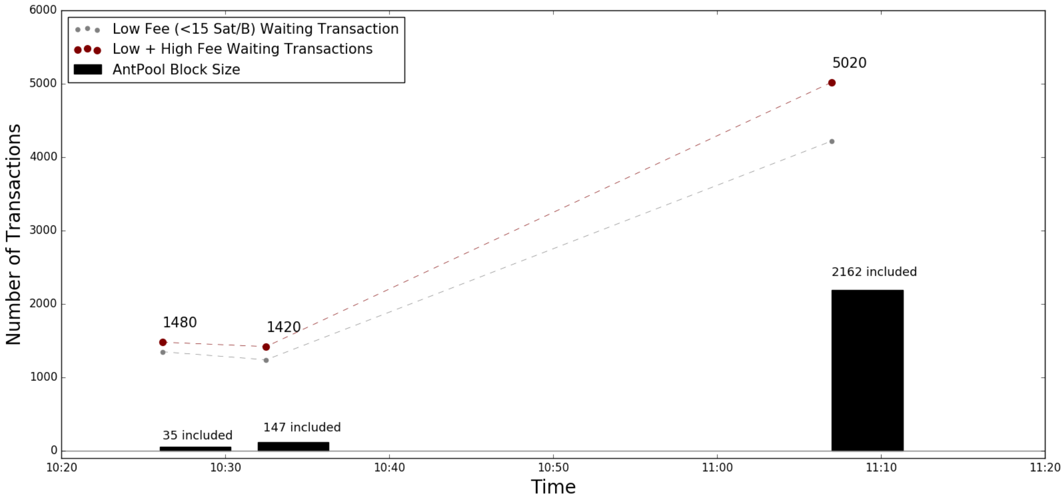
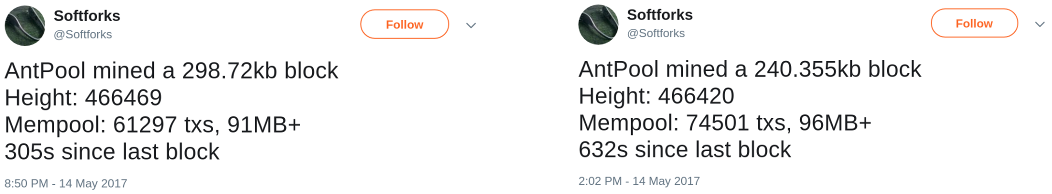


Figure 2 *Three blocks mined by AntPool on September 31, 2018. The first two blocks used up less than 10% of the block capacity, while more than 1000 transactions were in the queue. The third block utilized its full capacity when a significant number of high-fee-paying (> 15 satoshi/Byte) transactions were waiting, as indicated by the dark gray region.*

who admits to this practice and responds to the criticism by saying that their actions are within the rights provided by the consensus mechanism of Bitcoin. [2] We will show that the strategy of sacrificing present capacity for future earning surges is rational for large miners only. In line with this finding, Figure 5 shows that AntPool (which currently holds 20% of the overall mining power) has been underfilling blocks compared to all other miners.



AntPool mined a 298.72kb block
Height: 466469
Mempool: 61297 txs, 91MB+
305s since last block

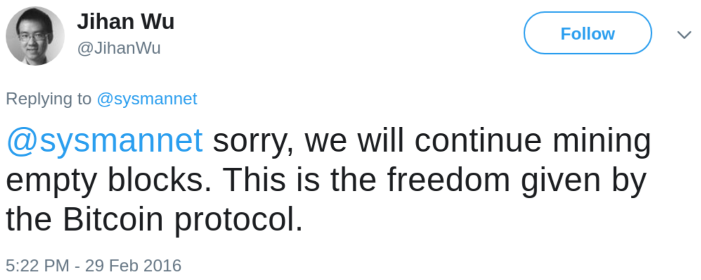
8:50 PM - 14 May 2017

AntPool mined a 240.355kb block
Height: 466420
Mempool: 74501 txs, 96MB+
632s since last block

2:02 PM - 14 May 2017

Figure 3 *Two instances where a large bitcoin mining pool (AntPool) was criticized on Twitter for mining partial blocks (298 KB and 240 KB) while more than 60000 transactions were waiting in the queue.*

[2] An alternative explanation for the partially filling of blocks could be related to the computation time advantage. A miner is required to prepare a new block with a set of awaiting transactions. This requires a small amount of computation time to validate all newly added transactions. Some miners prefer to forego the verification and therefore the transaction fees by preparing a small block to start puzzle solving as soon as possible. The validation time and puzzle-solving time trade-off should be equally beneficial for all miners. Partial block filling by a select few large miners contradicts this explanation.



Jihan Wu
@JihanWu

Replying to @sysmannet

@sysmannet sorry, we will continue mining empty blocks. This is the freedom given by the Bitcoin protocol.

5:22 PM - 29 Feb 2016

Figure 4 *The figure shows a reply by an AntPool representative with regard to criticism concerning the partial blocks.*

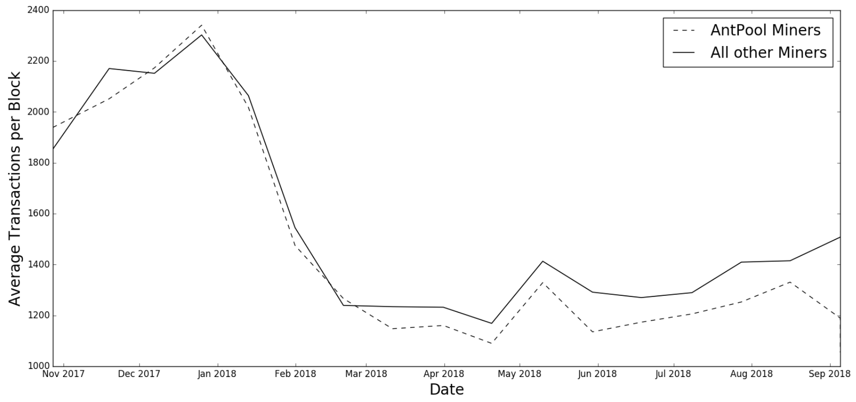


Figure 5 *In periods of high demand (Pre-January 2018), AntPool fully fills their blocks similar to any other miner. In periods of low demand (Post-January 2018), AntPool has underfilled blocks compared to those of all other miners.*

Miner collusion would effectively reverse back any capacity increase, which resembles an intermediary unilateral action to extract excess economic rents. The Bitcoin community can respond to miner collusion via technology interventions. Banning large miners is one potential intervention to make collusion unsustainable. We show that while eliminating collusion (via banning miners) may increase throughput and lower transaction fees, this makes the system less secure from double-spend attacks. In a double-spend attack, a miner sacrifices block revenue for a small probabilistic shot at stealing a payment value. This sacrifice of block revenue is a small price to pay once the collusion is eliminated. Powerful miners would then prefer to launch double-spend attacks rather than earn these small revenues. Eliminating collusion removes artificial capacity constraints; however, the security threat drives away the demand for transactions. Thus, these collusion and security threats place economic bounds on Bitcoin's scale. In fact, we will demonstrate a different perspective where collusion is a necessary endogenous mechanism for the participants to coordinate investment in security or respond to temporary exogenous demand fluctuations.

Our work has four major contributions. First, we provide economic limits to Bitcoin's scale. Industry debate and some early research have largely focused on technological cost of repetitive information propagation to keep multiple copies of the Blockchain ledger in sync. We instead focus on inability of decentralized platforms to offer differentiated service tiers to spectrum of users with dissimilar outside options, and willingness to bear security risk and delay. Second, we argue that decentralization and its many facets (number of miners, cumulative mining power, miner homogeneity) are not simultaneously achievable or even desirable. Third, we are one of the first to simultaneously model rational equilibrium decision choices for users and miners. We are the first to suggest miners' economic incentives to tacitly collude for coordinating on efficient security funding. Existing research largely focuses on strategies for one of these agents in isolation. We show how user preferences (banks vs Bitcoin) impact miner earnings and how miner block filling impacts user fees and security. These externalities are highlighted because we model the two sets of agents in a single game. Finally, we argue the generalizability of our findings and design suggestions to major peer-to-peer payment blockchains that broadly follow the paradigm kickstarted by Nakamoto 2008. We also expect elements in our model to form a basis for future I/O research in Decentralized Apps, smart contracts, and decentralized and dark web markets.

In the next section, we describe the Bitcoin ledger and consensus mechanism in greater detail and highlight relevant prior literature on the economics of Bitcoin. Section 3 presents the model to illustrate Bitcoin's transaction fees as an outcome of user competition to earn space on the limited ledger throughput. We extend this base model to show the equilibrium collusion strategy sustained by a group of large miners. Section 4 discusses potential methods to break the collusion and the perilous double-spend security implications. Section 5 discusses the robustness of our findings to a wider range of available blockchain design choices and provides guidance on potential redesign that tackles the underlying problem, i.e., accommodating users with vastly different willingness to pay for delay and security onto a single chain.

2. Background

A simplified example of Bitcoin's payment transaction ledger is shown in Figure 6. The ledger is broken into blocks, where each block records details of transactions including the sending and receiving account information, transaction fees offered by each sender for a transaction, and the cryptographic signature of the sender, which is needed to verify that the transaction is genuine. This ledger is secured by a group of *miners*, i.e., any willing individual with compute power connected to the Internet. A new block is added to the ledger on average every 10 minutes. Miners compete for the privilege of adding a block of transactions by solving a computationally expensive cryptographic puzzle (Nakamoto 2008). The cryptographic puzzle used by Bitcoin requires the miners to find a rare SHA256 hash, which is time consuming to identify even for the fastest available computing hardware. Each block also stores the hash of the previous block, which ensures that blocks are added chronologically. The fastest solver is able to add the next block on top of the existing chain. This miner broadcasts his or her new block to the network. All other miners update their copies of the ledger after validating the puzzle solution as well as the block content. If a majority of miners accept the block, the state of the ledger is seen to be updated. These block addition steps are repeated infinitely to grow the ledger.

The balance in an account is determined by adding up all transactions that appear in the unique longest chain of blocks. Because Bitcoin is a peer-to-peer network and because block propagation takes time, two miners who are far from each other in the network can often find a solution to the puzzle at approximately the same time. Both miners could propagate the solution to peers, which may lead to parallel chains. This situation is resolved endogenously. A chain with support from the miners with the largest combined computing power grows faster than others, eventually becoming the longest chain. This is because higher computing power leads to faster puzzle solving and therefore more block additions. Once a longer chain of blocks emerges, all shorter chains are typically abandoned.

Users make payments by broadcasting desired transactions to all miners. Because of the fixed block size, not all transactions pending at a given time can make it into the next block. The consensus mechanism of Bitcoin provides a miner full autonomy over which pending transactions to include in a block. Users hope that a sufficiently high fee offer will incentivize any miner to include their transactions over other pending transactions (Huberman et al. 2019, Easley et al. 2019). Thus, unlike banks, Bitcoin does not set a fee. Users seeking transaction processing enter an auction to find a spot on the limited throughput (transactions per sec) of the Bitcoin ledger. The winning miner collects these transaction fees. Every block also records a transaction called the “generation transaction”. The winning miner is awarded with a block reward (newly minted Bitcoins), which is recorded as a generation transaction. The fees offered by all transactions in a block plus the newly minted Bitcoin make up revenues to incentivize miner participation.

As a miner invests more in computing power, they become increasingly faster at solving the mining puzzle. A select few miners have gained access to high-quality ASIC hardware specialized for Bitcoin mining. These large miners make up a large proportion of the total computing power. The remaining computing power comes from a large crowd of small miners utilizing inferior hardware (e.g., GPUs and CPUs). The mining network overall has become faster at solving puzzles due to the increase in computational power devoted to mining Bitcoins. The puzzle difficulty is therefore adjusted endogenously to keep the average time to

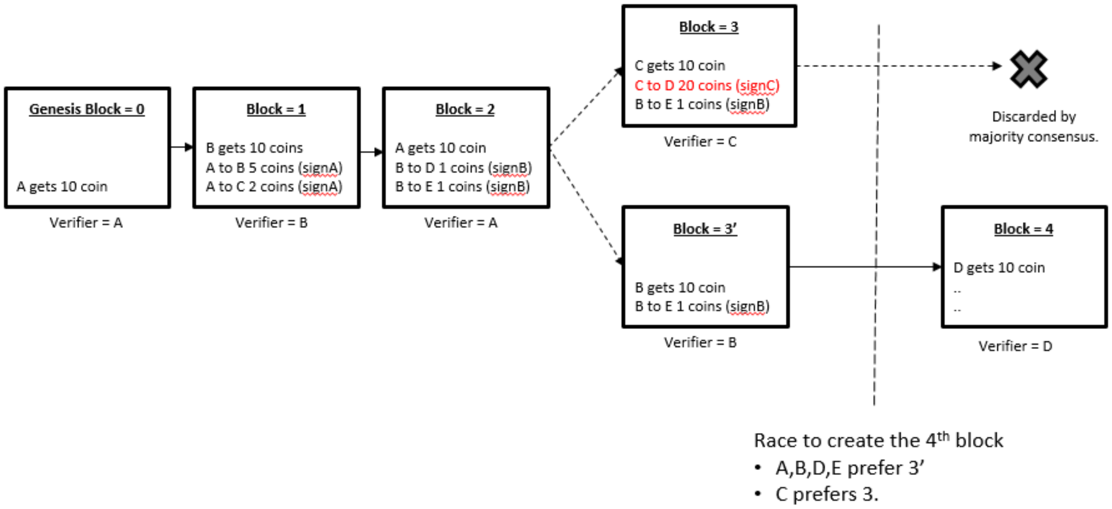


Figure 6 *Left to Right: The 0th or Genesis block records a single fixed block reward transaction. The 0th, 1st and 2nd blocks are mined by miners A, B and A, respectively. These respective miners are the quickest to solve those three puzzles. Each miner receives 10 coins as a fixed reward for block creation. They also collect fees offered by respective transactors. The 3rd block is proposed by C. Because of an invalid transaction whereby C attempts to spend 20 coins that they do not have, this block is rejected by all other miners. An alternative 3rd block is then proposed by the 2nd fastest puzzle solver B. The chain is extended on top of this block by a majority of the mining community. The rewards earned by miner C on their proposed block do not form a part of the longest unique chain.*

a solution at 10 minutes. A natural question arises: Why not have 1 minute between blocks? The puzzle-winning miner needs to transmit a new block to the entire mining network. Because of the peer-to-peer nature of the network, blocks propagate slowly, and a short duration allows too little time for all miners to sync with the new state of the ledger. A shorter time between blocks could lead to the creation of multiple forks as new blocks are being proposed faster than any single miner can sync with their peers. In addition, users would need to wait for a longer chain to emerge to determine whether their transactions have made it to a block in the longest chain. Hence, reducing the time between blocks would not necessarily reduce the effective transaction validation time. While our research focuses on the economic limit, the above discussion should underscore that it remains a technological challenge as well.

2.1. Prior Literature

The economics of blockchains has gained attention relatively recently. Halaburda and Haeringer 2018 provide an excellent literature review. Our work focuses on the trade-off between capacity, security and decentralization. Four key areas of research within economics of blockchain intersect with our model: user transaction fee models, miner entry models, collusion and security. Huberman et al. 2019 and Easley et al. 2019 were the first to model **user transaction fees**, and they serve as the starting point for our model. Both of these papers model users as heterogeneous in blockchain fee savings and delay costs. However, unlike our approach, neither users nor miners are rational or strategic about block fill levels and security. The goal of Huberman et al. 2019 is to explain why transaction fees exist even when the average blockchain capacity is greater than average demand. They point to stochasticity in demand and block arrival rates as the reasons. Our goal is to show why transaction fees exist even when the first mechanism provided by Huberman et al. 2019 is absent. Therefore, we assume nonstochastic demand arrival and block addition rate. This theoretically eliminates the primary mechanism described by Huberman et al. 2019, allowing us to show the secondary phenomenon.

A majority of the literature models **miner entry** with zero profits resulting in small homogeneous miners (Huberman et al. 2019, Easley et al. 2019, Prat and Walter 2018, Abadi and Brunnermeier 2018). Huberman et al. 2019 conclude that "small miners cannot affect user behavior" and that the "protocol lacks an easily workable mechanism to change prices, offerings and rules ...". Abadi and Brunnermeier 2018 say that "coordination requires dynamic punishment schemes, but in an environment without rents, such punishments are infeasible". We show participants' endogenous ability to make artificial protocol adjustments without designer action. This relies on the presence of large miners via a degree of miner heterogeneity, which is an outcome of the skewed distribution of mining hardware and cost efficiency, also argued by Arnosti and Weinberg 2018.

Cong and He 2019 and Malinova and Park 2017 highlight the unique public address identity on Bitcoin, which allows users to consistently exhibit their actions while maintaining anonymity. This is a crucial feature that allows small miners to sustain a grim trigger **collusion** in our model. However, unlike us, few existing examples of collusion consider the price of quantity setting among firms that use blockchain for payments instead of miners/gatekeepers of the blockchain itself. This also differentiates our work from standard collusion in Cournot markets (Green and Porter 1984). It is worth highlighting two key differences: (i) In the standard Cournot setting, all firms decide supply quantities simultaneously. In our setting, an individual firm makes a quantity decision for the entire market at different time slots. (ii) In the standard Cournot setting, all users observe the same quantity and price in the market. In our setting, a user bids first, which leads to user discrimination by private valuations.

The academic literature has discussed a few adversarial attack strategies – double-spend attack (Sompolinsky and Zohar 2015a), selfish mining attack (Nayak et al. 2016a), eclipse attack (Natoli and Gramoli 2017), etc. Gervais et al. 2016 simulate gains for double-spend attacker under different blockchain design choices. We build upon a simple double-spend attack strategy to analytically formulate economic trade-offs for an attacker – honest mining revenues versus low-probability attack payoff. Budish 2018 consider a similar trade-off, but they focus on settings that would make the trade-off unappealing to the attacker, e.g., non-repurposable mining hardware. Contemporary work by Chiu and Koeppl 2017 is closest to ours in considering different levels of **security** induced and security desired by small versus large payments. Unlike our approach, they focus on block rewards as the primary source of miner revenues and simplify transaction fees as exogenously given rather than the endogenous outcome of the auction and collusion. In fact, this leads them to some contrasting conclusions. We try to elucidate how these two works should be viewed in conjunction.

Finally, we add to the growing literature on information security economics (Gao et al. 2013, Cezar et al. 2013, Kannan and Telang 2005, Arora et al. 2007, August et al. 2014, Hsu et al. 2012, Dey et al. 2018). This literature focuses on the strategic interactions of firms and hackers. Our work is a unique setting whereby a firm is replaced by decentralized miners and users. Our work also intersects with the literature on P2P platforms (Asvanund et al. 2004, Wei and Lin 2016, Li and Agarwal 2016). Johar et al. (2011) study participant equilibrium sharing responses to lower congestion in P2P platforms. In Bitcoin, miners control the supply side; they instead prefer to increase congestion in this P2P setting.

3. Model

Our model captures three desirable features of Bitcoin – (i) Capacity: Large number of payments at low fees and small delay, (ii) Security: Low likelihood of payment value being hacked (double spent) and (iii) Decentralization: Zero-barrier entry of large number of miners. In this and the next section, we analyze trade-offs between capacity and security, treating decentralization to be a static requirement for a peer-to-peer blockchain-based payment solution. We consider the impact of both protocol design choices (e.g., block size, block duration, mining puzzle) and endogenous participant strategies (e.g., user fee offers and miner block filling) on the capacity-security trade-off.

3.1. Baseline Model

We model Bitcoin's blockchain as an infinitely repeated block creation game between miners and users. Users decide whether to use the blockchain for their payment needs and what fees to offer. Heterogeneous miners decide whether to participate in block creation and which user transactions to include on blocks. Our baseline model obtains the equilibrium transaction fee offered by users and the endogenous entry decision by miners.

Users: All users realize the need for a new payment value v drawn independently from a uniform distribution $g(v)$ with domain $[0, V_{max}]$ every period. A user can complete his or her transaction using either the blockchain or an off-chain fiat option. This choice is based on a comparison of (i) fees, (ii) delay and (iii) security risk on the two channels. **Fee** on the off-chain option is modeled as a proportional (ρv) fee structure. The proportional nature of this cost is driven by payment value-dependent factors such as fx margins and liquidity guarantee[3]. The fee on the blockchain, denoted by f, is endogenously determined by the demand and capacity of the blockchain.

Delay on the blockchain is the result of duration d between the addition of blocks of payments. This delay decreases the user's utility by a factor of $\delta = \delta(d)$ (< 1), i.e., the discount factor is increasing in duration d. Further, users also face an uncertainty in successful inclusion of their transaction on the blockchain. The platform has a hard upper limit on the number of transactions n_F that can be included on a single block. Users who take the on-chain option broadcast their transaction to all the miners with a fee offer. Next, miners prepare a block independently by selecting transactions from the waiting queue up to the blockchain's capacity. The block prepared by the puzzle-winning miner is added to the blockchain. A user's transaction may not be picked up on the new block. In this case, they revert to the off-chain option, albeit with delay (δ).

Security cost on the blockchain arises from a possibility of payment reversal[4]. Consider a user making a payment to purchase a product from a merchant. The merchant normally delivers the product after they see the payment recorded onto a block. In Section 4, we elaborate the likelihood that a (double-spend) attack *reverses* this payment after the product is delivered. The user needs to insure the merchant against a future (double-spend) attack. However, double-spend attacks are costly to execute for an attacker. A transaction should be of high enough value to attract a double-spend attack. Let V_{secure} represent the value of a transaction such that none of the transactions with $v \leq V_{secure}$ are attractive to the attacker. The remaining transactions with $v \geq V_{secure}$ face a likely threat. An attack effectively reverses the target payment after the merchant delivers the product. Let $S(v)$ represent this cost of insuring the merchant against such a reversal or double spend[5]. Then, we have

$$S(v) = \begin{cases} 0, & \text{if } v < V_{secure}, \\ v & \text{else} \end{cases} \tag{1}$$

This formulation implicitly assumes that payment $v \geq V_{secure}$ will be double spent by an attacker with certainty[6] and therefore need to be insured in full. In Section 4, we endogenize V_{secure} as rational expectation of (double-spend) security attack likelihood. For now, this is treated as exogenous and known to all users.

A combined consideration of fees, delay and security allows us to formulate users' utility on the available payment channels. Off chain, the user derives a utility $v - \rho v$. On chain, the user derives a utility of

[3] Shy and Wang 2011 show why traditional intermediaries prefer proportional fees to maximize profits. In practice, different off-chain options vary on proportional rate and time delays. Appendix A provides the motivation behind a single off-chain fee structure.

[4] We assume zero security risk off chain because intermediaries insure users against such risks

[5] For simplicity, we assume that the insurance is provided via some off-chain secure method

[6] In Appendix D, we discuss probabilistic threat

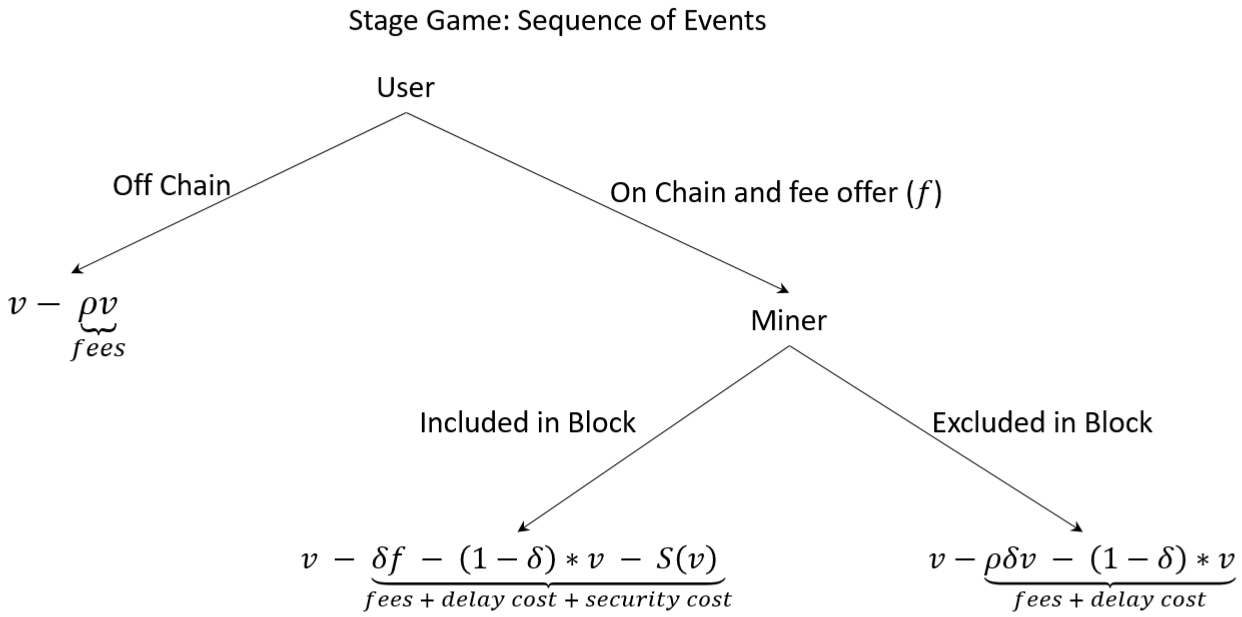

Stage Game: Sequence of Events
User
Off Chain
On Chain and fee offer (f)
$v - \underbrace{\rho v}_{fees}$
Miner
Included in Block
Excluded in Block
$v - \underbrace{\delta f - (1-\delta) * v - S(v)}_{fees + delay\ cost + security\ cost}$
$v - \underbrace{\rho\delta v - (1-\delta) * v}_{fees + delay\ cost}$

Figure 7 *The figure shows the single-period extensive form game between the users and miners that is infinitely repeated at every block creation period.*

$(v-f)\delta - S(v)$. This can be interpreted as fixed fees δf, delay cost $(1-\delta)*v$ and security cost $S(v)$ for completing payment v. If excluded on the chain, the user derives a utility of $(v-\rho v)\delta$. This can be interpreted as a discounted proportional fee $\delta\rho v$ and delay cost $(1-\delta)*v$. The user's utility from the available choices is shown in figure (7).

Let $\bar{f}$ represent the maximum fee that a user is willing to offer on chain, i.e., this is the fee offered by the user who is indifferent between on-chain and off-chain options. Thus, solving

$$v - \rho v = (v - \bar{f})\delta - S(v)$$

for $\bar{f}$, we obtain

$$\bar{f} = \frac{v(\rho + \delta - 1)}{\delta} - \frac{S(v)}{\delta}.$$

Users in high-risk regions (i.e., $v > V_{secure}$) will never find an on-chain channel viable for any positive fee offer f because $\bar{f} < 0$ for all $v > V_{secure}$. Users in the zero-risk region make up the overall demand for the blockchain. Define $V = \min\{V_{max},\ V_{secure}\}$. The payment-value uniform distribution $g(v)$ has a domain $[0, V_{max}]$. If $V_{max} \geq V_{secure}$, then the highest value vying for an on-chain payment is $V = V_{secure}$. Otherwise,

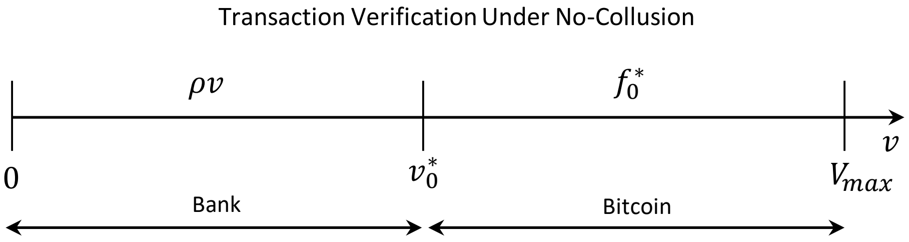


Figure 8 *Users with a transaction value higher than v_0^* transact on the Bitcoin network.*

if $V_{max} < V_{secure}$, then the highest value vying for an on-chain payment is $V = V_{max}$. The total on-chain payment demand is given by[7],

$$\int_0^V g(v)dv \text{ ; where } g(v) = d\frac{N_{max}}{V_{max}} \tag{2}$$

$$\int_0^V g(v)dv = d\frac{N_{max}}{V_{max}}V = N(d,V) = N \tag{3}$$

where $N(d,V)$ is the total number of users that draw a payment need between 0 and V in $d = 10$-minute block duration. We simplify the notation $N(d,V)$ to N since d and V are held constant for large parts of our model.

Figure 7 represents the sequence of events in a block creation game. We first investigate subgame perfect equilibrium (SPE) strategies for a single stage using backward induction. Miners will pick the top-n_F transactions in the order of fee offers if more than n_F users offer on chain fees. In an auction-like manner, users determine their fee offer based on rational beliefs of competing offers. Note that $\bar{f}$ is monotonically increasing for $v \in [0,V]$, i.e., users with a large value payment stand to gain most by avoiding the off-chain channel. It follows that in equilibrium, user fee offers will be increasing in $\bar{f}$ and therefore in v, i.e., $f(v_1) \geq f(v_2)$ if $v_1 \geq v_2$. Thus, we investigate equilibrium where the winning miner includes the top-n_F transactions in the order of payment value v, which is same as the order of fee offers.

We define the following constants, which will be used throughout the paper:

$$\gamma = \frac{n_F}{N},\ \alpha_h = \frac{\delta+\rho-1}{\delta\rho},\ \alpha_l = \alpha_h\left(\frac{1}{\gamma}-1\right),\ \text{and } \beta = \frac{(1-\gamma)(\alpha_h-\alpha)}{\alpha}. \tag{4}$$

Let v_0^* be the transaction value such that there are exactly n_F users that transact a greater value. Since N transactions are uniformly distributed between 0 and V, we have

$$\begin{aligned} v_0^* &= V\left(1-\frac{n_F}{N}\right), \\ &= V(1-\gamma), \end{aligned} \tag{5}$$

[7] A different distribution (e.g., exponential $g(v) = dN_{max}\lambda e^{-\lambda v}$) instead of uniform does not change our core findings. We discuss this in Appendix D

where γ is defined in (4). All users with the top-n_F values $(v \geq v_0^*)$ offer a fee such that the remaining users $(v < v_0^*)$ are driven out. The marginal user with $v = v_0^*$ is indifferent between the on-chain and off-chain option. Thus, we have

$$(v_0^* - f_0^*)\delta = v_0^*(1-\rho) \tag{6}$$

where f_0^* is the transaction fee offered by the user with transaction value v_0^*. Note that security cost $S(v) = 0$ for payment domain $[0, V]$. Using the above equation and expression of v_0^* from (5), we obtain

$$f_0^* = \rho\alpha_h(1-\gamma)V, \tag{7}$$

where α_h is defined in (4). Let R_0 represent the transaction fee revenue earned by miners. Then, we have

$$R_0 = f_0^* n_F = \rho\alpha_h(1-\gamma)\gamma VN. \tag{8}$$

Proposition 1 *Bitcoin limits entry of low-value users ($v < v_0^*$) due to the competitive fee auction for limited capacity n_F. It limits entry of high-value users ($v > V$) due to the security threat. The following inferences ignore collusion and security issues to be discussed later,*

- *As capacity n_F increases, Bitcoin fee f decreases. User cost savings are greatest at very high block size ($n_F \to N$)*[8]*.*
- *As block duration d increases, the Bitcoin fee decreases but delay disutility increases. User cost savings are greatest at very low block duration ($d \to 0$).*
- *As the proportional rate ρ levied on the off-chain option increases, the Bitcoin fee increases.*

Note that in practice, coinbase block rewards B distributes additional currency supply to miners. This is an additional cost borne by users. We model $B = 0$ because (i) most blockchains have a diminishing block reward schedule eventually expected to decrease to zero anyway. A perpetual block reward design, which leads to infinite supply, does not have a major practical example yet. (ii) Transaction fees are unavoidable even if a blockchain had a perpetual block reward because they provide an endogenous mechanism for prioritizing payments. It avoids flooding the blockchain with 1 cent payments. (iii) Our primary focus is user choice for means of payment not store of value. Therefore, we want to abstract away from myriad issues related to cryptocurrency valuation, e.g., investors, speculators, and fiat currency to cryptocurrency exchange rates. Chiu and Koeppl 2017 make the opposite modeling choice, i.e., abstracting away the transaction fee model, instead of delving deeper into endogenous user demand for coin holdings. In Appendix E, we discuss extension of our model with block rewards ($B > 0$) and contrast with Chiu and Koeppl 2017.

[8] Appendix E provides user cost charts for a more detailed explanation

Table 1 *Power and costs for popular computing hardware. ASICs are more expensive but have exponentially higher power.*

	h(m)	**c(m)**	
	Hash Power (GH/sec)	Power Consumption (Watt)	Hardware Cost (USD)
AntMiner S9 (ASIC)	14000	1375	2400
Avalon Batch 1 (ASIC)	66.3	620	1300
NVIDIA GTX 460 (GPU)	0.127	340	200
Intel Corei-5 (CPU)	0.014	95	82

Miners: We model miners as heterogeneous in terms of the quality of computing hardware[9]. We assume that when arranged in decreasing order of hardware quality, a miner with rank m has a mining power $h(m)$ (number of cryptographic hashes performed per second), which runs at a periodic cost $c(m)$ (USD per second). This hardware distribution $h(m)$ over rank is a monotonically decreasing convex function[10] with domain $m \in [0, \infty)$ and $h(\infty) = 0$. The convexity represents a long tail, i.e., a large number of potential miners own a small CPU. A few miners have access to high-quality ASIC hardware. High-quality hardware $h(m)$ also consumes greater cost $c(m)$. This periodic cost $c(m)$ is a combination of running costs (e.g., electricity) and the amortized cost of the hardware purchase. Table 1 provides a small comparison of computing hardware.

The best hardware (ASICs) costs up to 10 times more than a CPU but can deliver thousands of times more hash power. The hash power delivered is disproportionately larger than the higher cost. If this were not true, simply purchasing multiple CPUs would be preferable to buying a single ASIC.

$$m_1 < m_2 \implies h(m_1) > h(m_2); \quad c(m_1) > c(m_2); \quad \frac{h(m_1)}{c(m_1)} > \frac{h(m_2)}{c(m_2)} \tag{9}$$

The running cost component does not change any of our analysis. Without loss of generality, we set this cost to zero going forward. We focus on the upfront cost of hardware purchases, which is amortized periodically at c (USD/period). This is equivalent to an upfront fixed cost of $\frac{c}{1-\delta}$. This upfront fixed cost plays a significant role in miners' decisions to participate in Bitcoin mining.

Miners join the mining network to earn revenues from transaction fees (R) offered by users and block rewards (B) fixed in the protocol. A large number of hashes per second $h(m)$ means a greater likelihood of

[9] We model consensus achieved by commitment of computing power (PoW). Alternative p2p payment blockchains may use a different (costly and well-distributed) commitment resource, e.g., mining equipment, coin stake, and storage capacity. In all of these settings, miners must purchase mining equipment or coin stake upfront for frequent opportunities to decide ledger additions. The underlying intuition should remain valid. The heterogeneous miner enters up to a marginal miner who has a cost of purchasing commitment resource that is just low enough. We do not rigorously incorporate PoS into our model because (i) practical examples of PoS blockchains are still rare. (ii) Blockchains that have considered it usually rely on PoW to cold-start their mining ecosystem anyway. At the outset, compute power commitment still remains the only reasonably fairly distributed resource. (iii) PoS has decentralization concerns (the rich obtain a richer outcome) because it helps large coin holders to earn even more coins in revenue. As the research and applications of PoS mature, we indeed expect this to be an interesting direction of future research.

[10] $h(m) = \lambda e^{-\lambda m}$ is one potential distribution.

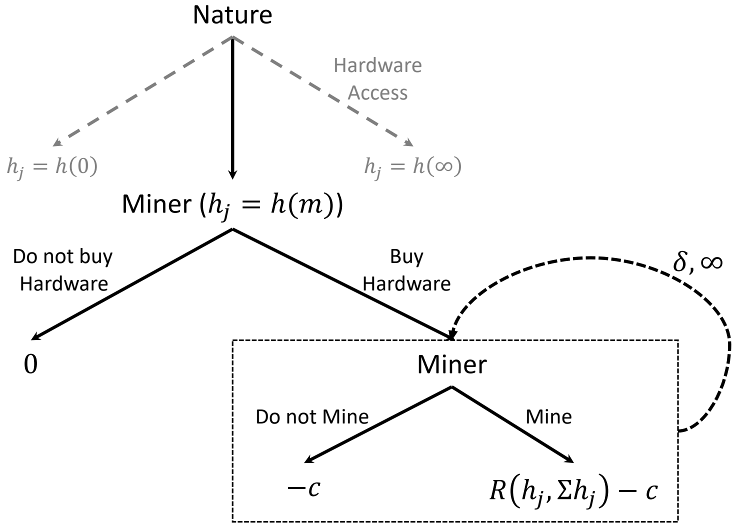


Figure 9 *Miner j first realizes access to hardware computing power h_j such that the miner of rank m has access to the computing power of $h(m)$. Next, the miner decides whether to buy the hardware for mining. If they buy, it is only rational to mine every block period thereafter. The period revenue is based on one's own power relative to everyone else's power. The same payoff is repeated infinitely with a discount factor δ.*

finding the Bitcoin mining puzzle solution faster than other miners. The expected revenue earned by a miner is given by block revenue $(R+B)$ multiplied by the miner's probability of creating any given block. This probability is a ratio of their own computing power $h(m)$ relative to the network's overall power. As more miners enter the network, each miner's puzzle-winning probability decreases, as they all share the same pie. We model miner entry as a zero-barrier event, i.e., miners enter until the marginal miner makes zero profits. Miners with efficient hardware (high $h(m)/c(m)$) will naturally crowd out inefficient hardware because of their competitive advantage. The marginal miner is the one who earns just enough expected revenue to equal his or her cost. Let m_0^* represent the rank of marginal miner and H_0^* represent the total mining power in the network. The network's overall power is the sum of power for all miners ranked between 0 and m_0^*, i.e., $\int_0^{m_0^*} h(m)dm = H(m_0^*)$.

We can uniquely identify the marginal miner m_0^*. The uniqueness is guaranteed since $R+B > c(0)$ and $\Lambda(m_0) = \frac{h(m)}{H(m)c(m)}$ is monotonically decreasing from 1 to 0 on its domain $m \in [0, \infty)$[11].

$$(R+B) \times \frac{h(m_0^*)}{H(m_0^*)} = c(m_0^*) \ ; \quad \text{where} \quad H(m_0^*) = \int_0^{m_0^*} h(m)dm \tag{10}$$

This simplifies to

$$m_0^* = \Lambda^{-1}\left(\frac{1}{R+B}\right) \tag{11}$$

[11] $\frac{h(m)}{c(m)}$ is monotonically decreasing, as discussed above. Additionally, $H(m)$ is a cumulative sum, thus increasing in m.

Table 2 *A summary of the main notation used in the analysis throughout the paper.*

Notation	Description
v	Value of a payment.
V_{max}	Maximum payment value in the market.
V_{secure}	Maximum payment value secure against double-spend attack.
V	Maximum value vying for payment via blockchain $V = \min\{V_{max}, V_{secure}\}$.
N_{max}	Total number of payments distributed in $[0, V_{max}]$ every period.
N	Total number of payments distributed in $[0, V]$ every period.
f	Transaction fee offered by a user.
ρ	Proportional transaction fee charged by a bank.
δ	Discount factor delay in verification of a transaction.
c	Per period amortized cost of mining.
n_F	Fixed capacity (block size) of Bitcoin, decided by its designer.
n_P	Size of a partially filled block.
γ	Ratio of Bitcoin capacity to its demand.
α	Proportion of computing power relative to the total mining network.
α_l	Minimum power for a single miner to profit from partial blocks.
$\hat{\alpha}$	Minimum power for an individual miner in a colluding group.
α_m	Power of smallest colluding miner.
α_s	Power of smallest free riding miner.

We consider the nonzero block reward in Appendix E. Earlier in this section, we derived the equilibrium expressions for transaction fee revenue R as follows,

$$R_0^* = \rho\alpha_h\gamma(1-\gamma)NV \tag{12}$$

Proposition 2 *The number of miners and the corresponding network computing power increase with increasing mining revenues. The mining power collapses at zero revenue*

- *As capacity n_F increases, revenue from the transaction fee first increases then decreases.*
- *As duration d increases, revenue from the transaction fee first increases then decreases.*
- *As hardware running cost c (USD/block) increases, the number of miners and network computing power decreases.*

3.2. Strategic Mining

We have shown that user cost savings are maximized when $n_F = N$. However, this raises zero revenues for miners. In this section, we consider the case whereby the blockchain designer has set capacity $n_F \in [\frac{N}{2}, N]$

above the revenue-maximizing value, i.e., $n_F = \frac{N}{2}$. We investigate if an individual strategic miner, with α proportion of the total network hash power, has a profitable strategy to extract excess rents, specifically by deviating to create partially filled blocks n_P. These partial blocks include only the top-n_P $(< n_F)$ transactions. Such a strategy could potentially create an artificial capacity constraint. A smaller capacity means that individual users need to beat greater competition to have their transaction included. It is initially unclear if the increased fee competition among users will make up for fewer transactions being included on the block. We will show that a strategic miner with at least α_l (defined in (4)) proportion of the total computing power is able to extract excess revenues via partial block filling. A miner smaller than α_l creates blocks too infrequently and, therefore, is not able to create sufficient congestion to make up for their sacrifice of block capacity.

The game tree provides the sequence of choices exercised by users and miners as well as payoffs in every block creation period. First, the user with payment v chooses to either make their payment off chain or to make a fee bid $f(v)$ on chain. Next, nature determines if the block is added by a passive miner with probability $(1-\alpha)$ or a strategic miner with probability α. If it is a strategic miner, the miner picks an optimal partial fill level n_P from a choice of levels in $[0, n_F]$. There are three critical points to note in this game description: (i) The strategic miner is forward-looking. They choose n_P to maximize lifetime discounted revenue over infinite repetitions of this stage game. (iii) Users are myopic about maximizing utility from current payment[12], and they have rational expectations about competing bids. (ii) Users know the identity of the strategic miner; they can see the history of partial blocks created by the strategic miner, and they have rational expectations of the strategic miner's actions.

Users now face uncertainty in whether the next block will be mined by a strategic or passive miner. They have three choices: (1) offer zero fee and take the off-chain option, (2) offer a low fee f_l that beats only $N - n_F$ other users hoping for a full block (n_F), or (3) offer a high fee f_h that beats $N - n_P$ other users to be included irrespective of a full (n_F) or partial (n_P) block. The users with the highest values stand to lose the most if excluded on the chain. High-value users $(v \geq v_h)$ unwilling to take this risk pay higher fees f_h. They are included in the immediate block, deriving a payoff $\delta(v - f_h)$. A middle tier of users $(v \in [v_l, v_h])$ will prefer to take the risk rather than pay a higher fee. A lower fee offer of f_l derives a payoff of $\delta(v - f_l)$ with probability $(1-\alpha)$ or a fallback off-chain payoff of $\delta v(1-\rho)$ with probability α. The lowest tier $(v \in [0, v_l])$ will continue to be left out of the chain as before. The three segments of users are illustrated in Figure 11. Thus, we are searching for an equilibrium bidding function for users $f^*(v)$ in the following parametric strategy space,

$$f(v) = \begin{cases} 0 & \text{if } v \in [0, v_l] \\ f_l & \text{if } v \in (v_l, v_h] \\ f_h & \text{if } v \in (v_h, V] \end{cases} \quad ; \quad v_l \leq v_h \leq V \tag{13}$$

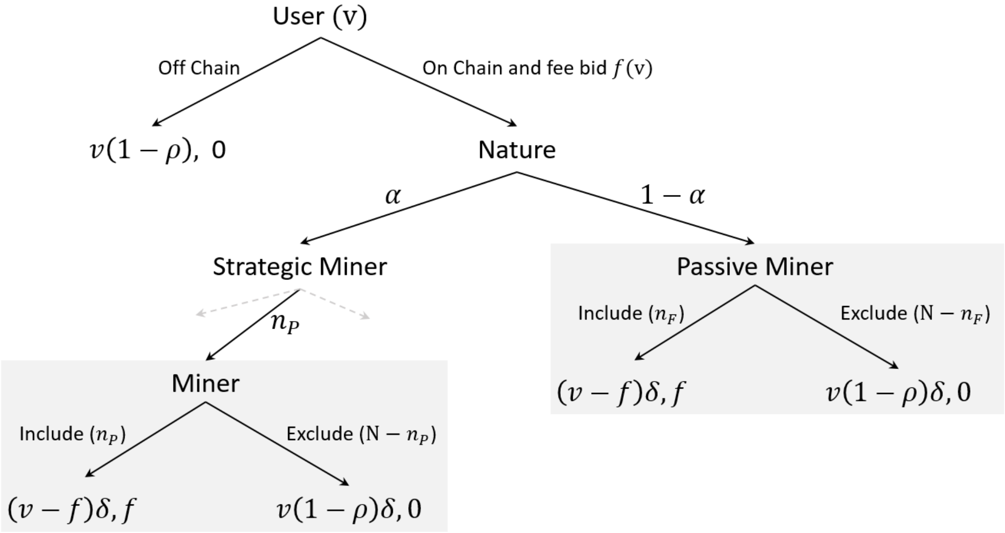


Figure 10 *User first draws a payment need v, then decides off chain or on chain bid. The next Block is added by either a strategic miner (probability α) or a passive miner (probability $1-\alpha$).Unlike the passive miner, the strategic miner has an additional decision to decide optimal partial fill level n_P.*

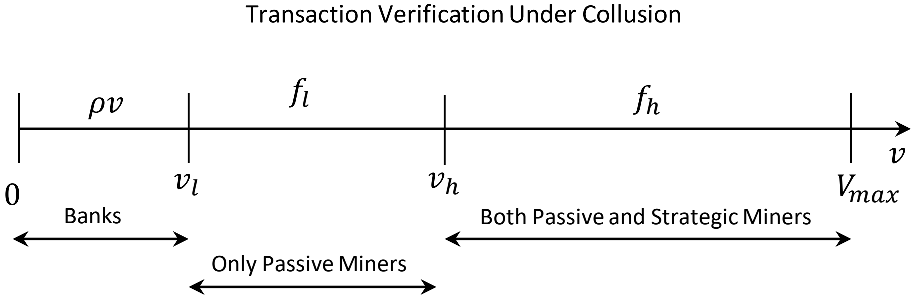

Transaction Verification Under Collusion
ρv
f_l
f_h
v
0
v_l
v_h
V_{max}
Banks
Both Passive and Strategic Miners
Only Passive Miners

Figure 11 *Three segments of users and their respective off-chain and on-chain fees.*

We must find an equilibrium bidding function for users $(f_l^*, f_h^*, v_l^*, v_h^*)$ and an equilibrium block fill level response for the strategic miner (n_P^*) such that neither the users nor the miner have profitable deviations from these strategies. This requires

(a) Low-value users $v \in [0, v_l]$ prefer off-chain bidding. They do not find it profitable to deviate to an on-chain bid $f > f_l$.

(b) Medium value users $v \in (v_l, v_h]$ prefer on-chain bid $f(v) = f_l$. They do not find it profitable to deviate to (1) going off chain, (2) bidding $f_h > f(v) > f_l$ or (3) bidding $f(v) \geq f_h$ on chain.

(c) High-value users $v \in (v_h, V]$ prefer on-chain bid $f(v) = f_h$. They do not find it profitable to deviate to (1) going off chain, (2) bidding $f(v) < f_h$ or (3) bidding $f(v) > f_h$.

(d) The strategic miner, who moves second, must not deviate from their equilibrium partial fill level n_P^* when faced with any off-equilibrium bid vector $f(v) \neq f^*(v)$. The strategic miner stands to earn R_{Dev}

[12] We relax this assumption in Appendix B.

by deviating to a partial fill level $n_P \neq n_P^*$ in the current period. This may be preferable to not deviating R_{noDev}. However, the subsequent lifetime payoff should be sufficient to compensate the strategic miner for this single-period loss ($R_{Dev} - R_{noDev}$). This ensures that partial fill commitment is credible even though the miner plays second in the stage game.

We know that the value of v_h (resp., v_l) should be such that there are exactly n_P (resp., n_F) transactions that have transaction values above v_h (resp., v_l). Thus, we have

$$v_h = V\left(1 - \frac{n_P}{N}\right), \tag{14}$$

$$v_l = V\left(1 - \frac{n_F}{N}\right). \tag{15}$$

The conditions (a), (b.1) and (c.1) are satisfied if the user at $v = v_l$ is indifferent between using on-chain low fee f_l and off chain. Consequently, all lower value users prefer off chain and all higher value users will prefer on chain at fees $f \geq f_l$. Thus, we have

$$(1-\alpha)\delta(v_l - f_l) + \alpha\delta v_l(1-\rho) \;=\; v_l(1-\rho).$$

From the above equation, we obtain

$$f_l = \frac{v_l(\delta + \rho - 1 - \alpha\delta\rho)}{(1-\alpha)\delta}. \tag{16}$$

Substituting the value of v_l in (16), we obtain

$$f_l = (1-\gamma)(\alpha_h - \alpha)(1-\alpha)\rho V \tag{17}$$

The conditions (b.2), (b.3) and (c.2) are satisfied if user at $v = v_h$ is indifferent between on-chain low fee f_l and on-chain high fee f_h. Thus, we have

$$\delta(v_h - f_h) \;=\; \delta(1-\alpha)(v_h - f_l) + \delta\alpha v_h(1-\rho).$$

From the above equation, we have

$$f_h = \alpha\rho v_h + (1-\alpha) f_l. \tag{18}$$

Using (17) and (18), we have

$$f_h = (\frac{n_F - n_P}{N}\alpha + \frac{N - n_F}{N}\alpha_h)\rho V \tag{19}$$

The condition (c.3) is trivially satisfied because user payment receives the exact same treatment whether a user's fee offer ranks in top n_P or top 1. There is no incentive to offer a fee higher than f_h.

The revenue for strategic miner $(f_h \times n_P)$ is maximized as follows,

$$n_P^* = \begin{cases} \frac{N}{2}((1-\gamma)\alpha_h + \gamma\alpha), & if \quad \alpha \leq \alpha_h \\ \frac{N}{2}, & else \end{cases} \tag{20}$$

We know that the strategic miner cannot choose a value of n_P higher than n_F, i.e., $n_P^* \leq n_F$. Using the expression of n_P^* from (20), this condition simplifies to

$$\alpha \geq \alpha_l^{(1)} \text{ ; where } \alpha_l^{(1)} = \alpha_h \frac{1-\gamma}{\gamma} \tag{21}$$

This condition implies that for a miner with low hash power $(\alpha < \alpha_l^{(1)})$, the partial block filling strategy is always dominated by a full block filling strategy. A small (less than $\alpha_l^{(1)}$) strategic miner adds blocks too infrequently to seriously threaten users to increase their transaction fee offers. If alpha is low at 0.05, there is only a 5% chance that any given block is partially filled. This creates too little risk for users to compete for top-n_P rank. An extremely powerful strategic miner $(\alpha \geq \alpha_h)$ can force users to either offer a high fee $(f_h = \alpha_h \rho V/2,\ f_l = 0)$ or not take a risk with the on-chain option at all. A miner who wants to add a full block n_F does not find any awaiting transactions beyond n_P paying a lower fee f_l. In other words, this miner can take complete control of Bitcoin and act like a revenue-maximizing monopolist by keeping the effective capacity at its optimal value, i.e., $n_P = \frac{N}{2}$. This partial filling strategy is irrelevant if the capacity on the chain is already constrained (say, $n_F \leq \frac{N}{2}$).

Consider a hypothetical situation whereby demand is 25% over capacity (i.e., $\frac{N}{n_F} = 1.25;\ \gamma = \frac{4}{5}$). A miner with power greater than 90.9% ($\alpha_h = \frac{\delta+\rho-1}{\delta\rho} = 0.909$ for $\delta = 0.99; \rho = 10\%$) shrinks the artificial capacity to the revenue-maximizing level $n_P = \frac{N}{2}$ by filling 62.5% (n_P/n_F) of their block. A miner with as little as 22.7% of the total computing power $(\alpha_l = \alpha_h \frac{1-\gamma}{\gamma})$ of the network fills 78% $(n_P/n_F = \frac{1+\beta}{2\gamma})$ of their block and extracts undue revenues. A miner with less than 22.7% of the total computing power is unable to unilaterally engage in a partial block filling strategy. Assume that the blockchain designer attempts to increase the block capacity to serve greater demand, say, $n_F = 0.9N$. The same miner $(\alpha = 0.227)$ performs partial filling. These partial blocks are only 59% full in the second setting compared to 78% in the first setting. Furthermore, at $n_F = 0.9N$, a miner with as little as 9.09% of the total network power strategically performs underfilling. They fill 60% $(n_P/n_F = \frac{1+\beta}{2\gamma})$ of their block. Therefore, an increase in block capacity allows an increasingly smaller miner to perform strategic underfilling. Figure 12 shows three scenarios for a large miner with power $\alpha = \alpha_h,\ \alpha_h/2,\ \alpha_h/4$. The largest miner is able to keep the effective block capacity at $N/2$. The effective capacity is given by $n_P * \alpha + n_F * (1-\alpha)$ when the individual miners have less than α_h power.

Finally, condition (d) is satisfied if not deviating in response to an off-equilibrium bid vector $f(v)$ dominates deviation. If the miner does not deviate, they earn R_{noDev} now and subsequently they earn partial fill equilibrium revenues R_P^* in all following periods where they add a block. The expected reward t periods ahead is given by probability (α) of winning the mining puzzle multiplied by the discounted reward value $(\alpha \times \delta_m^t R_P^*)$. The expected lifetime rewards correspond to summation $(\alpha\delta_m^1 R_P^* + \alpha\delta_m^2 R_P^* + ..)$. Similarly, if the miner deviates, they earn R_{Dev} now and subsequently earn complete fill equilibrium revenues R_0^* in all following periods where they add a block.

$$\underbrace{R_{noDev}}_{\text{No Deviation Revenue}} + \underbrace{\alpha\frac{\delta_m}{1-\delta_m}R_P^*}_{\text{Partial Fill Revenue}} \geq \underbrace{R_{Dev}}_{\text{Deviation Revenue}} + \underbrace{\alpha\frac{\delta_m}{1-\delta_m}R_0^*}_{\text{Complete Fill Revenue}} \tag{22}$$

This simplifies to

$$\alpha \geq \frac{1-\delta_m}{\delta_m}\frac{R_{Dev}-R_{noDev}}{R_P^*-R_0^*} \tag{23}$$

We can upper bound the right-hand side by considering the most enticing fee bid vector $f(v)$ that maximizes the single-period deviation payoff $R_{Dev} - R_{noDev}$. This most enticing fee bid vector corresponds to all mid-tier users $(v \in (v_l, v_h])$ offering highest possible fees such that they gain no on-chain savings compared to off chain, i.e., $f(v) = \bar{f} = \alpha_h \rho v \ \forall \ v \in (v_l, v_h]$. For this bid vector, the strategic miner stands to gain by deviating to a full block n_F composed of payments $(v_l, V]$ instead of n_P only composed of payments $(v_h, V]$,

$$R_{Dev} - R_{noDev} \leq \int_{v_l}^{v_h}(\alpha_h\rho v)\frac{N}{V}dv = \frac{\alpha_h\rho N}{2V}(v_h^2 - v_l^2)$$

Thus, we have

$$\alpha \geq \alpha_l^{(2)} \text{ ; where } \alpha_l^{(2)} = \frac{1-\delta_m}{\delta_m} \times \frac{\alpha_h\rho N}{2V} \times \frac{v_h^2 - v_l^2}{R_P^* - R_0^*} \tag{24}$$

The strategic miner's partial filling threat is credible if the miner adds blocks frequently. This is true either if the miner is large enough or sufficiently forward-looking (high δ_m). This miner will not be enticed away from partial filling because her or she is willing to sacrifice payoffs in the immediate block creation game for near-future payoff in subsequent block creation periods. Thus, $\alpha_l^{(1)}$ and $\alpha_l^{(2)}$ make up two constraints on the smallest size of the strategic miner[13].

Proposition 3 *A strategic miner with proportion of the total mining network's computing power α is able to conduct a partial block filling strategy and extract undue revenues from Bitcoin transaction fees if $\alpha \geq \alpha_l$, where*

$$\alpha_l = max(\alpha_h \times \frac{1-\gamma}{\gamma}, \quad \frac{1-\delta_m}{\delta_m} \times \frac{\alpha_h\rho N}{2V} \times \frac{v_h^2 - v_l^2}{R_P^* - R_0^*}). \tag{25}$$

[13] The first constraint $\alpha_l^{(1)}$ turns out to be the binding constraint, while the second constraint $\alpha_l^{(2)}$ is trivially satisfied in practical parameter ranges. The second constraint plays a more critical role in the next section.

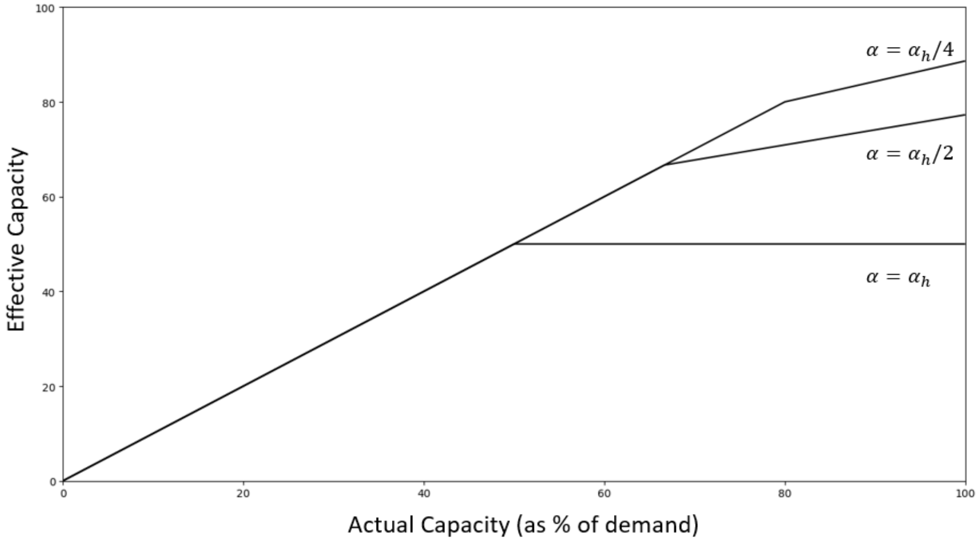


Figure 12 *A very larger miner ($\alpha = \alpha_h = 0.909$) does not let the effective capacity increase beyond 50% of demand. A small miner ($\alpha = \alpha_h/4 = 0.227$) performs partial filling when the actual capacity is raised beyond 80%. They do not let the effective capacity increase beyond 86% of demand. ($\delta = 0.99$, $\rho = 10\%$)*

- *As the strategic miner's power increases, the optimal partial fill level, i.e., n^*_P, decreases.*
- *As block size n_F increases (γ increases), strategic partial filling can be performed by an increasingly smaller miner.*
- *As block duration d increases (γ decreases but α_h decreases), strategic partial filling can only be performed by an increasingly larger miner.*

In practice, the accumulation of a high amount of computing power ($\alpha > 0.5$) may not be realistic for a single miner. This can be accomplished by a mining pool (e.g., AntPool). Although a mining pool allows for some coordination of mining resources, they do not have complete control. Individual participants can hop on and hop off at any time. In the next section, we discuss the possibility of small miners tacitly colluding to achieve computation power larger than α_l as a group.

3.3. Collusion

Partial block filling by a large strategic miner (with at least α_l fraction of the total computing power) achieves extra mining revenues. A group of small miners comprising the same computation power ($\Sigma\alpha_j = \alpha_l$) should similarly be able to achieve the same extra revenues. This requires miners to agree on the mutually beneficial partial block filling strategy. Such coordination is complicated by the low barriers to entry of new miners. In this section, we identify the conditions under which a tacit collusion is stable.

There is one key difference between the incentives of a single large miner and a group. The large strategic miner sacrifices nonzero fees. This creates artificial congestion and increased fee offers. The subsequent benefit of higher fee offers via this congestion is shared by the large miner as well as all other miners. The remaining miners $(1-\alpha_l)$ thus free-ride on congestion created by the large miner. Full blocks added by free riders gather larger fee revenues (R_F) than partial blocks (R_P) added by the large miner (equation 26). A group of small miners is not incentivized to make the same sacrifice. Every individual miner prefers that others perform the partial block filling while they free-ride on the resulting congestion. The mutually beneficial Nash equilibrium is therefore unstable in a single-block creation game, i.e., all miners want a free ride.

$$R_P = f_h n_P^*, \tag{26}$$

$$R_F = f_h n_P^* + f_l(n_F - n_P^*). \tag{27}$$

Bitcoin's block creation is repeated infinitely among the same set of miners. The infinite repetitions allow miners to sustain a tacit collusion. They can mutually commit to partial block filling under the threat of future punishment. A single deviation by an individual miner would result in no collusion for T block periods in order to wipe out single-period deviation profits. In a single-stage game, an individual miner has a profitable deviation. However, the deviation is no longer profitable when accounting for all future payoffs. We therefore focus on subgame perfect equilibrium (SPE) miner strategies in an infinitely repeated block creation game.

We obtain the following lower bound on the smallest colluding miner; the proof of this result can be found in the Appendix. We also discuss verifiability of a miner's action, which is a key requirement for tacit collusion. A small miner $(\alpha_j < \hat{\alpha})$ mines blocks infrequently, say, once a month or year. They are less threatened by punishment far off in the future. This can also be seen as a small miner's desire to make the most out of winning a mining puzzle once in a long while. A large miner sticks to the collusion strategy, as they expect frequent or near-term fee revenues. This difference in commitment of a small versus large miner arises because the decision to follow the collusion with partial blocks or deviate with full blocks is only revealed if the miner wins the puzzle race. In typical Cournot settings, the quantity produced by both small and large firms are observed in all periods.

$$\alpha_j \geq \hat{\alpha}, \quad \text{where} \quad \hat{\alpha} = \frac{1-\delta_m}{\delta_m(1-\delta_m^T)} \times \frac{\alpha_h \rho N}{2V} \times \frac{v_h^2 - v_l^2}{R_P - R_0}. \tag{28}$$

The constraints above $(\alpha_j \geq \hat{\alpha},\ \Sigma\alpha_j \geq \alpha_l)$ assure a colluding miner's commitment to the collusion. We could use this to check if an exogenously given distribution of mining powers (α_j) conforms to collusion requirements. However, miner entry is endogenously determined by available revenues and hardware distribution. Next, we identify corresponding constraints on mining hardware distribution for collusion. This is useful because a blockchain designer does not directly control mining powers (α_j), but they control hardware distribution via choice of the mining puzzle (e.g., SHA 256 Hash, Scrypt). It would allow us to discuss

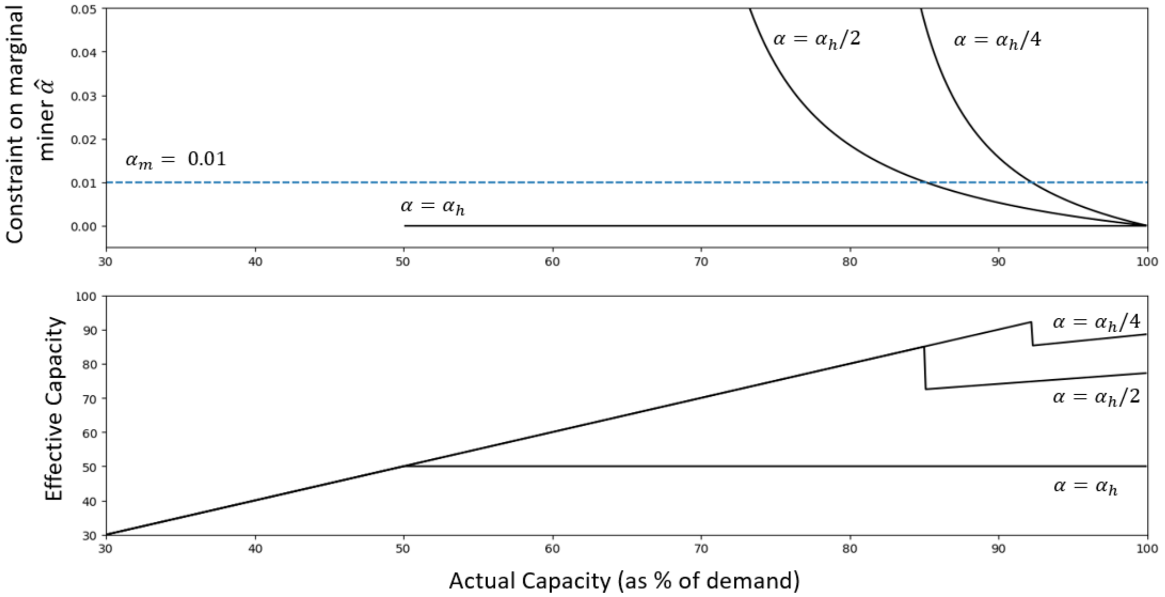


Figure 13 *A very large group ($\alpha = \alpha_h$) does not let the effective capacity increase beyond 50% of demand. They do not need to satisfy any $\hat{\alpha}$ constraint. A small group ($\alpha = \alpha_h/4$) does not let the effective capacity increase beyond 86% of demand. They only perform partial filling when the actual capacity is increased beyond 90%. The group is slightly less effective than a single miner because they also need to satisfy the $\hat{\alpha}$ constraint. ($\delta = 0.99$, $\rho = 10\%$)*

potential blockchain designs for averting collusion. We obtain a constraint on mining cost c and hardware distribution Λ. The proof can also be found in the Appendix; the proposition below captures the conclusion.

Proposition 4 *A group of miners with α_l proportion of the total mining power can sustain collusion if the hardware distribution (c, Λ) satisfies,*

$$c\left(\Lambda^{-1}(\frac{1-\delta_m}{\alpha_l R_P})\right) \geq \hat{\alpha}\frac{R_P}{1-\delta_m} \tag{29}$$

- *Under this collusion, all miners with greater than α_m proportion of the total mining power collude.*
- *The total number of miners s^* exceeds the total number of miners who joined under the no-collusion strategy m_0^*.* [14]

Note that the miner collusion described above has similarities with collusion among firms in Cournot markets with repeated interactions. It is worth highlighting two key differences: (i) In the standard Cournot setting, all firms decide supply quantities simultaneously. In our setting, the firm's quantity decisions (n_P) are instead separated in time. The standard Cournot collusion places a bound on the discount factor. Our

[14] Since $R_P > R_0$ and Λ^{-1} is monotonically decreasing, $s^* > m^* > m_0^*$.

setting places a bound on the colluding firm size because the firm's effective discount factor depends on size. This happens in our context because firms have the opportunity to set quantity for the entire market at a frequency proportional to their size. (ii) In the standard Cournot setting, all users irrespective of private valuations observe the same quantity and price in the market. In our setting, the user bids first with rational expectations of markets' output quantity levels. While users have the ability to affect the firms' collusive behavior being the first mover, it also leads to user discrimination. Users with different valuations bid differently for a homogeneous product. While colluding firms must only supply to high valuation users, noncolluding firms can supply to both high and low valuation users. A similar collusion equilibrium may be applicable in markets beyond p2p payment blockchains, with unique characteristics described above.

4. Security

The collusion equilibria discussed in the previous section results in a smaller effective block capacity, greater fees and larger revenues for the mining network. The Bitcoin community wants to eliminate this collusion to serve the entire user demand at low fees. A simple intervention could be to enforce block filling. Such an intervention is easily defeated, as the colluding miners could simply fill their remaining blocks with artificial transactions. They could add zero-fee-paying transactions sending coins back and forth between their own accounts. Such artificial transactions would achieve the same objective as partial block filling by keeping out actual user demand. A more realistic solution could be to effect a change in the hardware distribution.

The existence of large mining hardware ($> \hat{\alpha}$) helps sustain the collusion. Bitcoin's mining puzzle requires miners to perform SHA256 hash calculations. These calculations are computationally intensive, i.e., requiring high calculation speeds but minimal memory. Large miners deploy ASIC machines that are orders of magnitude faster than GPUs and CPUs at SHA256 hash calculations. A few blockchains – alternatives to Bitcoin – implement ASIC-resistant mining puzzles. ASIC-resistant mining puzzles are memory intensive, i.e., requiring substantial storage memory. ASIC machines deployed by large miners are efficient at computations but inefficient at memory-intensive operations. As ASICs become uncompetitive, miners with GPUs and CPUs are able to enter the mining network. Since GPUs and CPUs are more widely accessible than ASICs, this change has the potential of making miners more equitable. In the context of our hardware distribution model, this effectively means that mining is now performed by large numbers of miners all in the tail of the distribution.

The hardware distribution functions $h(m),\ c(m)$ and $\Lambda(m)$ are now updated to $h'(m), c'(m)$ and $\Lambda'(m)$, respectively. A colluding group could still form, thereby adding up to the same (α_l) proportion of the total computing power. The proportion of the total computing power for the marginal miner must still remain greater than $\hat{\alpha}$. This is satisfied if

$$c'(m^*) \geq \frac{R_P}{1 - \delta_m} \times \hat{\alpha} \tag{30}$$

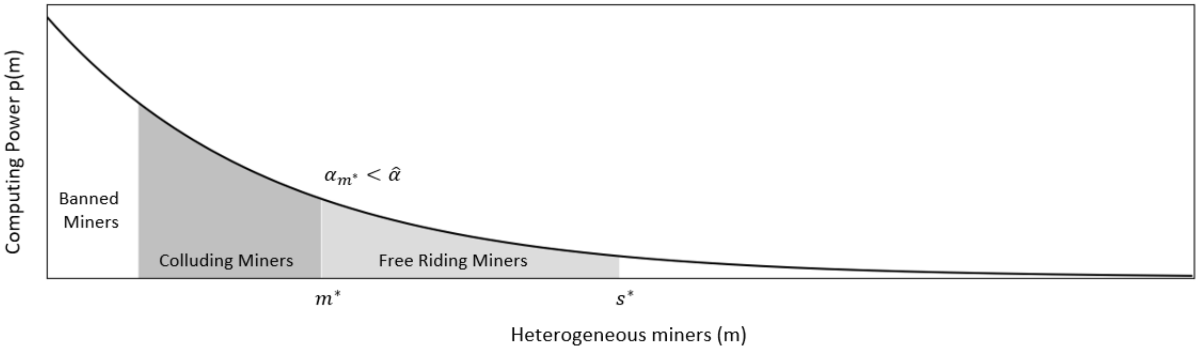


Figure 14 *The x-axis represents miners ranked by access to hardware. The ASICs on the left are blocked by an intervention. A colluding group with α_l power would need to include a larger number of small miners. The new marginal miner (m^*) may not be willing to sacrifice block capacity under tacit collusion.*

The updated cost function $c'(m)$, similar to $c(m)$, is a decreasing function of m.

$$c'(m) = \begin{cases} \infty & \text{if } m \text{ in banned region} \\ c(m) & \text{otherwise} \end{cases}$$

Prior to the ban, the marginal miner's cost was large enough to sustain collusion. As high-quality mining hardware is banned, the new marginal miner moves toward the right (see Figure 14). The cost of an infinitely small hardware (say, hand-held mobile devices) is near zero $c'(\infty) = 0$. There exists a unique level of hardware ban beyond which the inequality above is shattered. In summary, banning large miners invites smaller hardware that can be operated at lower cost. A lower cost attracts a more equitable community of miners – a larger number of miners each expending a smaller computing cost. The new marginal miner may be too small $(> \hat{\alpha})$. The marginal miner may not be willing to sacrifice block capacity under tacit collusion. At first glance, this appears to be an effective method to eliminate collusion and recover the full block capacity. We will show that the intervention (ASIC resistance or otherwise) to eliminate collusion threatens the security of user payments by inviting hacks.

4.1. Double-Spend Attack

A double-spend attack poses a threat to the validity of a payment. Note that the Bitcoin ledger is a unique longest chain of blocks. As discussed earlier, when multiple chains emerge, the chain that attracts the majority of the mining power emerges as the longest and thus the accepted chain. Once the longest chain emerges, all other parallel chains are abandoned. The transactions that appear in these shorter parallel chains but not in the longest chain are deemed invalid. An adversary with computing power $(\alpha_j = \theta)$ can create a parallel chain. In addition, if the adversary has majority power $(\theta > 0.5)$, his or her chain will always be the longest chain. Such an adversary essentially controls the blockchain. This adversary could prevent new transactions from gaining confirmation, add invalid transactions, or even reverse transactions that were completed to double-spend their coins. This is well known as a 51% attack. If any miner were to achieve 51% of the aggregate computing power of the Bitcoin network, Bitcoin would lose all trust and value. As a result, it is not in the interest of any miner to accrue that much power.

A smaller adversary $(\theta < 0.5)$ cannot arbitrarily add invalid transactions onto their block. Their block would be rejected in favor of a valid block. However, they can spend their account balance twice by creating parallel chains (Figure 15). The attack involves an adversary and a merchant. The adversary orders some service or product from a merchant and pays the merchant in Bitcoin. The adversary wants to receive the product but does not want to pay. The merchant would deliver the product only after they see the payment recorded on the single longest chain. In a double-spend attack, the adversary records payment on the longest chain, the merchant delivers the product; then, the adversary is able to respend the same payment on a different fork, which later emerges as the longest chain. The attack would be successful if the adversary could convince a majority of the mining community to switch away from the longest chain where they initially make payment to the merchant. The original chain (and the payment to the merchant on it) ends up being abandoned in favor of a new longest chain. The merchant has effectively delivered a valuable product for a payment that never really occurred.

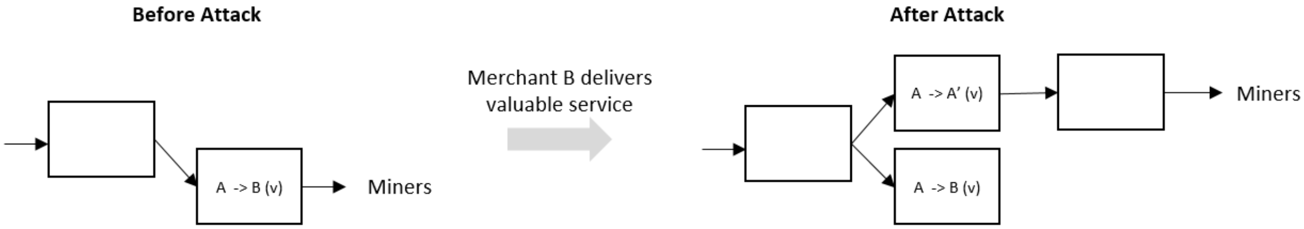


Figure 15 *Adversary (A) sends a payment v to the Merchant (B). The merchant delivers a valuable service once this transaction is added onto a Bitcoin block. The adversary then manages to resend the same value v to another account (A'). The second transaction is recorded on the eventual longest chain.*

We describe a simple double-spend attack strategy employed by an adversary. The attack is launched when the adversary has just mined a new block. The adversary decides to keep this block private. Honest miners are attempting to add a new block in parallel, which will eventually create a parallel chain. We denote the adversary chain by a and the honest chain by h. State 0 represents the launch of the attack when the height of a is 1 and that of h is 0. The adversary's goal is to mine a private chain a with more blocks than the honest chain h. If the adversary releases their private chain, all miners will accept this adversary chain as the longest unique chain. The shorter honest chain would be abandoned. This is a risky strategy, with power $\theta < 0.5$, the adversary is likely to be left behind. If the adversary becomes left behind, they will lose the block fees earned on their abandoned blocks. If the small probability that their chain wins is realized, they want to make the most out of it by duping a merchant.

The adversary adds a payment of value v to one of their own accounts. They do not broadcast this spend publicly; instead, they add this to their private block. The adversary also broadcasts a payment of value v. This payment is meant to purchase valuable goods or services from a merchant. As $h = 0$, the adversary's payment has not been picked up on the publicly visible chain yet. The adversary must continue mining privately since the merchant does not deliver the valuable service before they see the payment added to the longest publicly visible chain.

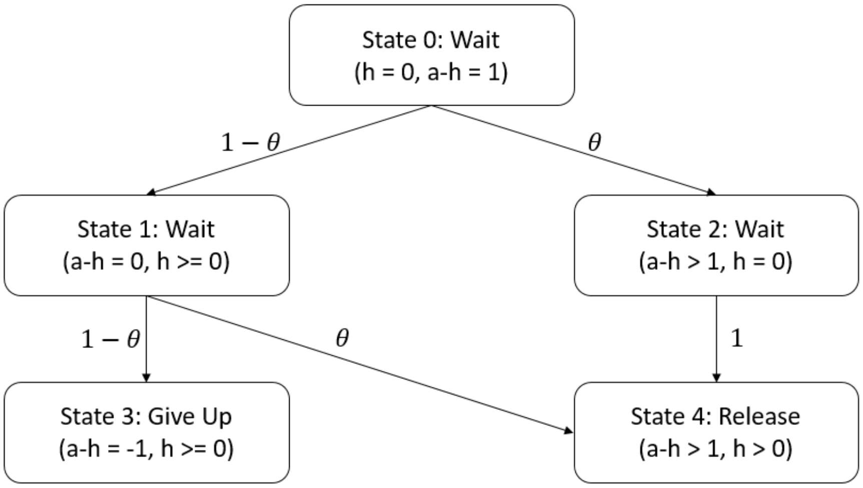


Figure 16 *Adversary double-spend Markov decision process (MDP). Each node represents a state and the corresponding action for the adversary. State 0 is the initial attack launch state, and States 3 and 4 are terminal states. The edges represent the state transitions and the corresponding probabilities.*

The next block is added to the honest chain with probability $1-\theta$ landing in State 1. In this state, the honest and adversary heights are equal. If the adversary adds the next block (probability θ to State 4), they succeed. If the honest miners add the next block (probability $1-\theta$ to State 3), the adversary gives up, and the attack fails.[15] If the next block from the initial State 0 is added by the adversary, we land in State 2. The two heights are updated to $h = 0$ and $a = 2$. The attack is not complete since the broadcasted transaction has not been added to the honest chain yet. However, the adversary is certain to succeed in his attack at this point. He or she will release the longer chain once a single block is added to the honest chain. All miners will accept this adversary chain as the longest unique chain. The transactions on the adversary's

[15] The adversary has other options: (1) release the private chain in State 1 and let the honest miners decide on a toss-up between two equal-height chains or (2) continue mining privately at State 3 in hopes of once again obtaining the advantage. These alternatives are preferable in certain parameter ranges but do not change the core result.

blocks remain *valid*. When the ledger entries on the unique chain of blocks are added up, no balance seems to be double spent once the honest chain is ignored.

It is important to reiterate that we have described a simple form of the double-spend attack MDP strategy. Following Eyal and Sirer 2018, the attacker could choose to keep mining privately in State 4 instead of ending the attack with one block advantage. This results in competing miners expending their power on a fork that eventually fails to become the single longest chain. A more comprehensive strategy would certainly improve the attacker's payoff. On the other hand, following Biais et al. 2018, the attack may have a lower chance of success if one or more large honest miners have received block revenues on the shorter fork in State 4 and thus have a "vested interest" in not moving to the attacker's longer fork. In fact, there are numerous variants of the attack strategies and equilibrium fork outcomes that would change the exact formulation of the attacker's payoff. We argue that the underlying intuition of an attacker's trade-off remains the same: (i) guaranteed honest participation revenues versus (ii) a costly attempt at creating an alternative profitable "state of the world" (fork) with significantly less than 51% consensus power. A similar trade-off is formulated by Chiu and Koeppl 2017 and Budish 2018.

In our simple MDP, the adversary fails in this attack with a probability of $(1-\theta)^2$. This corresponds to the probability of two consecutive blocks being added to the honest chain. As an example, an adversary with a 0.05 proportion of the total computing power has a success rate of $[1-(1-\theta)^2] = 9.75\%$. If the adversary succeeds, he or she obtains the double-spend value (v) in addition to the usual block fee. The adversary fails if honest miners add two blocks before the adversary is able to find a single block on top of their private chain. In this scenario, the adversary's private chain consists of a single block. This private block records a list of transactions that offer fees to the block creator (adversary). Unfortunately, this private chain is a short chain from the longest main honest chain. Therefore, none of these transactions are added up as the final balance of the adversary. The adversary forfeits the block revenue R on their private chain when they fail to perform their double-spend attack. A small adversary $(\theta << 0.5)$ fails more often than not. Equation 31 captures the profits for the adversary when they have just launched the attack. They earn $(v+R)$ on the first block if it is eventually included in the main chain with a probability of $[1-(1-\theta)^2]$. Subsequently, they earn the usual honest mining rewards in the future.

$$\pi_{\text{double spend}} = [1-(1-\theta)^2](v+R) - c + \delta_m(\theta R - c) + \delta_m^2(\theta R - c) + ... \tag{31}$$

As an alternative to this attack, an adversary could have simply engaged in the usual mining, i.e., not mining private blocks. The adversary would avoid losing the fee on their abandoned private block.

$$\pi_{\text{honest}} = (R-c) + \delta_m(\theta R - c) + \delta_m^2(\theta R - c) + ... \tag{32}$$

This attack is worthwhile if the likely loss of block fees on the adversary's abandoned private chain is compensated by the double-spend value v. Equation (33) represents the additional payoff in launching a double-spend attack.

$$\pi_{\text{double spend}} - \pi_{\text{honest}} = \Delta\pi = v[1-(1-\theta)^2] - R(1-\theta)^2, \tag{33}$$

To dissuade an adversary from performing a double-spend attack, we need $\Delta\pi \leq 0$. From (33), this condition simplifies to

$$v \leq V_{secure},$$

where

$$V_{secure} = R\frac{(1-\theta)^2}{1-(1-\theta)^2} \tag{34}$$

The above condition shows that only transactions with value $v \leq V_{secure}$ are safe from a double-spend attack with miner revenue R. If all users have knowledge of an adversary with power θ, merchants expect buyers to pay an extra security cost $S(v)$ to insure against an attack. We assume a binary security cost[16],

$$S(v) = \begin{cases} 0, & if \quad v \leq V_{secure}, \\ v & else \end{cases} \tag{35}$$

Thus, in equilibrium, users with transactions above V_{secure} will not consider the blockchain channel at all owing to the high security cost. Note that we had treated V_{secure} as an exogenous known security upper bound at the outset. We had modeled a payment upper bound V with a demand for N payments uniformly distributed in $[0, V]$. Now, to fully endogenize users' security consideration, we hope to derive the payment upper bound V with a demand for N payments uniformly distributed[17] in $[0, V]$.

$$N(V) = \int_0^V g(v)dv = \begin{cases} \frac{N_{max}}{V_{max}}V, & \text{if } V < V_{max} \\ N_{max}, & \text{else} \end{cases} \tag{36}$$

We use (8) to substitute miner revenue R generated from these payments, and we try to find conditions where all payments $[0, V]$ are secure, i.e., $V < V_{secure}$,

$$V \leq (1 - \frac{n_F}{N(V)})Vn_F \times \frac{1}{\eta(\theta)} \tag{37}$$

where

$$\eta(\theta) = \frac{1}{\rho\alpha_h}\frac{1-(1-\theta)^2}{(1-\theta)^2} \;; \quad \frac{\partial\eta(\theta)}{\partial V} \geq 0 \tag{38}$$

[16] We show a continuous security cost in Appendix D

[17] We also show an exponential payment density instead of uniform in Appendix D.

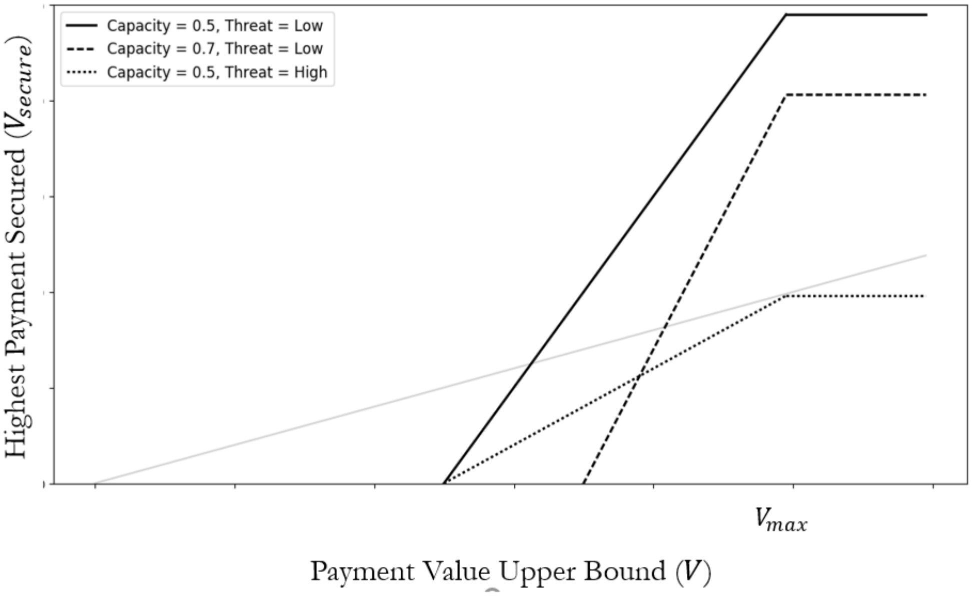


Figure 17 *The x-axis is payment upper bounds V. The y-axis is the maximum payment value V_{secure} secured by revenue generated from $N(V)$ payment in $[0,V]$. This is shown for alternative threat levels (θ) and block capacities (n_F). If the resulting V_{secure} is less than V (gray trend), the corresponding setting ($[0,V],\theta,n_F$) is infeasible.*

Equation (37) simplifies to,

$$\eta(\theta) \leq (1 - \frac{n_F}{N(V)})n_F \tag{39}$$

The RHS is increasing in payment upper bound V until V_{max} because the number of payments $(N(V))$ vying for execution via blockchain $[0,V]$ grow linearly with V. Thus, if there exists a $V \in [0, V_{max}]$ such that demand $[0,V]$ is secure, this implies that full demand $[0,V_{max}]$ must also be secure. On the other hand, if even full demand $[0,V_{max}]$ cannot raise sufficient miner revenue and is insecure, then no payment demand $[0,V]$ will be secure. Therefore, we replace upper bound V by the highest revenue-raising upper bound V_{max} and demand $N(V)$ by full demand N_{max}. We obtain

$$(n_F)^2 - (n_F)N_{max} + \eta(\theta)N_{max} \leq 0 \tag{40}$$

This is satisfied for

$$n_F \in [1-(n_F)_{max}, (n_F)_{max}] \quad ; \qquad (n_F)_{max} = \frac{N_{max}}{2} + \frac{N_{max}}{2}\sqrt{1 - \frac{4\eta(\theta)}{N_{max}}} \tag{41}$$

The miner revenue is highest at a block capacity n_F close to $N_{max}/2$. A feasible range of block capacity around $N_{max}/2$ raises sufficient revenues to secure the payment demand $[0,V_{max}]$. This range of feasible block capacities depends on adversary threat $\eta(\theta)$. If the adversary is extremely powerful (high θ,$\eta(\theta)$), the

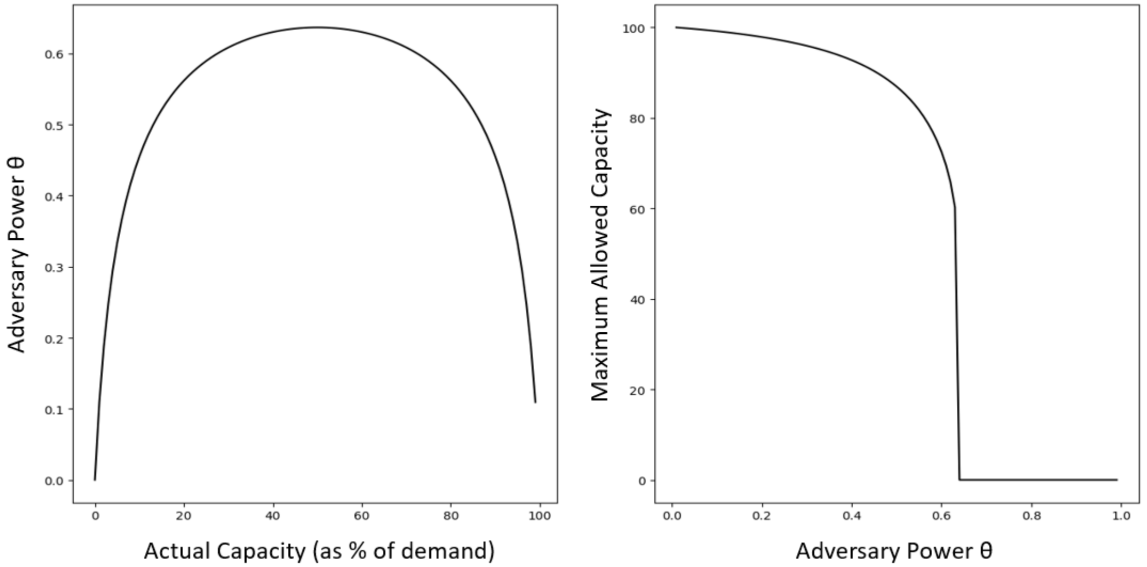


Figure 18 *(Left) The adversary power required to perform double-spend attacks decreases if the capacity is close to 100% or 0% of demand. (Right) The maximum allowed capacity n_F decreases with adversary power. Note that even a $\theta > 0.5$-adversary may not perform an attack if there are sufficient revenues to be acquired.*

feasible range shrinks and no real-valued n_F exists. This occurs when $\eta(\theta) < N_{max}/4$, which simplifies to $\theta > \theta_{UB}$, where

$$\theta_{UB} = 1 - \sqrt{\frac{4}{4 + \rho\alpha_h N_{max}}} \tag{42}$$

In this case, even the revenue-maximizing capacity ($n_F = N_{max}/2$) may not be sufficient to avert such a powerful adversary ($\theta > \theta_{UB}$). More generally, the existence of a smaller adversary (say, $\theta < \theta_{UB}$) places a bound on the maximum (and minimum) block capacity. The maximum possible capacity decreases in θ. Figure 18 shows the trade-off between block capacity and minimum adversary power required for a worthwhile double-spend attack.

Note that collusion among a large group of miners ($\Sigma\alpha_j > \alpha_h$) shrank the block capacity artificially to half the demand. We had discussed possible intervention in mining hardware distribution to avert such a collusion. However, given the security implication, such an intervention could make the blockchain open to the double-spend attack. Further, the intervention specifically required blocking out miners with disproportionately more efficient hardware (high $h(m), c(m)$), and these banned miners may even be the most likely double-spend adversary if they cannot earn honest mining revenues at all. Given the importance of security, we could in fact consider collusion as a positive characteristic. Security on the platform is an outcome of total mining power, which may often be below desired level. Collusion among miners raises additional revenues; a chunk of these revenues go toward adding more mining equipment by endogenous zero-barrier new miner entry.

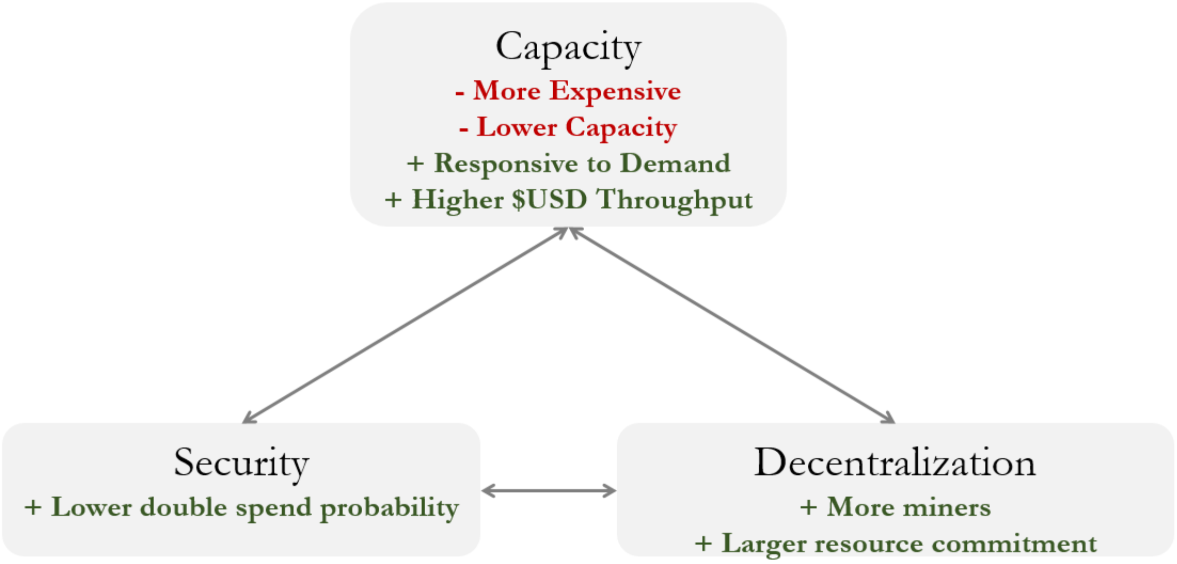


Figure 19 *A summary of collusion's impact on Capacity-Decentralization-Security trilemma. [Capacity] Collusion allows fewer payments at higher fees, but dynamically adjusts the throughput in response to demand. [Capacity → Decentralization] Collusive rent extraction partially goes to colluding miner, but also spills over into entry of more miners. [Capacity + Decentralization → Security] A higher cumulative commitment of resources mean that an attacker has smaller proportion of network's power (low θ) and higher incentive to gain honest revenues (higher collusion revenue R). Thus affording a higher level of Security. [Security → Capacity] Higher security means larger payments are facilitated (higher V_{secure}).*

Thus, economic rent extraction of collusion spills over into investment in greater security. Additionally, note that block capacity n_F to avert the adversary depends on demand level N_{max}; the latter can fluctuate in practice, while the former is relatively inflexible. Collusion acts as an endogenous mechanism for participants to make artificial adjustments (e.g., n_P) in response to small fluctuations in demand (N_{max}), payment spectrum ($[0, V_{max}]$) or off-chain fee rates (ρ).

Proposition 5 *If collusion is eliminated, the Bitcoin block size remains bounded by the threat of double-spend attacks:*

$$(n_F)_{max} = \frac{N_{max}}{2} + \frac{N_{max}}{2}\sqrt{1 - 4\frac{\eta(\theta)}{N_{max}}} \tag{43}$$

- *An adversary greater than θ_{UB} would perform an attack, even if the block size is set at the block-revenue-maximizing level of $\frac{N_{max}}{2}$.*

In July 2014, Bitcoin faced the first and perhaps only realistic risk of a 51% double-spend attack. A single mining pool, Ghash.IO, accumulated close to 51% of the total computing power. While mining pools are composed of a large number of contributing miners, a pool owner typically coordinates block addition. The pool's total power would have allowed Ghash.IO to easily launch a double-spend attack. Ghash.IO voluntarily promised to reduce its power below 40% to minimize the potential risk of double spend (ArsTechnica 2014). Ghash.IO acted rationally since they most likely expected to earn more from continued honest block revenues than a double-spend attack. The average transaction value on Bitcoin was $5000, while a single-block reward was worth $100,000 (bitinfocharts 2018). Ethereum is the 2nd-largest public blockchain by usage and market capitalization after Bitcoin. Ethereum has also not yet faced a double-spend attack. It is not surprising that miners preferred to continue earning the reward instead of launching a double-spend attack.

Bitcoin Gold is a fork of Bitcoin launched to eliminate large ASIC miners. Bitcoin Gold has a substantially smaller demand from users and is valued 100-times less than Bitcoin. In May 2018, Bitcoin Gold experienced double-spend attacks. A payment worth $17.5 Mn (Osborne 2018) was stolen at a time when Bitcoin Gold blocks paid less than $1000 per block (bitinfocharts 2018). Ethereum classic is a fork of Ethereum valued at 30-times less than Ether (Ethereum coin) and pays $20 per block (40 blocks every 10 minutes). In January 2019, Ethereum classic suffered a theft of $1.1 Mn from a double-spend attack (Casey 2019). Crypto51 (2019) estimates the cost to rent sufficient hashing power to match the current network hashing power for an hour. This cost is estimated at $239,263 for Bitcoin, $79,699 for Ethereum, $3,969 for Ethereum Classic and $737 for Bitcoin Gold. This cost is a reasonable proxy for comparing the cost of a double-spend attack. Both Ethereum Classic and Bitcoin Gold have low mining power due to a combination of smaller revenues and their ASIC-resistant mining puzzles.

5. Generalization Beyond Bitcoin

While we connect our model to Bitcoin, our findings are generalizable to a large variety of peer-to-peer payment blockchains (Table 3). These blockchains broadly follow a broad paradigm kickstarted by Nakamoto 2008 with the following design elements: (i) Limits on ledger addition size and frequency to allow sufficient time for dissemination (syncing) over the peer network ($n_F,\ d$). (ii) Rewards (transaction fee auction f or block reward R) for validating peer payments when making ledger entries. (iii) Consensus among participants by commitment of costly and fairly distributed resources (compute $h(m), c(m)$ or coin stake) for a fair and distributed opportunity to make ledger entries. We model a large space of design choices followed by these p2p blockchains: (i) Block sizes are flexible and their optimal choice is central to our model, (ii) Block durations are flexible, and we discuss their impact, (iii) Miner revenue is modeled as an outcome of the transaction fee auction, but we discuss intuition and show (in Appendix E) the robustness of findings even if some revenues come from block rewards in perpetuity, (iv) the PoW hashing algorithm is flexible in our model via choice of compute power distributions ($h(m), c(m)$), and (v) the PoW consensus mechanism is fixed, but we discuss the intuition behind the robustness of our findings for proof of stake (PoS) systems.

Note that among these design choices, block size and block duration are still direct levers to adjust available capacity for ledger growth rate. Other design features have a role to play by changing the incentives

Table 3 *Major peer-to-peer payment blockchains and their block size, block duration, miner rewards and consensus methods.*

Blockchain	Cap (Bn$)	Proof	Block Reward	Daily Fee	Capacity
Bitcoin	202.8	PoW	12.5 coin (4 years halving)	$545,000	1 MB per 10 min
Bitcoin Cash	6.2	PoW	12.5 coin (4 years halving)	$185	8 MB per 10 min
Bitcoin Gold	0.5	PoW	12.5 coin (4 years halving)	$908	1 MB per 10 min
LiteCoin	6.3	PoW	25 coin (4 years halving)	$1,700	1 MB per 2.5 min
Ethereum	28.8	PoW	2 coin (Variable)	$70,000	20 KB per 10 sec
Ethereum Classic	0.73	PoW	4 coin (Variable)	$77	20 KB per 10 sec
EOS	4.3	PoS	1% Inflation per year	Stake	1 MB per 0.1 sec
Steem	0.09	PoS	0.95% Inflation per year	Stake	65 KB per 3 sec

for collusion and security. First, miner revenue from block rewards weakens the incentive to fill blocks partially. However, as long as some portion of revenues comes from transaction fee offers, collusion on underfilling still remains relevant. Second, cryptographic hash puzzles can significantly change heterogeneity among participating miners, thus weakening (or strengthening) their ability to collude and double spend. Third, consensus commitments in addition to computing power (e.g., coin stake or storage space) may change some formulations, but they do not change the underlying double-spend trade-off.

Permissioned enterprise blockchains (e.g., Ripple, HyperLedger) and intermediated marketplace platforms (e.g., Airbnb, Uber, Amazon) have some similarities with public p2p blockchains. Our model does not extend to centrally permissioned enterprise blockchains. These chains may have miner revenues that are entirely outside the blockchain ecosystem. User competition for space may be obviated since extremely large storage space can easily be synced between a handful of mining nodes. These settings do not conform to zero entry barrier decentralization, which is a necessary requirement for our model. Intermediated marketplace platforms (e.g., Airbnb, Uber, Amazon) have replaced traditional large firms. These marketplaces provide low-barrier entry to small cost efficient firms (e.g., home owners, cabs, small sellers), thus significantly reducing costs. In doing so, the platform intermediary gains significant power to extract economic rents. Blockchains disintermediate this concentrated power while keeping the low costs arising from low barrier entry to firms (miners). Unfortunately, the low costs and disintermediate market power come at the cost of consensus limits, firm anonymity and lack of unilateral incentive to invest in product quality (security). This trade-off will become critical with the emergence of decentralized apps (DApps) or smart contracts that go beyond payment and try to disintermediate ride-sharing, house rentals, lending, insurance, etc. Our paper forms the basis for future research in these disintermediated marketplaces.

Design Guidance: We have shown that as long as blockchains follow the design elements discussed above, capacity will be limited by collusive economic rent extraction or security threat. Collusion overlooks small transactions, while double-spend security threatens large transactions. Therefore, it is untenable to include the entire range of payment values with vastly different willingness to pay for delay and security

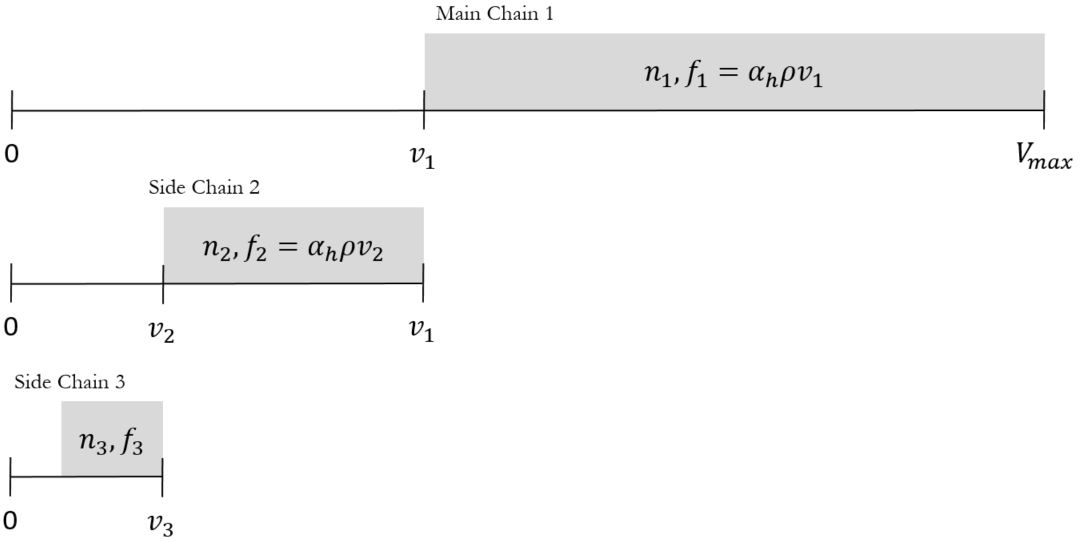


Figure 20 *(Top) The main chain caters to n_1 payments in $[V_1, V_{max}]$ at a fee of f_1. (Middle) The second chain caters to n_2 payments in $[V_2, V_1]$ at a fee of f_2. (Bottom) The third chain caters to n_3 payments in $[V_3, V_2]$ at a fee of f_3.*

onto a single chain. A natural solution is "side chains". Here, the primary chain serves the highest segment of payments with greatest willingness to pay, while smaller side chains serve lower payment value segments. Figure 20 shows the blockchain broken down into multiple chains. The first or main chain caters to the highest value payments between V_1 and V_{max}. The block size (n_1) on this chain is set below full demand (N) in order to raise sufficient transaction fee revenue ($n_1 f_1$) and mining power to secure the largest payments V_{max}. Subsequently, the second side chain caters to payments between V_2 and V_1. The block size (n_2) on this chain is set below full remaining demand ($N - n_1$) in order to raise sufficient transaction fee revenue ($n_2 f_2$) and mining power to secure the largest payments V_1. The same follows for smaller side chains. Side chains lower down the hierarchy have lower fees and lower security. Low-value payments are not incentivized to pay high fees on higher chains, while high-value payments are not incentivized to risk high-security risk on lower chains.

This design paradigm effectively discriminates among users on willingness to pay for delay and security. In fact, this starts to resemble service levels offered by financial intermediaries in lieu of proportional fee rates. Each chain offers a different delay and security guarantee. We assume that cryptographic mining is adjusted to ensure that miners are too small to sustain collusion on these chains. Note that protocol designs ($n_1, n_2, ..$) are hard to adjust in response to demand shocks ($N - \Delta N$). This may create temporary settings where, say, the main chain capacity n_1 is greater than demand ($N - \Delta N$), thus raising zero fee revenue and security guarantees. The prospect of zero no-collusion mining revenue incentivizes even small miners to engage in collusion. This acts as an endogenous mechanism for the participants to make artificial adjustments over the inflexible protocol in response to small market fluctuations. Note that this is merely a guidance, which would benefit from full-fledged technology design and adversarial attack proposals. Similar ideas albeit with different terminology or definitions have in fact also received some attention in practice.

6. Conclusion

Our work places economic bounds on the capacity of Bitcoin. In contrast to popular belief, merely upgrading the block size does not scale Bitcoin. We show why miners will underutilize block capacities to force users into competition on fees. Bitcoin mining puzzles could be redesigned to eliminate collusion-driven artificial constraints. This approach results in too little mining revenue and security guarantees. Interestingly, collusion can be seen as a platform attribute rather than a shortcoming. The purpose of p2p blockchains is to provide disintermediated payment service. Disintermediation inherently creates challenges with investment in security and inflexibility of protocol. (i) Security on the platform is an outcome of total mining power. No individual service provider (miner) has incentive to invest in security because upgrade in mining power raises payment security on blocks created by any miner, not just the miner who does incremental investment in mining power. Collusion raises additional revenues; a chunk of these revenues go toward adding more mining equipment by new miners. Therefore collusion among firms is a useful mechanism for firms to coordinate investment in security of payments. (ii) Protocol (e.g., block size, duration, mining puzzle) is inflexible because any change requires offline agreement between participants to upgrade their software. Collusion acts as an endogenous mechanism for the participants to make artificial adjustments (e.g., n_P) over the inflexible protocol (e.g., n_F) in response to small exogenous fluctuations in demand, payment spectrum or off-chain fee rates.

A few limitations of our research are worth highlighting. First, we largely ignore Bitcoin features, such as privacy, and traditional off-chain services (e.g., customer service and credit) in the user's utility. Researchers in the future could model user choice more thoroughly. Second, investors and speculators have played a major role in Bitcoin valuation and demand. We model users and miners and discuss some options for Bitcoin designers. However, we refrain from modeling investors. Nevertheless, we expect the core intuition in the paper to be broadly applicable to a wide range of peer-to-peer payment blockchain solutions that roughly follow the paradigm kickstarted by Nakamoto 2008.

References

Abadi, J., and M. Brunnermeier. 2018. Blockchain economics. Technical report, National Bureau of Economic Research.

Arnosti, N., and S. M. Weinberg. 2018. Bitcoin: A natural oligopoly. *arXiv preprint arXiv:1811.08572*.

Arora, A., A. Greenwald, K. Kannan, and R. Krishnan. 2007. Effects of information-revelation policies under market-structure uncertainty. *Management Science* 53 (8): 1234–1248.

ArsTechnica 2014. Bitcoin pool GHash.io commits to 40% hashrate limit after its 51% breach. *arstechnica.com*.

Asvanund, A., K. Clay, R. Krishnan, and M. D. Smith. 2004. An empirical analysis of network externalities in peer-to-peer music-sharing networks. *Information Systems Research* 15 (2): 155–174.

August, T., M. F. Niculescu, and H. Shin. 2014. Cloud implications on software network structure and security risks. *Information Systems Research* 25 (3): 489–510.

Biais, B., C. Bisiere, M. Bouvard, and C. Casamatta. 2018. The Blockchain Folk Theorem.

Bitcoin Wiki 2017. SegWit2x. Available at:
`https://en.bitcoin.it/wiki/SegWit2x`.

bitinfocharts 2018. Bitcoin Avg. Transaction Value historical chart. *https://bitinfocharts.com*.

Blockchain.com 2018. *https://www.blockchain.com/en/btc/tx/261d69b25896034325d8ad3e0668f963346fd79baefb6a7*

BlockchainExplorer 2018. *Blockchain.com*.

Budish, E. 2018. The Economic Limits of Bitcoin and the Blockchain. *Chicago Booth Research Paper*.

Buntinx, J. 2017. Bitcoin's Transaction fee exceeds $50 as Network Issues Remain. *newsbtc.com*.

Carlsten, M., H. Kalodner, S. M. Weinberg, and A. Narayanan. 2016. On the instability of bitcoin without the block reward. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 154–167. ACM.

Casey, M. J. 2019. The Ethereum Classic Attacker Has Sent a Bigger Message. *coindesk.com*.

Cezar, A., H. Cavusoglu, and S. Raghunathan. 2013. Outsourcing information security: Contracting issues and security implications. *Management Science* 60 (3): 638–657.

Chiu, J., and T. V. Koeppl. 2017. The economics of cryptocurrencies–bitcoin and beyond. *Available at SSRN 3048124*.

Cong, L. W., and Z. He. 2019. Blockchain disruption and smart contracts. *The Review of Financial Studies* 32 (5): 1754–1797.

Cong, L. W., Z. He, and J. Li. 2018. Decentralized mining in centralized pools.

Cong, L. W., Y. Li, and N. Wang. 2018. Tokenomics: Dynamic adoption and valuation.

Crypto51 2019. PoW 51% Attack Cost. *crypto51.app*.

Dey, D., A. Kim, and A. Lahiri. 2018. Online Piracy and the "Longer Arm" of Enforcement. *Management Science*.

Easley, D., M. O'Hara, and S. Basu. 2019. From mining to markets: The evolution of bitcoin transaction fees. *Journal of Financial Economics*.

Eyal, I., A. E. Gencer, E. G. Sirer, and R. Van Renesse. 2016. Bitcoin-NG: A Scalable Blockchain Protocol. In *NSDI*, 45–59.

Eyal, I., and E. G. Sirer. 2018. Majority is not enough: Bitcoin mining is vulnerable. *CACM* 61 (7): 95–102.

Gao, X., W. Zhong, and S. Mei. 2013. Information security investment when hackers disseminate knowledge. *Decision Analysis* 10 (4): 352–368.

Gervais, A., G. O. Karame, K. Wüst, V. Glykantzis, H. Ritzdorf, and S. Capkun. 2016. On the security and performance of proof of work blockchains. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 3–16. ACM.

Glazer, E. 2018. Justice Department Probing Wells Fargo's Wholesale Banking Unit. *Wall Street Journal*.

Glazer, E., and M. Farrell. 2018. Big U.S. Banks Face Increase in Attempted Cyberattacks. *Wall Street Journal*.

Green, E. J., and R. H. Porter. 1984. Noncooperative collusion under imperfect price information. *Econometrica: Journal of the Econometric Society*:87–100.

Hagiu, A., and J. Wright. 2015. Multi-sided platforms. *International Journal of Industrial Organization* 43:162–174.

Halaburda, H., and G. Haeringer. 2018. Bitcoin and blockchain: what we know and what questions are still open. *NYU Stern School of Business, Forthcoming*.

Hinzen, F. J., K. John, and F. Saleh. 2019. Proof-of-Work's Limited Adoption Problem. *NYU Stern School of Business*.

Hsu, C., J.-N. Lee, and D. W. Straub. 2012. Institutional influences on information systems security innovations. *Information systems research* 23 (3-part-2): 918–939.

Huberman, G., J. Leshno, and C. C. Moallemi. 2019. An economic analysis of the Bitcoin payment system. *Columbia Business School Research Paper* (17-92).

Johar, M., S. Menon, and V. Mookerjee. 2011. Analyzing sharing in peer-to-peer networks under various congestion measures. *Information Systems Research* 22 (2): 325–345.

Jordan, D., and S. S. Kerr. 2018. Baffled by Bitcoin? How Cryptocurrency Works. *Wall Street Journal*.

Kannan, K., and R. Telang. 2005. Market for software vulnerabilities? Think again. *Management science* 51 (5): 726–740.

Kroll, J. A., I. C. Davey, and E. W. Felten. 2013. The Economics of Bitcoin Mining, or Bitcoin in the Presence of Adversaries. In *Proceedings of WEIS*, Volume 2013, 11.

Li, Z., and A. Agarwal. 2016. Platform integration and demand spillovers in complementary markets: evidence from Facebook's integration of Instagram. *Management Science* 63 (10): 3438–3458.

Malinova, K., and A. Park. 2017. Market Design with Blockchain Technology.

McMillan, R. 2018. Thieves Can Now Nab Your Data in a Few Minutes for a Few Bucks. *Wall Street Journal*.

Nakamoto, S. 2008. Bitcoin: A Peer-To-Peer Electronic Cash System.

Natoli, C., and V. Gramoli. 2017. The Balance Attack or Why Forkable Blockchains Are Ill-Suited for Consortium. In *Dependable Systems and Networks (DSN), 2017 47th Annual IEEE/IFIP International Conference on*, 579–590. IEEE.

Nayak, K., S. Kumar, A. Miller, and E. Shi. 2016a. Stubborn Mining: Generalizing Selfish Mining and Combining with an Eclipse Attack. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, 305–320. IEEE.

Nayak, K., S. Kumar, A. Miller, and E. Shi. 2016b. Stubborn mining: Generalizing selfish mining and combining with an eclipse attack. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 305–320. IEEE.

NeonVest 2018. The Scalability Trilemma in Blockchain. *Medium.com*.

Osborne, C. 2018. Bitcoin Gold suffers double spend attacks, $17.5 million lost. *zdnet.com*.

Popper, N. 2017. Bitcoin Hasn't Replaced Cash, but Investors Don't Care. *The New York Times*.

Prat, J., and B. Walter. 2018. An equilibrium model of the market for bitcoin mining.

Shy, O., and Z. Wang. 2011. Why do payment card networks charge proportional fees? *American Economic Review* 101 (4): 1575–90.

Sompolinsky, Y., and A. Zohar. 2015a. Secure High-Rate Transaction Processing in Bitcoin. In *International Conference on Financial Cryptography and Data Security*, 507–527. Springer.

Sompolinsky, Y., and A. Zohar. 2015b. Secure high-rate transaction processing in bitcoin. In *International Conference on Financial Cryptography and Data Security*, 507–527. Springer.

Tasca, P. 2018. The Hope and Betrayal of Blockchain. *The New York Times*.

TransferWise 2018. How long does an international wire transfer take? *transferwise.com*.

Wei, Z., and M. Lin. 2016. Market mechanisms in online peer-to-peer lending. *Management Science* 63 (12): 4236–4257.

Wilmoth, J. 2018. The First 8MB Bitcoin Cash Block Was Just Mined.

Yli-Huumo, J., D. Ko, S. Choi, S. Park, and K. Smolander. 2016. Where is current research on blockchain technology?—a systematic review. *PloS one* 11 (10): e0163477.

Zaitsev, D. 2018. Quarterly Cryptocurrency and ICO Market Analysis. *Medium.com*.

Appendices

Proof of Proposition 4:

The verifiability of a miner's action is a key requirement for tacit collusion. Colluding miners should be able to verify each other's partial block filling actions. New blocks on the Bitcoin network – full or partial – are public. Colluding miners can perform their actions using a unique public address to add their blocks. An alternate method is for colluding miners to pool their computational power. For example, Antpool is a group of miners who have decided to pool their resources together for mining. Each miner participating in a pool contributes a verifiable amount of computational power and submits blocks under the pool ID. The pool then distributes the rewards from winning blocks based on the work performed. Regardless of whether the miners gather together in a pool or perform their actions using a unique public address, they can still deviate. However, such a deviation is easily detectable (Cong and He 2019, Malinova and Park 2017). If the colluding miners use the same public address, transactions included in any block that they win will reveal if they are deviating. However, it is possible for the miner to use a different public address when cheating. In a pool, this would not be possible because the pool can verify the computational power contributed by the miner to the pool and also the transactions included by the miner in a block. Outside of a pool, a drop in the win rate of a large miner (identified by the public address) may indicate that he is cheating using another address to submit fully filled blocks. Further, a colluding miner does not need to verify every other miner's actions. Overall, if α_l fraction of the total computational power is assumed to be involved in collusion, a participating miner simply needs to check that on average α_l fraction of blocks are partially filled to ensure that no one is deviating.

Every miner prepares a block of transactions and then starts to look for the solution to the Bitcoin mining puzzle. Miners need to decide their filling action – full or partial – before starting to find the mining puzzle solution. A colluding miner considers a trade-off (equation 44) between (a) immediate revenues from a partial fill n_P^* facing a bid vector $f(v)$ i.e., R_{noDev}, followed by expected collusion profits R_P for T periods or (b) deviation to maximize revenues from current block R_{noDev}, followed by expected no collusion profits R_0 for T periods. Recall that δ represents the time discounting factor for Bitcoin users, which can potentially be different from that for miners (δ_m). A colluding miner of size α_j will not deviate from collusion if

$$R_{noDev} + \alpha_j R_P * \frac{\delta_m(1-\delta_m^T)}{1-\delta_m} \geq R_{Dev} + \alpha_j R_0 * \frac{\delta_m(1-\delta_m^T)}{1-\delta_m}, \tag{44}$$

This simplifies to,

$$\alpha_j \geq \frac{1-\delta_m}{\delta_m(1-\delta_m^T)} \times \frac{R_{Dev} - R_{noDev}}{R_P - R_0}. \tag{45}$$

If the miner observes equilibrium fee bid $f^*(v)$ then R_{noDev} corresponds to R_P, and R_{Dev} corresponds to R_F. Further, miner must be able to sustain the collusive partial fill level n_P^* in response to any off equilibrium enticing bid vector $f(v) \neq f^*(v)$ as well. We can upper bound the right hand side $(R_{Dev} - R_{noDev})$ by considering the most enticing fee bid vector $f(v)$ similar to single strategic miner setting in the last section,

$$\alpha_j \geq \hat{\alpha}, \quad \text{where} \quad \hat{\alpha} = \frac{1-\delta_m}{\delta_m(1-\delta_m^T)} \times \frac{\alpha_h \rho N}{2V} \times \frac{v_h^2 - v_l^2}{R_P - R_0}. \tag{46}$$

This represents different trade-off faced by heterogeneous miners. A small miner $(\alpha_j < \hat{\alpha})$ mines blocks infrequently, say, once a month or year. They are less threatened by punishment far off in the future. This can also be seen as a small miner's desire to make the most out of winning a mining puzzle once in a long while. A large miner sticks to the collusion strategy, as they expect frequent or near-term fee revenues. The lower bound on the smallest colluding miner $(\hat{\alpha})$ can be reduced by a longer punishment strategy $(T \to \infty)$. This can be useful if small miner participation is necessary to attain a colluding group with total power α_l.

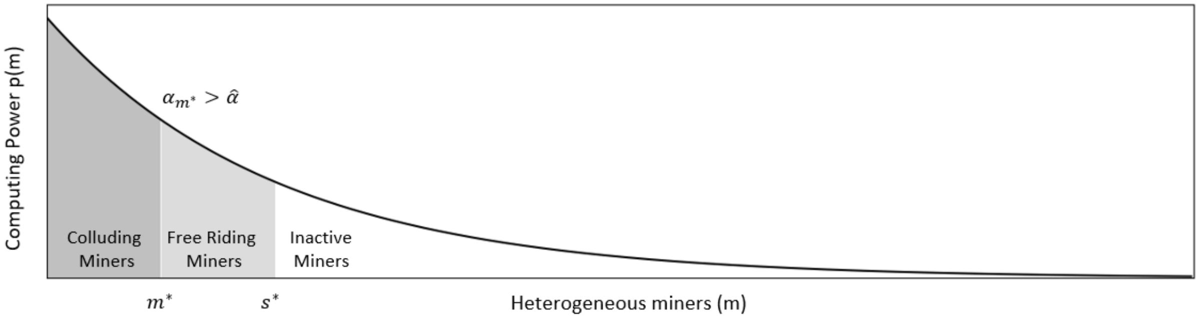


Figure 21 *The x-axis represents miners ranked by access to hardware. The y-axis represents the computing power of the respective miner. An example of a monotonically decreasing concave function $h(m) = \lambda e^{-\lambda m}$ with a long tail. Three categories of miners, from left to right, - (1) Colluding by partial block filling, (2) Free Riding by full block filling, and (3) Inactive miners. m^* and s^* represent the boundary between these groups in a collusion equilibrium.*

The constraints above ($\alpha_j \geq \hat{\alpha}$, $\Sigma\alpha_j \geq \alpha_l$) assures a colluding miner's commitment to the collusion. We could use this to check if an exogenously given distribution of mining powers (α_j) conforms to collusion requirements. However, miner entry is endogenously determined by available revenues and hardware distribution. Next we identify corresponding constraints on mining hardware distribution for collusion. This is useful because a Blockchain designer does not directly control mining powers (α_j), but they control hardware distribution via choice of mining puzzle (e.g., SHA 256 Hash, Scrypt). It would allows us to discuss potential Blockchain designs for averting collusion. Figure 21 illustrates our model of miner heterogeneity with respect to access to computing hardware. A large miner has access to hardware capable of faster hash calculations at the same cost. Given the favorable trade-off for the large miners, we focus on collusion among a group $(\Sigma\alpha_j = \alpha_l)$ of the largest miners. We must ensure that no miner (colluding, free riding, or outside) deviates from their action in equilibrium.

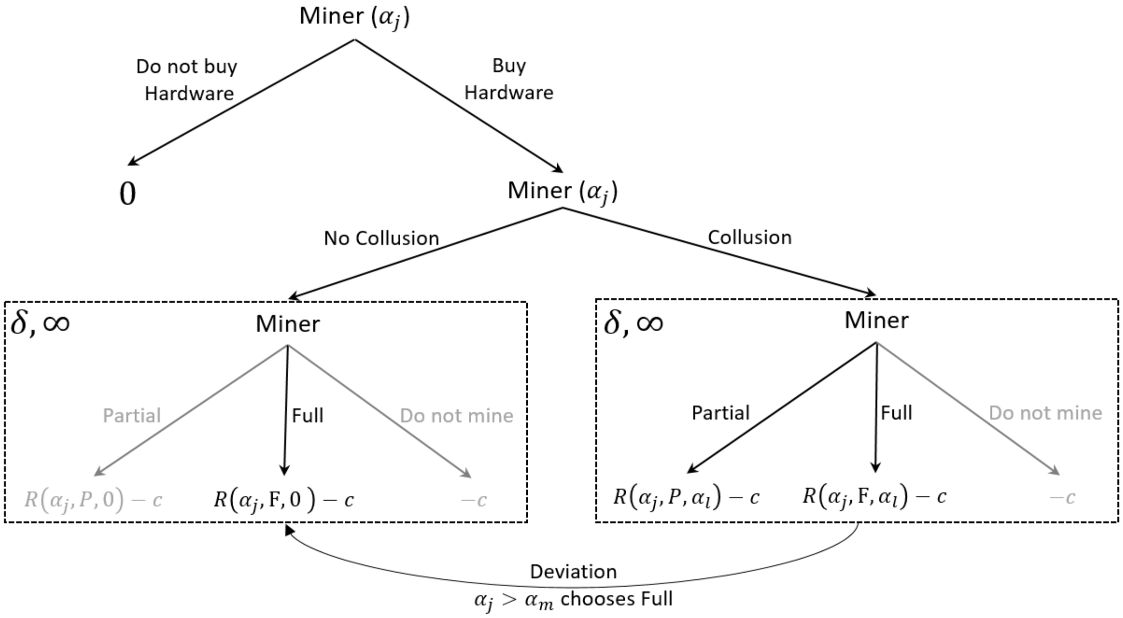


Figure 22 *Miner j with access to computing power (α_j) decides whether to buy hardware. Next, the miner decides whether to follow the collusion strategy. The colluding group of miners chooses Partial fill. The free-riding group chooses Full. Deviation by a colluding group member leads to a no-collusion setting whereby everyone always fully fills their block. The block creation sub-games on equilibrium (right) and off equilibrium (left) are repeated infinitely with a discount factor δ_m.*

Figure 22 depicts the sequence of choices for a miner. First, they decide whether to purchase the hardware. Second, they decide whether to follow the collusion strategy. The second choice is repeated over infinite block creation periods. These choices are made by all miners simultaneously. We want to identify conditions under which the top m^* miners collude and the next $(s^* - m^*)$ miners participate as free riders. In a subgame equilibrium, these participating miners should be willing to purchase the hardware and stick to the collusion strategy. All miners beyond s^* should be better off not purchasing the hardware. Individual miners have rational expectations of equilibrium strategies followed by all other miners. The focal miner first decides whether to buy the hardware. Under collusion equilibrium, a rational miner expects the top s^* miners to buy hardware. Next, they decide whether to follow the equilibrium collusion strategy or the no-collusion strategy. On the collusion path, a rational miner expects the top m^* miners to add partial blocks. On the no-collusion (off-equilibrium) path, a rational miner expects all active miners to add full blocks. On both the collusion equilibrium and the no-collusion off-equilibrium path, the focal miner exercises a choice of partial filling, full filling or no mining. The block creation sub-game on-equilibrium (right) and off-equilibrium (left) paths are repeated infinitely with a discount factor δ_m.

Single-block-fee payoffs are denoted by $R(*, *, *)$ with three arguments. The first argument represents the mining power of the focal miner. The second argument represents the focal miner block fill action, i.e.,

Table 4 *Single-period expected payoffs corresponding to three actions (Partial, Full, and No Mining) off and on collusion equilibrium paths.*

	On Equilibrium (α_l)	**Off Equilibrium** $(\alpha_l = 0)$
Partial $R(\alpha_j, P, *)$	$f_h n_P$	$f_0 n_P$
Full $R(\alpha_j, F, *)$	$f_h n_P + f_l(n_F - n_P)$	$f_0 n_F$
No Mining $R(\alpha_j, 0, *)$	0	0

Table 5 *List of all constraints that ensure that no miner has a profitable deviation in an SPE. Three pairs of constraints (1a,1b), (2a,2b) and (3a,3b) correspond to three types of miners - (1) colluding miners, (2) free-riding miners and (3) inactive miners, respectively.*

Focal Miner	**Constraint**
$\alpha_j > \alpha_m$	**1a** $R(\alpha_j, P, \alpha_l) + \alpha_j \frac{\delta_m}{1-\delta_m} R(\alpha_j, P, \alpha_l) \geq R(\alpha_j, F, \alpha_l) + \alpha_j \frac{\delta_m}{1-\delta_m} R(\alpha_j, F, 0)$
	1b $\frac{1}{1-\delta_m} R(\alpha_j, P, \alpha_l) - c_j \geq 0$
$\alpha_m \geq \alpha_j \geq \alpha_s$	**2a** $R(\alpha_j, F, \alpha_l) + \alpha_j \frac{\delta_m}{1-\delta_m} R(\alpha_j, F, \alpha_l) \geq R(\alpha_j, P, \alpha_l + \alpha_j) + \alpha_j \frac{\delta_m}{1-\delta_m} R(\alpha_j, P, \alpha_l + \alpha_j)$
	2b $\frac{1}{1-\delta_m} R(\alpha_j, F, \alpha_l) - c_j \geq 0$
$\alpha_j \leq \alpha_s$	**3a** $\alpha_j \frac{1}{1-\delta_m} R(\alpha_j, F, 0) - c_j \leq 0$
	3b $\frac{1}{1-\delta_m} R(\alpha_j, P, \alpha_l + \alpha_j) - c_j \leq 0$

partial (P) or full (F). The third argument represents the colluding group power, i.e., collusion (α_l) or no collusion (0). Table 4 provides the single-period expected payoffs corresponding to all actions. Single-period payoffs are strictly better under the full block filling action for all miners. If the focal miner has power $\alpha_j \geq \alpha_m$ but decides to add a full block under collusion, they expect to be punished. All miners would move to the no-collusion sub-game if a single miner deviates from the collusion.

Table 5 lists all constraints that ensure that no miner has a profitable deviation in an SPE. For a large focal miner $(\alpha_j > \alpha_m)$, constraint 1a ensures that they prefer to add partial blocks rather than a full block in the repeating block creation sub-game. This is fulfilled if the marginal miner satisfies $\alpha_m \geq \hat{\alpha}$. Constraint 1b represents their preference to buy the hardware at the start of the game. This is satisfied for the marginal miner making at least zero profits.

$$\alpha_m = \frac{c_m(1-\delta)}{R_P} \geq \hat{\alpha}; \quad \text{where} \quad R_P = f_h n_P \tag{47}$$

A miner with $(\alpha_m \geq \alpha_j \geq \alpha_s)$ proportion of the total power free rides. The lower limit α_s denotes the smallest miner that joins the mining network. Constraint 2a represents the preference to free ride over joining the colluding group. Joining the colluding group would increase the power of the colluding group to

$\alpha_l + \alpha_j$ and therefore the partial block revenues. If the marginal miner $(\alpha_j = \alpha_m)$ is large enough, they may increase the partial block revenues $R_P(\alpha_l + \alpha_j)$ to be higher than the full block revenue $R_F(\alpha_l)$. In this section, we are interested in settings whereby even the largest miner is too small to perform partial block filling without a threat of punishment [18]. Large miners unilaterally perform partial block filling as shown in Section 3.2.

Constraint 2b represents free-riding miners' preference to buy the hardware at the start of the game. This is satisfied for the smallest miner making positive profits.

$$\alpha_s \geq \frac{c_s(1-\delta)}{R_F}; \quad \text{where} \quad R_F = f_h n_P + f_l(n_F - n_P) \tag{48}$$

For a focal miner who stays out $(\alpha_j \leq \alpha_s)$, constraint 3a represents their preference to not join as a free rider. Joining in as a free rider reduces the power of the colluding group below α_l and makes collusion unprofitable. Since the smallest miner makes zero profits when free riding, they are guaranteed to make negative profits when collusion breaks. Finally, constraint 3b represents their preference to join the colluding group. Similar to the free rider, this increases the power of the colluding group to $\alpha_l + \alpha_j$ and therefore the partial block revenues. These partial block revenues $R(\alpha_l + \alpha_j, P, \alpha_j)$ must be smaller than zero. This is automatically satisfied since $\alpha_s < \alpha_m$.

We now proceed to obtain the equilibrium expressions for the smallest colluding miner, denoted by m^*, and the smallest free-riding miner, denoted by s^*. For the smallest colluding miner, we need constraint 1b to become an equality. Specifically, we need

$$\frac{1}{1-\delta} R(\alpha_m, P, \alpha_l) - c_m = 0.$$

We know that $R(\alpha_m, P, \alpha_l) = \alpha_m \times R_P$. Thus, from the above equation, we have

$$\alpha_m \times R_P = c_m(1-\delta)$$

or equivalently

$$\alpha_m = \frac{c_m(1-\delta)}{R_P}. \tag{49}$$

From the definition of α_l, we know that

$$\alpha_l = \frac{H(m^*)}{H(s^*)}.$$

Thus, we have

$$H(s^*) = \frac{H(m^*)}{\alpha_l}. \tag{50}$$

[18] An equilibrium at α_l is only valid when individual miners are relatively small: $\alpha_m \leq R_P^{-1}(R_F(\alpha_l)) - \alpha_l)$

From the definition of α_j, we also know that

$$\alpha_m = \frac{h(m^*)}{H(s^*)}.$$

Substituting the expression of $H(s^*)$ from (50), we obtain

$$\alpha_m = \frac{h(m^*)\alpha_l}{H(m^*)} = \Lambda(m^*)c(m^*)\alpha_l,$$

where $\Lambda(m) \equiv \frac{h(m)}{H(m)c(m)}$. Using (49), we have

$$\Lambda(m^*)c(m^*)\alpha_l = \frac{c(m^*)(1-\delta)}{R_P},$$

or

$$m^* = \Lambda^{-1}\left(\frac{1-\delta}{\alpha_l R_P}\right). \tag{51}$$

The monotonically decreasing function $\Lambda \equiv \frac{h(m)}{H(m)c(m)}$ ensures unique solutions. The marginal colluding miner earns zero profit and must be large enough such that future punishment is a credible threat ($\alpha_{m^*} \geq \hat{\alpha}$).

$$\frac{c(m^*)(1-\delta_m)}{R_P} \geq \hat{\alpha} \tag{52}$$

Free-riding miners present in equilibrium enter until colluding miners have exactly α_l proportion of the total computing power. Using (50), we have

$$s^* = H^{-1}\left(\frac{H(m^*)}{\alpha_l}\right), \tag{53}$$

where m^* is given in (51). In addition, the smallest free-riding miner must be large enough to make positive profits.

$$\alpha_{s^*} = \frac{h(s^*)}{H(s^*)} \geq \frac{c(s^*)(1-\delta_m)}{R_F} \tag{54}$$

■

The subgame perfect equilibrium above focuses on a colluding group with exactly α_l power. All miners adding partially filled blocks ($\Sigma\alpha_j = 1$) is yet another SPE. In such a case, users either stay off the chain or offer a high fee (f_h); no one offers a low fee (f_l). As a result, a deviation to add a full block (R_F) with low-fee-paying transactions is not an option. We have not observed such full collusion on the Bitcoin network. We do not provide a specific justification for one equilibrium over other; however, we focus on the α_l collusion as a more interesting and practical setting. In addition to these extreme cases, collusive equilibria with $\alpha_l \leq \Sigma\alpha_j \leq \alpha_h$ may also be possible. If a single miner with power $\alpha_{j'}$ deviates from such collusion, the remaining group is left with power $\Sigma\alpha_j - \alpha_{j'}$. Punishing the deviating miners requires this group to not engage in collusion permanently. This is not necessarily a rational strategy for the remaining $\Sigma\alpha_j - \alpha_{j'}$ group at this stage. They are better off colluding on partial block filling if $\Sigma\alpha_j - \alpha_{j'} > \alpha_l$. This rational strategy to collude with a smaller group does not constitute a threat to the deviating miner. The deviating miner is thus better off by continuing to free ride. In this paper, we do not validate or reject the existence of miner strategies that sustain such equilibria.

A. Off Chain Payment

In practice, off-chain alternative for Bitcoin is a combination of Credit Cards, PayPal and Wire Transfer (SWIFT). Following is a very rough estimate of proportional fees charged and time to complete for these off-chain modes on an international payment. While Credit Cards offer almost instant payment faster than Bitcoin, Wire transfers are much slower than Bitcoin. Our writing did not provide this detail on off-chain alternatives.

Mode	Fees	Time
SWIFT (Wire)	1-3% (fx markup) + \$50 (payer bank) + \$20 (correspondent bank) + \$20 (receiver bank) +	3 Days
PayPal	3% (to send) + 4.4% (to receive)	3+ Days
Credit Card	1-5% (to send) + 3% (to receive)	Instant

The utility of making an international payment v using mode m could be modeled as $U_m(v)$, where ρ_m is the proportional fees and $\delta(t_m)$ is the discount factor over the payment completion time. The off-chain modes offer different combination of fees and time to completion (ρ_m, t_m). We assume a single proportional cost $\rho = \rho_m + \delta(t_m)$ as a combination of fees and time delay. This is reasonable because typically low proportional rate options have large delay and vice versa.

$$U_m(v) = v - \rho_m v - \delta(t_m)v$$

$$U_m(v) = v - \rho v \qquad \text{where } \rho = \rho_m + \delta(t_m)$$

Modeling the on-chain time discount factor is critical because time to completion is endogenously determined by user action (e.g., fee offer) and competition. Modeling the off-chain discount factor is not necessary because it is fixed in advance irrespective of user action. It is accounted for via (ρ). The choice of off-chain features (ρ_m, t_m) provided by Banks or Credit Cards in response to user choices are outside the scope of our research.

B. User Collusion

Let us consider a scenario where miners act passively by adding full blocks (n_F). If users compete every period, top n_F users each pay fees $f_0 = \alpha_h \rho(1-\gamma)V$. Let us consider users collusion such that the top n_F users pay a lower fees $f_c < f_0$ while the remaining users stay off-chain instead of competing with the top n_F users on fee. The marginal user $v = v_0$ has the highest value payment among users who stay off chain. This user has greatest incentive to deviate by offering a fees $f_c + \epsilon$ to complete their payment on chain. This user can be prevented from deviating to $f_c + \epsilon$ fee bid by a punishment threat i.e. no collusion in future. This threatens the marginal user because they will likely draw a payment need $v \in [v_0, V_{max}]$ in subsequent periods. A break down in collusion means that they will not be able to benefit from the low collusion fees f_c $(< f_0)$. The marginal user pays $f_c + \epsilon$ by deviating and expects to pay no collusion fees $E_v^0[f(v)]$ in future periods. Otherwise, they stay off chain and pay ρv_0 in the current period and expects to pay lower collusion fees $(E_v^c[f(v)])$ in future. The collusion is sustained if the future punishment would lead to larger lifetime fee payments compared to no deviation.

$$f_c + \frac{\delta}{1-\delta} * E_v^0[f(v)] \geq \rho v_0 + \frac{\delta}{1-\delta} * E_v^c[f(v)] \tag{55}$$

where,

$$E^c_{v \sim U[0,V]}[f(v)] = (1 - \frac{v_0}{V_{max}}) * f_c + \int_0^{v_0} \rho \frac{v}{V} dv \tag{56}$$

and

$$E^0_{v \sim U[0,V]}[f(v)] = \underbrace{(1 - \frac{v_0}{V}) * f_0}_{\text{On Chain Fee}} + \underbrace{\int_0^{v_0} \rho \frac{v}{V} dv}_{\text{Off Chain Fee}} \tag{57}$$

This simplifies to,

$$\delta > \frac{1}{1 + \frac{n_F}{N}} \;; \quad \text{or} \quad n_F > N \times (\frac{1-\delta}{\delta}) \tag{58}$$

Interestingly this condition on the discount factor δ is independent of the equilibrium choice of f_c $(< f_0)$. In fact $f_c = 0$ is pareto optimal collusion fees for all users. This condition captures the intuition that the user collusion is averted if - (i) the block capacity is extremely small compared to the demand. This means that the marginal user v_0 strongly prefers to get on-chain now, instead of waiting for the low likelihood event of being one of top $n_F << N$ users in the near future. (ii) users transact infrequently (small δ) i.e. the demand N every period is made up of payment needs for a small fraction of overall users. (iii) user payment value v is sampled once instead of being randomly sampled from $[0, V_{max}]$ every period. Small value users are always small and large value are always large. If none of these three settings hold then user collusion at zero fees is trivial.

Theoretically, colluding miners can respond to this user collusion by only including payments that bid $f_h = \alpha_h \rho V/2$. Remember that f_h was the fee offered when all miner collude to add revenue maximizing blocks at half the demand $n_P = N/2$. Since both users and miners are forward looking, both are willing to forego payments via Blockchain and revenue from Blocks respectively for a few periods. The relative values of discount factors (δ, δ_m) and collusion savings would determine which side comes out on top in threatening the other into deviating from collusion. We skip delving deeper into this formulation because of limited evidence of forward looking user fee biding or strategies - counter strategies between users and miner. Future research could delve into some related questions - (i) Would forward looking (even colluding) users account for security risks and pay fees higher than the outcome of competitive auction? (ii) Would large users that comprise say 20% of all payment demand (e.g. major Bitcoin-USD exchanges) have strategies to unilaterally alter fees, delay and security to their advantage?

C. User Waiting

Our primary model assumes that users are impatient. They bid on chain and wait a maximum of one block period for their payment to be included. If not, they complete the payment off chain. In this section we ground this assumption to rational user behavior. Note that we originally modeled user payment values as continuously distributed in $[0, V]$. This was done to simplify the exposition of our research question. Here we take a more realistic set up wherein, N users that arrive every period have payment that take on one of V possible discrete values $1, 2, ..., V$. There are exactly N/V users at every payment value $v \in 1, 2, ..., V$. The discrete levels can be arbitrarily closely spaced. Similarly, fee bids are offered in discrete increments with ϵ (e.g. 1 cent or 1 satoshi) being the smallest increment. We will show a stable equilibrium where no user makes an on chain bid that results in a wait longer than single block period. Users either bid rationally expecting to be included on the immediate block or they go off chain.

At the outset let us assume that n_w users are present in the waiting queue, while N new users arrive in any given period. Every period miners add a block on chain, including top n_F $(< N + n_w)$ payments by fee offers. We want to find an equilibrium bid function $f^*(v)$ and the expected wait time $w^*(v)$ for a user with payment value v. Let v_0 be a payment value such that,

$$N\frac{V - v_0}{V} \leq n_F \leq N\frac{V - v_0 + 1}{V} \tag{59}$$

Figure 24 shows corresponding three segment of users - high value, low value and marginal users.

Let $\bar{f}(v, w)$ be the maximum fees a user is willing to bid for a payment of value v and a wait time w. At this maximum fees the user is indifferent between on chain and off chain proportional fees. $\bar{f}(v, w)$ is naturally increasing in v, we investigate equilibrium fee bid function $f^*(v)$ that is monotonically increasing in v since a user with larger payment value has a higher willingness to bid owing to costlier off chain option ρv and greater per period delay cost δv. All **high value** users $v \geq v_0 + 1$ are in top n_F new payment value arrivals, they offer an equilibrium fee higher than all users with payment values $v \leq v_0$. They do not need to wait i.e. $w^*(v) = 1 \ \forall \ v \geq v_0 + 1$. Note that none of the bids in the waiting queue belong to payment value $v \geq v_0 + 1$, thus do not compete with these high value users anyways.

All **low payment value** users $v \leq v_0 - 1$ are outside the top n_F new payment value arrivals. They are competing against more than n_F higher value users in this period. In equilibrium, where fee offers are monotonically increasing, these users do not expect to find space on the immediate block. Even if they bid and wait in queue, they will be competing against more than n_F higher value payments in all future periods. Their attempt to wait in queue will be futile in perpetuity. All these low value users are better off going off chain.

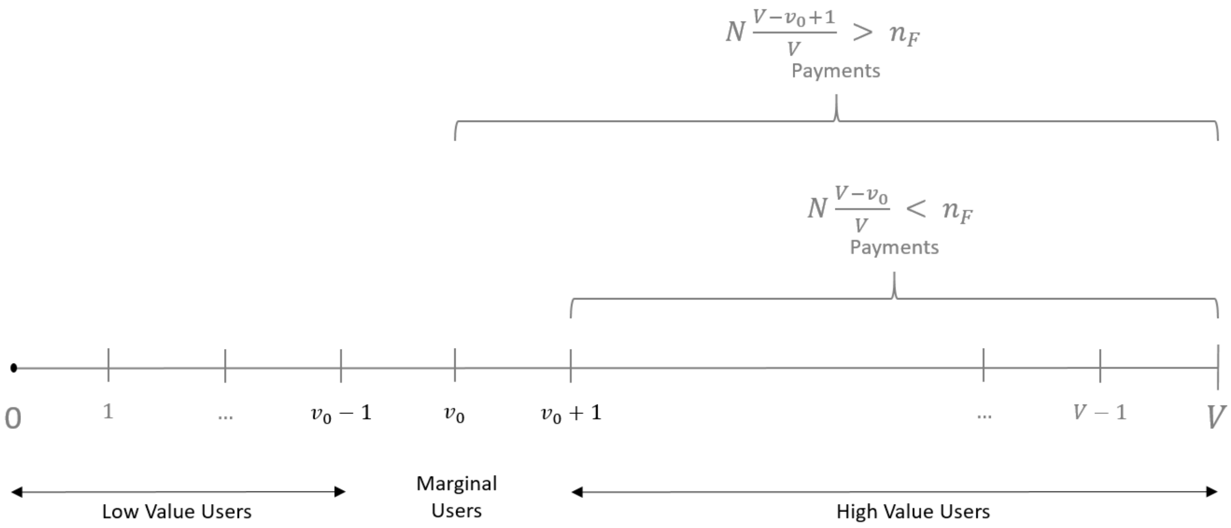


Figure 23 *Every block period N user payments arrive distributed uniformly over values $1,2,...,V$. Users with payment value v_0 are alluded to as marginal users. Some of these users but not all can find space on the immediate block the rest may have to either go off chain or remain in waiting queue. Higher value users ($v \geq v_0+1$) always find space on the immediate block, lower value users ($v \leq v_0-1$) never find space on any future block.*

Finally, all **marginal** users with payment value $v = v_0$ have a probabilistic shot at getting into top n_F bids. Let n_a $(= N/V)$ be number of users with payment value v_0 that arrive every period. Let n_w be number of users with payment value v_0 that have been waiting in queue for at least one period. n_d $(= n_F - N(V-v_0)/V)$ is the remaining capacity on any given block after the miner has included all payment bids with $v \geq v_0+1$. Only n_d out of $(n_a + n_w)$ can get onto the immediate block. New arriving users can either go off chain or place a bid on chain for these limited spots. We search for a mixed strategy equilibrium where these users randomize i.e. a fraction of these users $n_{a,on}$ place an on chain bid, while the remaining go off chain $n_{a,off}$ $(n_a - n_{a,on})$. If equilibrium $n^*_{a,on}$ or $n^*_{a,off}$ turn out to be zero, it collapses to a pure strategy.

In steady state, the total number of users vying for an on chain spot $(n_{a,on}+n_w)$ must equal the users that are included plus the users that are left waiting $(n_d + n_w)$ i.e. $n_{a,on} = n_d$. Thus in equilibrium n_d out of n_a users with payment value v_0 make an on chain fee offer $f^*(v_0)$ while the rest $n_a - n_d$ go off chain. Figure 24 shows a single block period arrival and payment completions of these marginal users $v = v_0$. The expected waiting time for these users is given by,

$$w^*(v_0) = 1 \times \frac{n_d}{n_w+n_d} + 2 \times \frac{n_w}{n_w+n_d}\frac{n_d}{n_w+n_d} + 3 \times (\frac{n_w}{n_w+n_d})^2 \frac{n_d}{n_w+n_d} + ... \tag{60}$$

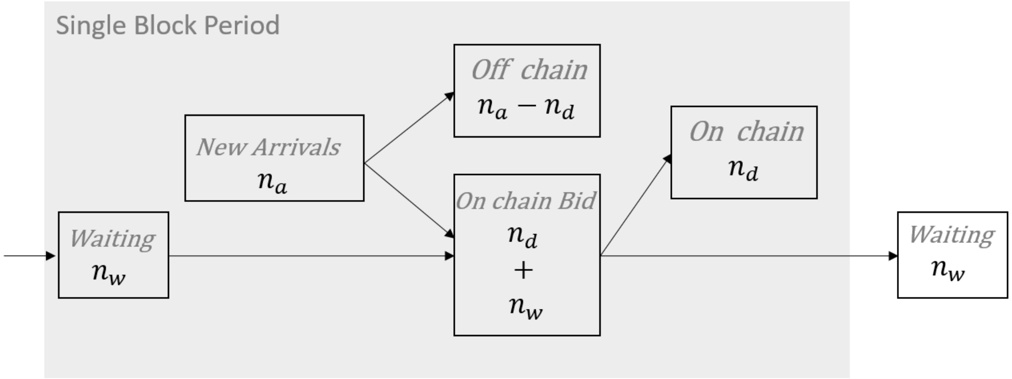


Figure 24 *Every period starts with n_w waiting on chain bids and n_a new payment arrivals. Some the new arrivals ($n_a - n_d$) go off chain, while the rest (n_d) make an on chain bid. Miners pick n_d random payments from the total $n_d + n_w$ on chain bids. This leaves n_w payments waiting for the next block period.*

From the setup above, we investigate equilibrium in following strategy space with unknowns (f_h^*, f_m^*, f_l^*) resulting in steady state waiting queue length (n_w^*) and waiting time (w^*),

$$f^*(v) = f_h^*, w^*(v) = 1 \text{ where } v \geq v_0 + 1 \tag{61}$$

$$f^*(v) = \begin{cases} f_m^* \text{ with probability } \frac{n_d}{n_a}, \\ f_l^* \quad \text{ with probability } \frac{n_a - n_d}{n_a} \end{cases}, w^*(v) = w^* \text{ where } v = v_0 \tag{62}$$

$$f^*(v) = f_l^*, w^*(v) = \infty \text{ where } v \leq v_0 - 1 \tag{63}$$

This equilibrium is stable if,

- High value users $v \geq v_0 + 1$ bid greater than all other users and are better off than the off chain option.

$$\bar{f}(v_0 + 1, 1) > f_h^* > f_m^* > f_l^* \tag{64}$$

- Low value users $v \leq v_0 - 1$ are better off with the off chain option rather than competing with marginal or high value users.

$$\bar{f}(v_0 - 1, 1) < f_m^* < f_h^* \tag{65}$$

- Marginal user could deviate to a higher bid $(f_m^* + \epsilon)$. By doing so, they overcome all other waiting or new arrival marginal users who all bid f_m^*. Thus guaranteeing inclusion on the immediate block without additional wait $(w = 1)$. This deviation is not profitable if the increased bid is greater than maximum willingness to bid for an immediate inclusion $\bar{f}(v_0, 1)$.

$$f_m^* + \epsilon > \bar{f}(v_0, 1) \tag{66}$$

Further, marginal user $v = v_0$ must bid less than or equal to their maximum willingness to bid $(\bar{f}(v_0, w^*))$ at a wait of w^*.

We also know that $\bar{f}(v_0, w^*) \leq \bar{f}(v_0, w = 1)$ since user always willing to bid more for minimal waiting $w = 1$. All the constraints above can be written as,

$$\bar{f}(v_0 + 1, 1) > f_h^* > f_m^* + \epsilon > \bar{f}(v_0, 1) \geq \bar{f}(v_0, w^*) > f_m^* > \bar{f}(v_0 - 1, 1) \tag{67}$$

This condition is satisfied for an arbitrarily small ϵ for following equilibrium strategy,

$$f_h^* = \bar{f}(v_0, 1) + \epsilon \quad ; \quad f_m^* = \bar{f}(v_0, 1) \quad ; \quad f_l^* = \bar{f}(v_0, 1) - \epsilon \tag{68}$$

In this equilibrium, marginal users randomize between on chain and off chain payments with probability n_d/n_a and $1 - n_d/n_a$ respectively. But when bidding on chain they do not expect any wait time ($w^* = 1$) or any waiting queue ($n_w = 0$).

Underlying the formulation above is the simple intuition that the marginal user is competing against all other marginal users in the current block period as well as marginal users that arrive in future period. If they bid anything less than full willingness to pay $\bar{f}$ than they will be superseded by the competition in this period as well as all subsequent periods in perpetuity. Note that this happens in particular because number of arriving users N and block capacity n_F are not stochastic. Waiting queues will emerge if these are considered to be stochastic. This is in-fact the case with Huberman et al. 2019 and Easley et al. 2019, who incorporate stochasticity but assume that users bid without knowledge of the exact bids in the waiting queue, instead relying on expectations of average waiting bids.

D. User Security Utility

In this section we discuss alternative payment value distribution and security cost distribution. We show in Section 4.1 the largest payment secured V_{secure} when the demand is made up on $N(d, V)$ payments uniformly distributed in $[0, V]$. Now we contrast with a setting where demand is made up of $N(d, V)$ payments exponentially distributed in $[0, V]$. The left half in Figure 25 shows two different payment distributions. The right half in Figure 25 shows the largest payment secured when demand is made up of $N(d, V)$ payments distributed uniformly or exponentially in $[0, V]$. In case of uniform distribution, the largest payment secured increases linearly with payment upper bound V until V_{max}. Participation of large value payments both raises and needs greatest level of security.

This is not the case with exponential distribution of payment demand. Increase in revenue with larger payments grows slowly because there are fewer large payments. This revenue growth may not be sufficient to provide security to increasingly higher payments. A middle segment of user with high enough fee savings and low enough security risk may have greatest willingness to participate on the Blockchain. Contemporaneous work by Chiu and Koeppl 2017, consider settings where a large number of small payment demand are compared against small number of large payment demand. Not surprisingly, such a setting resembles exponential distribution and leans in favor of the former i.e. small payments.

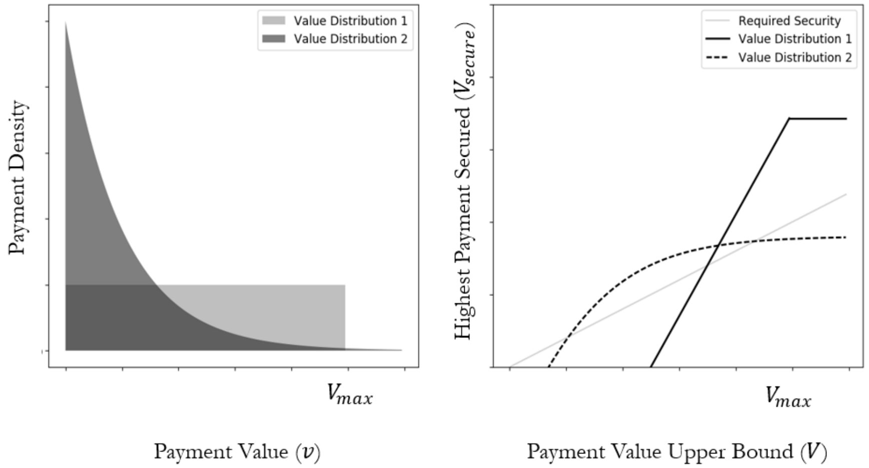


Figure 25 *The left figure shows two payment value distributions uniform ($U[0,V_{max}]$) and exponential ($\lambda e^{-\lambda v}$). The right figure plots highest payment secured when the demand is made of payments in $[0,V]$ following the two distributions. In the region where highest payment secured is less that required security, at least some of the large payments in $[0,V]$ are insecure.*

Beside the payment value distribution, we also make an assumption on the functional form of the security cost. In Section 3.1 we assume only transactions with value $v \leq V_{secure}$ are safe from a double-spend attack when miner revenue is R. We go on to endogenize this upper bound V_{secure} in Section 4.1. Throughout the analysis we assume a binary security cost i.e. payments above V_{secure} are certain to be double spent and payments below V_{secure} are certain to be safe. This happens if users have knowledge of specific adversary with power θ. In practice, users may only have beliefs of adversary power i.e. a distribution over possible adversary power $P(\theta)$. The probability of a payment v double spent will be,

$$\int_0^1 \mathbb{1}\left[v > R\frac{(1-\theta)^2}{1-(1-\theta)^2}\right] \times P(\theta)d\theta, \tag{69}$$

where $\mathbb{1}[\cdot]$ is an indicator variable that takes value 1 when the condition inside it is true, and value 0 otherwise. Thus the security cost takes a continuous form instead of binary,

$$S(v) = \int_0^1 v \times \mathbb{1}\left[v > R\frac{(1-\theta)^2}{1-(1-\theta)^2}\right] \times P(\theta)d\theta \tag{70}$$

We defined $\bar{f}$ as the maximum on chain fees a user is willing to offer in order to avoid off chain proportional rate i.e.f where $U_{off-chain} = U_{on-chain-included}$.

$$\bar{f} = \alpha_h \rho v - S(v)/\delta \tag{71}$$

Under a binary security cost $\bar{f}$ is monotonically increasing for $v \in [0,V]$ i.e. users with large value payment stand to gain most by avoiding the off chain channel. This results is an outcome where Blockchain is used by largest value payments that crowd out smaller payments. $\bar{f}$ is monotonically increasing for a security cost distributions if,

$$\frac{\partial S(v)}{\partial v} \leq \delta \alpha_h \rho \;; \quad \forall \quad v \in [0, V] \tag{72}$$

This is trivially satisfied for a binary distribution, but may not be true for all security cost distributions. If not satisfied, a middle tier of values will have the greatest willingness to pay for payment via Blockchain, instead of the highest tier of values.

Figure 26 shows three possible security cost distributions - binary distribution ($S^1(v)$) similar to one considered in the main text, and two continuous distributions ($S^2(v), S^3(v)$). All these security cost distributions take a monotonically increasing form since larger payments offer greater double spend incentive for any adversary θ. Distributions $S^1(v)$ and $S^2(v)$ satisfy (72) because the largest value $v = V$ still has the highest willingness to pay after accounting for the security cost. This can be interpreted from the figure as the gap between the (fee savings - delay cost) and the security cost. Distributions $S^3(v)$ does not satisfy (72), in fact a different payment value in the middle has the largest gap or the largest willingness to offer Blockchain fee ((fee saving - delay cost) - security cost). Such a security cost distribution would change our analysis such that a middle tier of values will get preferential treatment for payment via Blockchain.

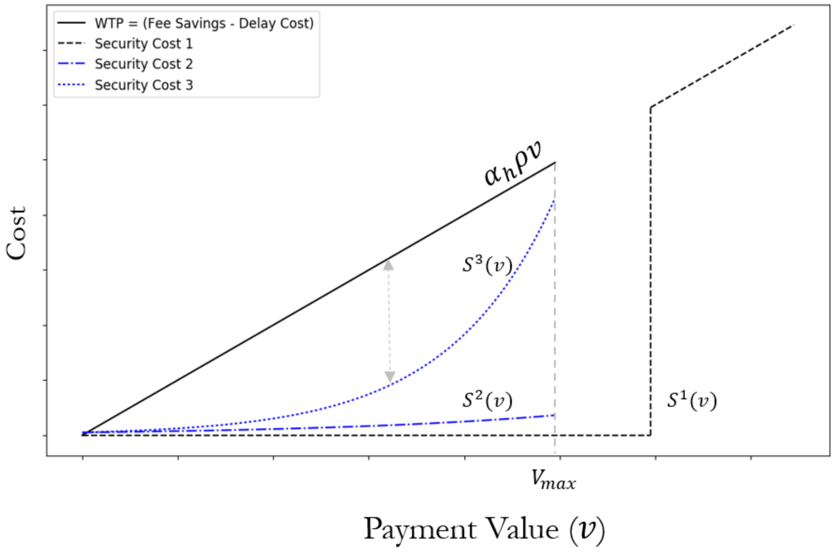


Figure 26 *The willingness to pay on chain incorporates fee savings and delay. This is compared against security cost $S(v)$ with three different formulations. $S^1(v)$ is the binary formulation used in the main text. $S^2(v)$ and $S^3(v)$ are two alternative formulations.*

Alternative models for payment value distribution or security cost distribution do not change the primary insight. Users are heterogeneous (off chain fee, delay, security risk) in willingness to pay. A group of strategic miner can benefit from creating artificial capacity constraints to keep out low willingness to pay users. A re-design to weaken the miner, reduces the security and thus all users willingness to pay. The double edged sword of collusion and security threat keeps the Blockchain accessible to only a fraction of the overall payment demand.

E. Users Costs and Block Reward

The total cost borne by Bitcoin users can be categorized into the following three components.

On Chain Transaction Fee $= \int_{v_0^*}^{V} f_0^* \frac{N}{V} dv = n_F f_0^*$

Off Chain Proportional Fee $= \int_0^{v_0^*} (\rho v) \frac{N}{V} dv$

Delay Cost $= \int_{v_0^*}^{V} (1-\delta) v \frac{N}{V} dv$

Figure 27 shows block size n_F varied between 0-100% of demand N as a protocol design choice. This changes equilibrium outcomes v_0^* and f_0^*. The left figure shows three components of total user costs – Transaction Fee, Off Chain Proportional Fee and Delay dis-utility. The right figure shows all three components combined into total costs with Blockchain, compared against total cost in a world without Blockchain option. The combined cost always dominates a setting without Blockchain i.e. more payment channel options are strictly better for users. Figure 28 shows block duration d on x-axis varies between 0.5 to 1.5 times the current duration $d = 10$ minutes. We assume that an increase in duration d, linearly decreases demand N in every Block period and linearly decreases the discount factor δ. A cursory analysis indicates that user surplus is greatest at very high block size n_F or very low block duration d, but we caveat this with collusion and security issues to be discussed in later section.

Note that we model users choices of channel determined by fees, delay and security. We made two implicit assumptions in doing so - (i) Users do not have a need for exchanging native Blockchain crypto currency for fiat currency. In our model, all users draw periodic payment needs from the same distribution. We assume that all users carry sufficient stock of crypto currency to send and receive payments without needing to exchange from fiat currency. Since all users are homogeneous in their long term payment needs, all of them hold the same average stock of crypto currency. (ii) Users do not face any inflation in crypto currency prices relative to fiat currency on their coin stock. In practice, coinbase Block rewards B distributes additional currency supply to miners. Given constant demand for payments, an increasing supply of coins leads to inflation. Effectively all users pay a combined B in rewards to miners every period. This would be a fourth cost component beside fees, delay and security.This reward is proportional to gross activity or total value of payments on the platform,

$$B = k \int_{v_0^*}^{V} v \frac{N}{V} dv = \frac{kN}{2V}(V^2 - {v^*}_0^2) \tag{73}$$

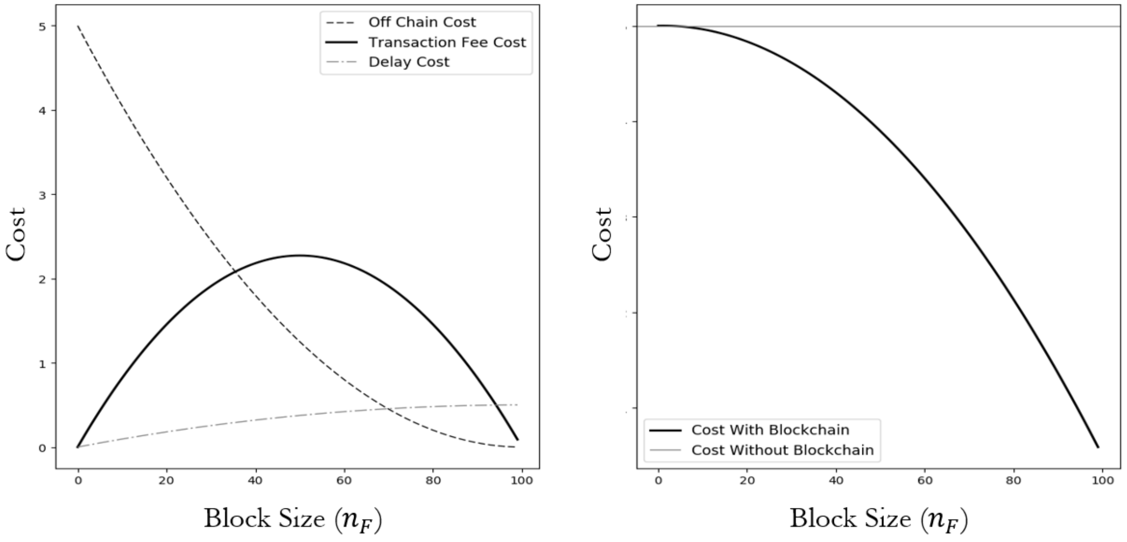


Figure 27 *The left figure shows three components of total user costs against block size n_F. The right figure shows the total cost (sum of three components), compared against payment costs in a world without Blockchain option.*

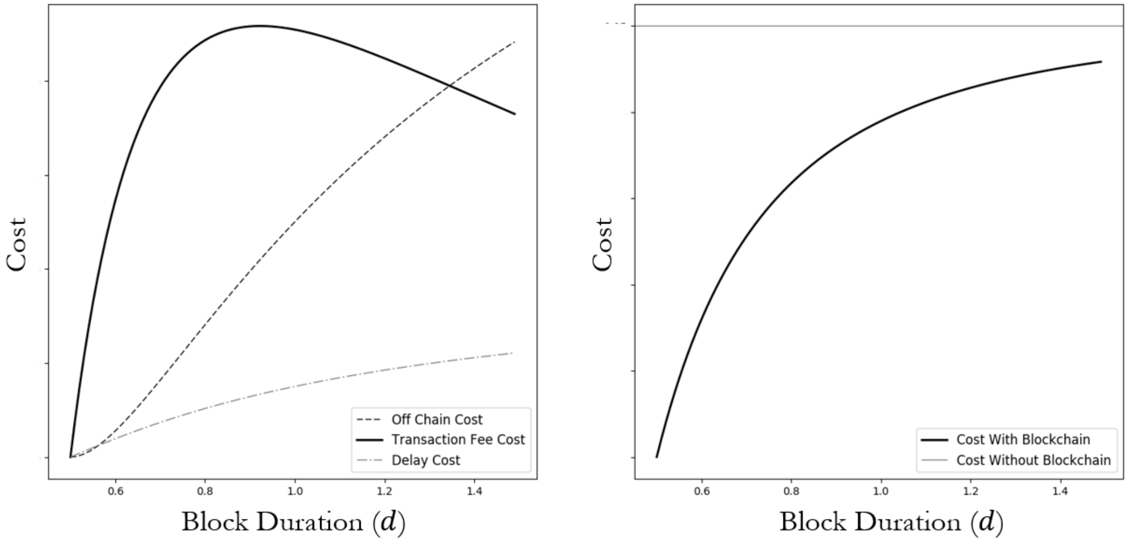


Figure 28 *The left figure shows three components of total user costs against block duration d. The right figure shows three components combined, compared against payment costs in a world without Blockchain option.*

The proportionality constant k captures markets inherent currency velocity i.e. number of payments where the same coin is used and inflation level or block reward level set on the protocol.

In our model, since all users are homogeneous in their long term payment needs, all of them hold the same average stock of crypto currency and thus face the same block reward inflation cost. Individual user is paying B/N in inflation cost to miners every period in addition to transaction fees. Users pays this per period average cost of holding inflationary crypto currency even if they happen to use the fiat channel for payment in a given period. Our primary model found that a setting with additional payment channel option is strictly better for users compared to a no Blockchain world. This is not necessarily the case anymore where we account for Block Reward Inflation cost in addition to fees, delay and security costs. Figure 29 is similar to an earlier Figure 27, albeit with this additional fourth cost component. A rational user will keep stock of coins for regular payments via Blockchain channel if these combined four cost components are strictly better for users compared to a no Blockchain world.

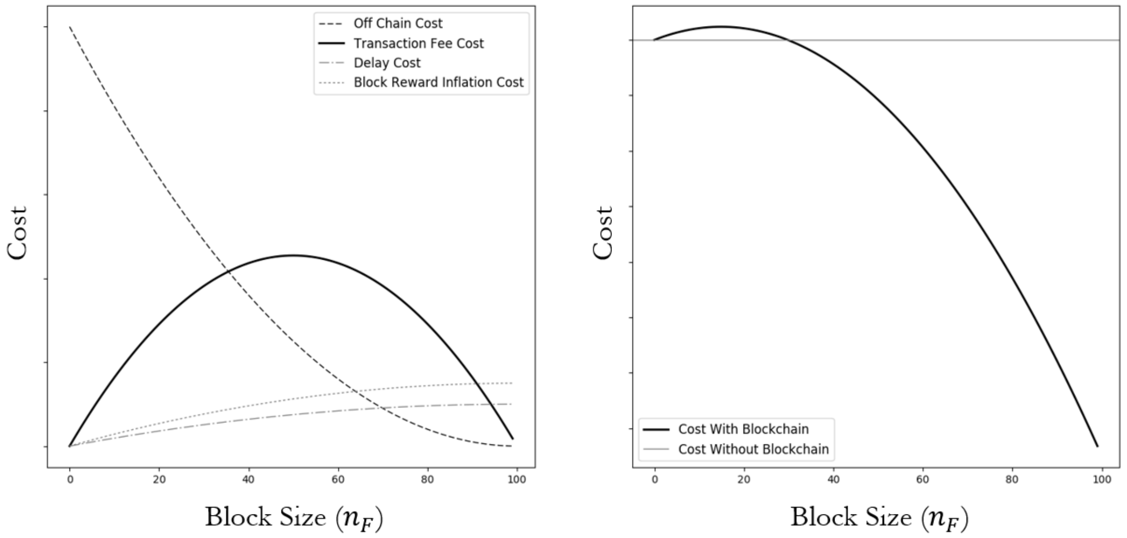


Figure 29 *The left figure shows four components of total user costs against block size n_F. The right figure shows four components combined, compared against payment costs in a world without Blockchain option.*

This leads to an additional constraint,

$$Cost_{\text{on chain fees}} + Cost_{\text{off chain fees}} + Cost_{\text{delay}} + Cost_{\text{block reward inflation}} \quad > \quad Cost_{\text{off chain only}} \tag{74}$$

$$\int_{v_0^*}^{V} f_0^* dv + \int_0^{v_0^*} \frac{\rho v}{V} dv + \int_{v_0^*}^{V} \frac{(1-\delta)v}{V} dv + \int_{v_0^*}^{V} \frac{kv}{V} dv \quad > \quad \int_0^{V} \frac{\rho v}{V} dv \tag{75}$$

Non zero Block Rewards $(k > 0, B > 0)$ may have total costs greater than a no Blockchain world. This happens at very low Block capacity n_F because, users are required to keep a stock of periodically inflating native coins for infrequent payment needs. Figure 30 shows user costs and miner revenues as block reward inflation levels are varied on the protocol between $k = 0\%, 1\%, 2\%$. Miner revenues increase with higher Block Reward. A group of colluding miner now prefer a block size higher than before $(n_P > N/2)$ where they earn larger revenues thus raising security. The overall impact for users is mixed - negative impact of additional cost component but positive impact of greater collusion capacity and greater security.

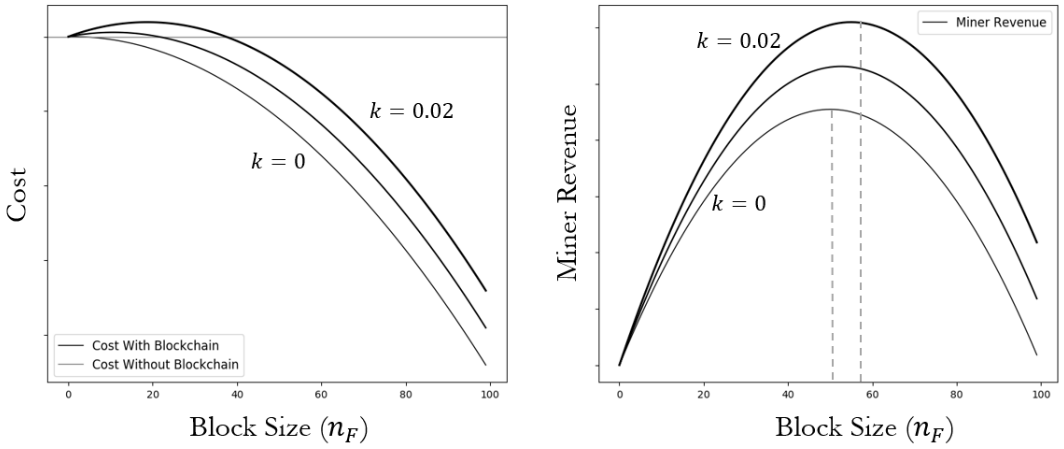


Figure 30 *The left figure shows four components combined against total user costs in a no Blockchain world. As Block rewards increase ($k = 0\%, 1\%, 2\%$), the overall cost for users are higher at the same block capacity n_F. The right figure shows total miner revenues (auction fees and block rewards). As Block rewards increase ($k = 0\%, 1\%, 2\%$), colluding miner will prefer an increasing partial fill level. With increasing Block rewards users face greater costs but less severe collusion prospects.*

While it is not the focus of our work, future researcher can venture further in direction to find optimal design balance between Block rewards and capacity. In fact its possible that very high Block Rewards and near zero transaction fee (n_F or $n_P < N, f = 0$) are overall more optimal in certain market parameter ranges (e.g. N, V, ρ). In-fact such a design is proposed by Chiu and Koeppl 2017. Transaction fee in our model are determined by competitive auction i.e. entry of a large value user increases bids by all other users. Chiu and Koeppl 2017 set transaction as an exogenous proportional rate i.e. entry of a large value user only adds a small fee paid by that individual user. Under this assumption they show that high Block Rewards and near zero transaction fee (n_F or $n_P < N, f = 0$) dominate. Naturally, such a design will need a different mechanism to prioritize payments without auctions. Since Block capacities are limited and miner endogenously prioritize payments via the auction mechanism in most existing p2p Blockchain designs, we would argue that our work caters more to current market conditions and popular design paradigms.

F. Readings on Bitcoin

Bitcoin's ecosystem is composed of four major components: (1) users, (2) miners, (3) the platform protocol, and (4) the cryptocurrency. We suggest Huberman et al. (2019) for a deeper understanding of user waiting queues and transaction fee decisions. Cong et al. (2018) provides an in-depth discussion of the arms race by miners for computing hardware and their organization into mining pools. The protocol itself is most accurately described by Satoshi Nakamoto – the creator of Bitcoin (Nakamoto 2008). Finally, Cong et al. (2018) can be used as a resource for a primer on cryptocurrencies and their adoption.

Table 6 *Major P2P payment Blockchains and their design choices.*

Blockchain	Cap(Bn$)	Proof	Block Reward	Daily Fee	Capacity
Bitcoin	202.8	PoW	12.5 coin (4 years halving)	$545,000	1MB per 10 min
Bitcoin Cash	6.2	PoW	12.5 coin (4 years halving)	$185	8MB per 10 min
Bitcoin Gold	0.5	PoW	12.5 coin (4 years halving)	$908	1MB per 10 min
LiteCoin	6.3	PoW	25 coin (4 years halving)	$1,700	1MB per 2.5 min
Ethereum	28.8	PoW	2 coin (Variable)	$70,000	20KB per 10 sec
Ethereum Classic	0.73	PoW	4 coin (Variable)	$77	20KB per 10 sec
EOS	4.3	PoS	1% Inflation per year	Stake	1MB per 0.1 sec
Steem	0.09	PoS	0.95% Inflation per year	Stake	65KB per 3 sec

G. Generalizability to Blockchains beyond Bitcoin

Nakamoto 2008 proposed Bitcoin as a "peer to peer electronic cash that would allow payments without going through financial intermediaries". This requires a ledger of payment transactions without a trusted third party responsible for its integrity. Bitcoin attempts to reach this goal by making the ledger public, allowing anyone to add to the ledger in return for committing a resource and designing a resource fairly distributed in small quantities, yet prohibitively expensive to accumulate in large proportions for one individual. Bitcoin executes this paradigm by making following design choices – (i) Limit on ledger additions size and frequency to allow sufficient time for dissemination (sync-ing) over peer network. (ii) Rewards for validating large number of peer payments when making ledger entries. (iii) Compute power as commitment proving resource for fair and distributed opportunity to make ledger entries. Beside Bitcoin, this broad paradigm is followed by numerous public Blockchains, albeit with variations in design choices.

We consider below each of these design choices to discuss – alternative choices beside those taken by Bitcoin, whether we model the choice space or keeps it fixed to Bitcoin's choice. If fixed, are our findings robust to relaxed assumptions on the design choice.

Block size and block duration: Block size directly changes capacity and it is central to our paper already. Our previous version keeps Block duration fixed. We implicitly assumed that halving the duration between blocks is effectively equivalent to doubling the block size. This happens because both scenario's will result in the same transaction capacity (per unit time), fees, miner revenues and costs (per unit time). We make use of gamma ($\gamma = n_F/N$) across our derivations to underscore that most of our results effectively depend on the ratio of capacity to demand (per unit time). We now recognize that a halved block duration does provide users greater utility via faster payments. This outcome is not replicated by doubling the size. Now we model Block duration choice explicitly. In so far as 10 minutes discount factors are close to 1 (e.g., $\delta \sim$ 0.99 or 0.999), this update does not have a significant impact but nevertheless offers completeness.

Miner Revenues: There are two distinct choices for rewarding miners – taxing transactions and taxing coin holders. In the first case, transactions are typically taxed by forcing users to bid in a competitive auction for limited throughput. Blockchains (Table 6) differ on current levels of transaction fee rewards due to different degree of user competition and throughput limits. Nevertheless, they all follow similar auction models. Note that in some cases, the user may stake coins or stake compute resources or burn coins as an auction bid. Naturally, this endogenous auction mechanism is central to our paper already.

Alternatively, coin holders can be taxed by giving miners coins as block reward. This increases the total supply of coins and thus causes inflation, if demand is held constant. Effectively, users who hold coins transfer a portion of their real value (inflation adjusted) to miners. We have added an extension in Appendix [D] to show robustness of collusion results even with Block reward. In summary, colluding miners face two opposite forces when deciding optimal partial fill level - increasing transaction fee bids by creating fierce auction competition and decreasing demand for coin holding due to fewer payments facilitated on the platform. Our main text assumes away the second effect, but an extension in Appendix E incorporates both effects. This has a moderating effect on miner collusion, they may not want to restrict capacity as low as $N/2$. Nevertheless some degree of partial filling will remain as long as a chunk of revenue comes from transaction fee auction.

Contemporary research focused on crypto currency valuation is equally important to build a complete picture. For example Chiu and Koeppl 2017 make the opposite modeling choice i.e. abstracting away the transaction fee model, instead delving deeper into endogenous user demand for coin holdings. We contrast our findings with theirs.

We make a conscious choice of not modeling demand for coin holding in our primary model, instead we assume that coin holdings remain static and large enough to allow users to complete random periodic payment needs ($v \sim U[0, V_{max}]$). We do so because – (i) most Blockchains have a diminishing block reward schedule eventually expected to go down to zero anyways. A perpetual block reward design, which leads to infinite supply, does not have a major practical example yet. (ii) Transaction fee are unavoidable even if a Blockchain had perpetual block reward, because they provide a secondary benefit of prioritizing payments. It avoids flooding the Blockchain with 1 cent payments. Fee auctions for limited capacity and miner incentive to collude on artificially lower capacity are therefore unavoidable even with Block reward. (iii) Our primary focus is user choice for means of payment not store of value. Therefore we want to abstract away from myriad issues related to crypto-currency valuation e.g., investors, speculators, fiat currency to crypto currency exchange rates etc.

Consensus mechanism: The consensus mechanism is meant to prove commitment of a fairly distributed resource. Proof of Work (PoW) Blockchains use computing power as this resource. These blockchains differ on cryptographic hash puzzle e.g., SHA256 (Bitcoin and Bitcoin Cash), Equihash (Bitcioin Gold), Scrypt (Litecoin) and Ethash (Ethereum and Ethereum Classic). The primary motivation for these alternatives is to make cryptographic hash puzzle solving resistant to Application Specific Integrated Circuits (ASIC), thus changing distribution of committed resource $(h(m), c(m))$. We indirectly consider alternative choices for hash puzzles when we discuss choices of compute power distribution $(h(m), c(m))$ that would make miner collusion in-feasible. Our results are therefore applicable over alternative PoW hashing algorithm design choice.

The second most popular consensus mechanism is Proof of Stake (PoS) which require miners to show commitment by staking coins instead of compute power [19]. We do not incorporate PoS as a choice in our model. However as long as PoS chains have the same auction mechanism to raise transaction fees, miners would collude on a smaller artificial capacity. Our double spend model assumes that adversarial miner faces no cost from loss of community confidence after attack. This is reasonable if wider community views double spend attack as a rational action by large miner on an ill-advised large payment rather than a threat to normal payments. Under this assumption, adversarial miner does not need to accounts for loss in future value of their mining equipment (if on PoW) or coin stake (if on PoS) when launching the attack. Thus PoS systems will have same incentives for double spend attacks i.e. honest mining revenues versus double spend value.

Naturally this is only a broad inference about PoS systems. We do not rigorously incorporate PoS into our model because - (i) practical examples of PoS Blockchains are still rare. (ii) Blockchains that have considered it, usually rely on PoW to cold start their mining ecosystem anyways. At outset, compute power commitment still remains the only reasonably fairly distributed resource. (iii) PoS has decentralization concerns (a rich get richer outcome) because it helps large coin holders to earn even more revenue. As research and applications of PoS mature we indeed expect this to be an interesting direction of future research.

The third most popular consensus mechanism is Practical Byzantine Fault Tolerance (pBFT). This consensus does not require a costly commitment from miners as it expects miner entry to be centrally permissioned. This is used in enterprise or permissioned Blockchains. These chains may have miner revenues that are entirely outside the Blockchain ecosystem. User competition for space may be obviated since extremely large storage space can easily be synced between a handful of mining nodes. In fact these chain do not intend to attain the same goal of “peer to peer electronic cash for payments without intermediaries”. As enterprise adoption of these systems increase, future researchers could study economics of permissioned Blockchains as a contrast to distributed information systems (or databases), instead of P2P currency-payment mode.

H. Application beyond Blockchains and contribution to IO Theory

We first highlight differences in our model from standard collusion in repeated cournot markets. We use these differences to predict three examples of markets where our models and findings will be applicable. Finally, we summarize the key drivers in the Blockchain setting that form the basis for our model and findings.

[19] Proof of Burn (PoB) typically have similar miner commitment as PoS but may require miners to burn coin permanently instead of staking coins temporarily or they involve users to burn coin instead of direct fees or block reward

Differences from collusion in repeated cournot markets: As pointed out by a few review comments, tacit collusion on quantity among firms using punishment strategies over repeated interactions is well established. The first part of our paper uses a similar idea to achieve higher price for recording transactions. There are some key differences compared to standard cournot repeated game setting – (i) Firm's quantity decisions are not simultaneous, instead separated in time. The standard cournot collusion places a bound on discount factor. Our setting places a bound on colluding firm size because firm's effective discount factor depends on size. This happens in our context because firms have opportunity to set quantity for the entire market at a frequency proportional to their size. (ii) In standard cournot setting, arbitrarily small homogeneous firms could collude. Firm heterogeneity plays a distinct role in our setting as greater heterogeneity allows handful of top firms to constitute large enough market power and sustain collusion even with zero barrier entry of new firms. This happens because new entrants always have inefficient cost structure compared to incumbents (iii) In standard cournot setting, all users irrespective of private valuations observe the same quantity and price in the market. In our setting, user bid first with rational expectations of markets output quantity levels. While users have ability to effect firms's collusive behavior being the first mover, it also leads to user discrimination. Users with different valuations bid differently for a homogeneous product. While colluding firms must only supply to high valuation users, non colluding firms can supply to both high and low valuation users.

Application beyond Blockchains - Firm Supply Decision: We expect such collusion to be more broadly applicable with decentralized markets where services, that are otherwise provided centrally, would be provided by individual crowd participants chosen randomly. Some examples being – fx rate reporting on decentralized currency exchanges, physical event reporting on decentralized betting markets, etc. These setting are uniquely different from oligopolies (e.g., airplane manufacturer) or inter-mediated marketplaces (e.g., Uber, Airbnb) where supply is a summation of quantities offered by individual sellers. In decentralized markets supply is not a summation of quantities. Instead it is agreement of individual service providers or even unilateral decision of a single randomly chosen service provider.

Application beyond Blockchains - Consensus Capacity: In the past decade inter-mediated marketplace platforms (e.g., Airbnb, Uber, Amazon) have replaced traditional large sellers. These marketplaces provide low barrier entry to small firms (e.g., home owners, cabs, small sellers) thus significantly reducing costs. In doing so, the platform intermediary gains significant power to extract economic rents. Blockchains "dis-intermediate" this concentrated power, while keeping the low costs arising from low barrier entry to firms (miners). Unfortunately, the low costs benefits are watered down due to high cost of consensus without an intermediary. This trade-off will become critical with the emergence of Decentralized Apps (Dapps) or Smart Contract that go beyond payment, and try to disintermediate ride sharing, house rentals, lending, insurance, etc. Our paper form basis for future research in dis-intermediated marketplaces. As long as consensus capacity remain limited, participants with high willingness to pay will continue to crowd out smaller users on such dis-intermediated marketplaces.

Application beyond Blockchains - Investment Spillovers: Decentralized marketplaces (e.g., Airbnb, Uber, Amazon) differentiate product quality using history of ratings and reviews. It is interesting to note that perfectly homogeneous product (a payment transaction), still has a notion of quality i.e. security (or finality) of payment. However, quality (security) offered by all firms (miners) is at the same level. No individual firm has incentive to invest in quality because of spill overs i.e. it raises not only their own quality but quality of all competitors. Therefore collusion among firms is a necessary mechanism for firms to coordinate investment in quality of payments. Collusion raises additional revenues, a chunk of these revenues go toward adding more mining equipment by new miners. This additional investment into mining costs raises security of payments. Role of intermediary (e.g., Airbnb, Uber, Amazon) in marketplace platforms is critical when investment in product has strong spillovers (Hagiu and Wright 2015). The same role is played by tacit collusion in dis inter-mediated Blockchain setting. Thus collusion in our setting is not just a mechanism for firms to extract economic rents, rather to invest back in product security (quality).

Summary of key drivers: Keeping specific details of Bitcoin aside, we believe following economic incentives are key drivers for our model - (i) Product supply is not a summation of quantities simultaneously produced by competing firms, rather decided by a randomly chosen participant firm. (ii) Eliminating economic power of intermediary comes at a high cost of consensus. (iii) A lack of unilateral incentive to invest in product quality (security) because of high spillover effects of investment. (iv) Firms with greatest market power, also constitute greatest threat to market stability/security.

Considering the key drivers stated above one by one, we expect broader applicability of our model to - (i) Decentralized services like fx rate reporting, event reporting to betting markets. These are unique settings where supply in market is not a summation of quantities simultaneously produced by competing (or colluding) firms, rather agreement among them. (ii) Decentralized Apps (DApps) or Smart contracts that attempt to dis-intermediate currently inter-mediated marketplaces (e.g., Uber, Airbnb, Lending). These future DApps face high information consensus costs similar to our Bitcoin model. (iii) Markets or products where investment in quality or security has strong spillovers. (iv) Markets with anonymity guarantees where market power does not necessarily result in long term reputation value or legal scrutiny (e.g., Dark web markets, money laundering or mixing services)

I. Blockchain Economics Literature

The table below summarizes recent Blockchain literature with connections to our work. We have discussed individual items in detail in response to specific review comments and the literature review section.

Domain	**Paper**	**Major similarities and differences**
User Transaction Fee Models	Huberman et al. 2019	Same as us, users with heterogeneous willingness to pay compete for Block capacity. But neither users nor miners are rational or strategic about block fill levels and security.
	Easley et al. 2019	
	Hinzen et al. 2019	Unlike us, endogenizes block duration as a function of mining community size.
Miner Entry Models	Arnosti and Weinberg 2018	One argues that cost efficient miners will keep raising their hash power leading to high degree of heterogeneity among miners. Other argues for high degree of homogeneity. We keep degree of heterogeneity as a flexible design choice.
	Prat and Walter 2018	
Collusion	Cong and He 2019	Same as us, they use transparency of Blockchain payment to sustain collusion. But unlike us, they look at collusion among firms that use Blockchain for payments instead of miners/gatekeepers of the Blockchain itself.
	Malinova and Park 2017	
Security Models	Chiu and Koeppl 2017	Same as us, they model attackers tradeoff between honest revenues and double spent payment value. Unlike us, they simplify the transaction fee revenue model and instead focus on block reward miner revenue. Our and their paper can be seen in conjunction for a complete picture.
	Budish 2018	They focus on designs to weaken double spend attack.
	Gervais et al. 2016	They run simulations double spend attacks under variety of design choices.
	Carlsten et al. 2016	They argue for block rewards instead of transaction fee revenues to weakend double spend attack.
Trilemma	Abadi and Brunnermeier 2018	Same as us, they discuss tradeoffs between scale, security and decentralization. Unlike us, they claim very different mechanisms to argue the tradeoffs.
	NeonVest 2018	
Forks in Equilibrium	Biais et al. 2018	Some of our security formulations can be fine tuned by incorporating equilibria other than longest chain or more sophisticated attack and defense strategies. We do not think it changes the underlying trade-off for an adversary - honest mining revenue vs double spend payment value.
	Kroll et al. 2013	
Attack Strategies	Eyal et al. 2016	
	Sompolinsky and Zohar 2015b	
	Nayak et al. 2016b	

Chapter 3:
Does Machine Learning Amplify Pricing Errors in Housing Market? : Economics of ML Feedback Loops

Abstract

ML pricing models (Zillow's Zestimate, Redfin Estimate) have been deployed in the last decade to make house price predictions. These ML models are revised regularly using recent sample of sales. The recent sales are themselves confounded by previous version of the ML model. We theoretically show how this Feedback Loop creates a self fulfilling prophecy where ML over estimates its own prediction accuracy and market participants over rely on ML predictions. We formulate size of resulting pricing bias. We identify conditions on ML and market characteristics such that participants are worse off after introduction of ML. We use data from Zillow's Zestimate for empirical evidence for necessary primitives of our theoretical model.

1. Introduction

Buyers and Sellers in the housing market need to accurately determine price of a house. Zillow's Zestimate and Redfin's Estimate are at least two examples of Machine Learning (ML) algorithms meant to assist market participants. Zillow (Redfin) state a median error of 1.9% (1.7%) between ML price of an on-market house and its eventual sale price [12, 17]. Both these platforms caution about price errors, and suggest that participants rely on professional appraisals or real estate agents. In spite of the risk, ML price has some obvious attractions – (i) costless accessibility, unlike a professional appraisal, (ii) sense of impartiality, unlike agents[1] and (iii) low error rates (1.7-1.9%) reported by the platform. Empirically we find an average reliance of 15% on ML price i.e., if ML price of a $200,000 house was displayed as $220,000 (+10% error) then the expected sale price of house is $203,000 (+0.15*10% error). Across submarkets this reliance varies from 10% to as high as 50%.

A market participant may use multiple signals to construct a private valuation for a house. These private valuations are noisy and are likely to create disagreement among participants. Participants need to spend time in the market to resolve these disagreement. For example, a

[1] Prior housing market literature provides evidence of mis-aligned pricinpal-agent incentives.

seller enters the market over-optimistic about how buyers will value a backyard swimming pool. The seller observes low buyer offers and corrects their valuation. Thus valuations disagreement, while costly, are corrected as long as the these private valuations are unbiased and uncorrelated across the participants. Introduction of ML price plays a unique role, in that it is a neutral and pervasive[2] signal to all buyers and sellers. If accurate, it is powerful in alleviating costly price disagreement among participants. If inaccurate, participants can not learn and correct since all other participants hold the same erratic signal. Ideally, an oracle social planner could tune a dial to set market's degree of reliance on ML price based on perfect knowledge of ML accuracy. In practice, we show that the calculated ML accuracy and market's reliance are intertwined in a Feedback Loop that doesn't converge to the socially optimal level.

Feedback Loop Definition: Buyers and Sellers in the market interact to complete sale of a house with feature X at price p. The house sale process can be summarized as (X, p) sampled from a probability density function $f_{\Phi}(X, p)$ parameterized by market structure Φ. Some of the elements that make up the market structure include - Hedonic value of floor area or neighborhood, sellers uncertainty about value of their house, variance among buyer offers, rental income options and agent costs. The ML pricing model $f_{\omega}(X; \theta)$ is retrained every period in order to minimize expected prediction loss. Here θ are model parameters and ω are model design hyperparameters. The expected loss is approximated using observed sale samples of (X, p). For example, our reverse engineering of Zillow's algorithm shows that a single house X that sells at $p =$ \$210,000 (say 5% over its actual value of \$200,000) shifts ML model θ such that ML price Zestimate $z = f_{\omega}(X; \theta)$ for 10-500 (25 on average) peer houses moves up by 1-1.25%.

$$\theta_t = \underset{\theta}{argmin} \underset{(X,p) \sim f_{\Phi(\theta_{t-1})}}{E} \left[loss\big(p, f_{\omega}(X; \theta)\big) \right] \tag{1}$$

Some elements of the market structure Φ are sensitive to ML model θ_{t-1}, and ML model θ_t naturally depends on samples drawn from f_{Φ}. In essence, ML model θ is chasing a distribution

[2] ML price is not the only such pervasive signal. Global events and expert opinions can have the same effect.

shift imposed by its own previous version. We call this repeating process where model training samples are confounded with models own past version as the Feedback Loop. This formulation[3] of Feedback Loop ($\theta \leftrightarrow f_\Phi$) is a general description for any online ML platform. We focus on feedback between two key elements ($\sigma_z \leftrightarrow \alpha$) – (i) ML accuracy σ_z which is part of the ML model θ and (ii) participants reliance on ML α which is part of the market structure f_Φ. In a "self fulfilling prophecy" houses sell closer to the ML price, than they would if ML price was hidden. Thus ML accuracy σ_z is increasingly underestimated with high reliance on ML α. Market reliance on ML α is reinforced by the over estimated accuracy σ_z.

Rigorous empirical evidence of Feedback Loop is hard to establish because Zillow's proprietary algorithm is opaque. In this paper, we show anecdotal comparison of Zillow's stated ML confidence interval compared against actual observed ML errors for a large sample of houses across three US counties. We note that Zillow made a conservative upward adjustment to its confidence interval in July-August 2019 while the actual error rate has been steady or declining. This would suggest that stated average confidence intervals in first half of 2019 of $\leq \pm 9\%$ may have be over-optimistic compared to actual error rates of 13% or more.

We identify conditions on ML and market characteristics such that introduction of ML makes participant expected payoff worse off. ML model, capable of high explanatory power in a hypothetical unconfounded setting, can surprisingly make participants worse off under the Feedback Loop. This happens as ML error estimate collapses to zero ($\sigma_z \rightarrow 0$) and reliance on ML goes to one ($\alpha \rightarrow 1$). A low complexity ML model works better under the confounded Feedback Loop. A "self-sufficient" market can resolve disagreement in participant private valuations, but it is more likely to be worse off after introduction of ML. We also provide a feedback correction procedure by setting aside ρ proportion of houses where ML signal is hidden. But the solution comes with the obvious caveat that the ML platform may not have the necessary incentive to sacrifice revenue in favor of this correction procedure. We also

[3] Perdomo et al (2020) introduced this formulation to derive convergence properties.

extrapolate from our findings to discuss implications for participants heterogeneous in - ability to accurately price and patience to spend time in the market.

2. Literature Background

ML pricing has a significant impact if buyers and sellers are not able to accurately price houses. Linneman, P. (1986) show that market participants have a large variance of house value estimates because – they transact infrequently (say once in 10 years), and, unlike other assets, housing has large diversity of features. Prior field survey and experimental research has tried to determine size of these valuation errors. Goodman Jr, J. L., & Ittner, J. B. (1992), Kiel, K. A., & Zabel, J. E. (1999) and Ihlanfeldt, K. R., & Martinez-Vazquez, J. (1986) find average absolute valuation errors of 14%, 5.3% and 16% respectively[4]. The heterogeneity comes from different sub markets and whether the individual participated in the market recently.

In theory, agents, brokers or other market experts can correct valuation errors (Han, L., & Strange, W. C. 2015). In practice, sellers and expert agents contract under information asymmetry. Since a seller cannot observe how much effort his agent is putting into selling his house, a Moral Hazard problem arises in the principal (seller) – agent contract (Anglin, P. M., & Arnott, R. 1991). The agent has an incentive to under value the house, and save on search cost. In a competitive market, agents may have incentive to over value the house in order to win the listing against competing agents. Further, Adverse Selection problem makes the seller unable to judge perfectly whether the agent is knowledgeable about the state of the market. Thus agents can not reliably reveal their expertise on house valuation. Beside agents, experts may still be able to identify price bias in a sub market. For example, a 5% overpricing sustained for 12-36 months in 1000-2000 sqft single family homes in Charlestow neighborhood of Boston. Expert may identify such pricing bias by observing summary level demand-supply across comparable neighborhoods or using knowledge of ML Feedback (as described in this paper). We do not model such external monitoring offered by market experts. In fact, Cheng, I. H. et al. (2014)

[4] Some of the literature also measures bias, i.e., different between actual value and expected value among prospective buyers and sellers. We model private valuations as unbiased, in order to assume away this second layer of potential behavioral anomaly

show that securitized home loan managers were themselves unaware of a large-scale housing bubble in 2004-06, preceding the 2007-08 collapse. ML Feedback Loop is a statistically complex phenomenon, similar to risk pricing of securitized home loan assets. It would not be surprising if pricing bias from the ML Feedback Loop similarly remain opaque to experts for an extended period of time.

ML Feedback Loops are not unique to the housing market. For example, Google Maps traffic routing model imposes a change in driver behavior, which are then consumed by Google Map to update its routing model. In fact, feedback labels are useful in a wide range of online learning settings where an ML algorithm learns by making mistakes. While innocuous is some settings, they can slowly accrue as technical debt (Sculley, D. et al. 2015) in large ML platforms posing a significant danger (Amodei, D.et al. 2016). Bottou, L. et al. (2013) discuss ML scoring of Ad placement options using training data from past Ad placements imposed by previous version of the ML model. Wager, S. et al (2014) discuss ML click through rate prediction using click data from past search engine result imposed by previous version of the ML model. They show how to detect the loops by introducing small amount of random noise in search result rankings. Sinha, A. et al. (2016) discuss ML recommender systems that model user-item matrices using training data from past user-item rating feedback provided by users under recommendation sets imposed by previous version of the ML model. They identify sufficient assumptions on the recommender system feedback process to discern intrinsic preferences from the confounded user-item model.

These examples from Ad placement, Search Engine rankings and recommendations focus on settings where individuals interact with ML predictions in isolation. For example, a user doesn’t have an alternative mechanism, beside ML, to interact with a crowd of her peers and determine relevant search results. A crowd of housing market participants, in absence of ML, can interact and determine pricing. Thus introduction of ML in housing market is unique in replacing wisdom of the crowd with a single correlated signal. Literature on implications of ML Feedback Loop on the overall market is limited. Chaney, A. J. et al (2018) is a rare examples that demonstrates homogenization of crowd behavior and loss of utility for at least some subset of users under recommender system Feedback Loop.

Our model of market participants has similarities with Schmit, S., & Riquelme, C. (2018). In a recommendation system setting, they assume that – users are naïve in believing that ML scores are unbiased estimates, and users are myopic and honest about their current action without regard to its impact on future state via the Feedback Loop. We make similar assumptions to rule out potential actions of agents or experts who may take strategic actions based on knowledge of the Feedback Loop and its associated biases. Perdomo et al (2020) identify conditions where repeated empirical risk minimization (RERM) converges to an optimal and stable point. They frame the problem as a unique case of covariate distribution shift, often studied in Computer Science literature on robustness of ML. In short, convergence requires ML loss function to be "nicely behaved" and market actions to be sufficiently insensitive to ML. Since we also model housing market ML pricing as RERM algorithm, we discuss if Perdomo et al (2020)'s convergence conditions are satisfied. However, we are not as concerned with robustness of ML accuracy at this converged state. Instead we want to discern what part of the covariate distribution shift is coming from evolution of intrinsic housing preferences and what part is artificially created by ML Feedback. If the latter starts to dominate, social surplus may be lost even as ML accuracy reaches its optimal.

3. Theoretical Framework

In this section, we formally model the Housing Market, the Machine Learning platform and the Feedback Loop. We identify necessary conditions under which introduction of Machine Learning makes the market participants better off.

3.1 House Value

We define the value $v_{k,t}$ of a house k at time t as the expected sale price $E[p_{k,t}]$ if the house is put up for sale. The realized sale price for a house may deviate from the expected sale price because of heterogeneity among seller and the buyers who visit the house. Individual seller i, who values the house higher (or lower) than actual $v_{i,k,t} > v_{k,t}$, asks a price above (or below) the expected sale price. A house on the market may encounter visits from a set of buyers who value the house higher (or lower) than actual and offer a price above (or below) the expected sale price. Thus, the expected sale price $E[p_{k,t}]$ depends on average valuation across market participants $E[v_{i,k,t}]$. If all participants were to start valuing a house 5% higher, it would increase the expected sale price $E[p_{k,t}]$ and therefore the value $v_{k,t}$.

$$v_{k,t} = E[p_{k,t}] = E[v_{i,k,t}] \tag{2}$$

$$p_{k,t} \; = v_{k,t} + \epsilon_{k,t} \quad ; \quad \epsilon_{kt} \sim N(0, \sigma_\epsilon^2) \tag{3}$$

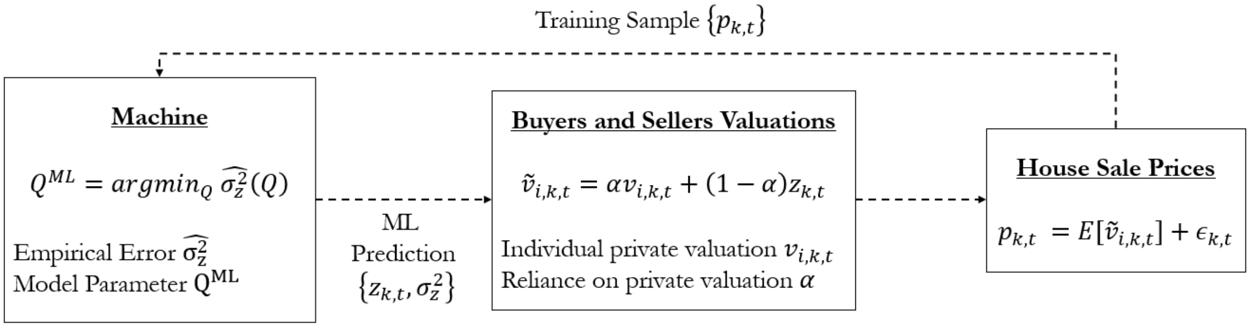

Training Sample $\{p_{k,t}\}$
Machine
$Q^{ML} = argmin_Q \; \widehat{\sigma_Z^2}(Q)$
Empirical Error $\widehat{\sigma_Z^2}$
Model Parameter Q^{ML}
ML Prediction $\{z_{k,t}, \sigma_Z^2\}$
Buyers and Sellers Valuations
$\tilde{v}_{i,k,t} = \alpha v_{i,k,t} + (1-\alpha) z_{k,t}$
Individual private valuation $v_{i,k,t}$
Reliance on private valuation α
House Sale Prices
$p_{k,t} \; = E[\tilde{v}_{i,k,t}] + \epsilon_{k,t}$

Figure 1: Feedback cycle from ML to buyer-seller valuations to house sales and back to ML.

A seller (or buyer) i estimates house value $v_{i,k,t}$ to decide ask (or offer) price for her house. This estimate is made up of – her private valuation of keeping ownership of house and her guess of the expected sale price of the house. Similarly, a buyer estimate is made up of – his private valuation of ownership of house and his guess of the expected sale price of the house. If all buyers had the same private valuation and all participants were rational about this homogeneity, then $e_{i,k,t} = 0$. Subsequently, all houses would sell at their expected sale price $\epsilon_{k,t} = 0$. In practice, participants are heterogenous in private valuations and have noisy guess of the expected sale price.

$$v_{i,k,t} = v_{k,t} + e_{i,k,t} \;\; ; \;\; e_{ikt} \sim N(0, \sigma^2) \tag{4}$$

Individual buyers (seller), due to erroneous valuation $e_{i,k,t}$, offer (asks) above or below the expected sale price. Large errors in value estimates make instances of such erroneous offers (asks) more likely[5]. Thus dispersion in value estimates σ^2 increases dispersion in realized sale prices σ_ϵ^2. Some of the dispersion in realized sale prices may be moderated as market participants learn from each other. For example, a seller with incorrect high estimate will eventually be able to correct her mistake after she encounters buyers with low offers. This correction happens as deviations $e_{i,k,t}$ among participants are constructed independently and therefore uncorrelated. We model a linear relationship $\sigma_\epsilon^2 = \delta\sigma^2$, where δ captures moderation of valuation errors as participants learn during search and bargaining in the market.

The value of a house depends on its unique features – location (neighborhood, zip code, county), size (area, bedrooms, floors) and finer structural details (year of construction, style, flooring, patio, etc.). A house k can be fully described using $\bar{Q}$ dimensional feature set $\overrightarrow{X_k}$. Every feature $x_{q,k}$ takes a binary value of 1 or 0 e.g., whether house has a patio or not[6]. A cumulative valuation house feature $g_t\left(\overrightarrow{X_k}\right)$ make up value of house $v_{k,t}$ at time t. The value of the features and therefore the value of the house evolve as a random walk over time.

$$\overrightarrow{X_k} = \left[x_{1,k}, \dots x_{q,k}, \dots, x_{\bar{Q},k}\right] \quad ; \quad x_{q,k} \in \{0,1\} \tag{5}$$

$$v_{k,t} = v_{k,t-1} + e_{k,t}^{rw} \quad ; \quad e_{k,t}^{rw} \sim N(0, \sigma_{rw}^2) \tag{6}$$

3.2 Machine Learning (ML) Feedback Loop

An ML model predicts house value $z = f_\omega(X;\theta)$, where θ are model parameters (e.g., weights in linear model or boundaries in a decision tree) and ω is model design hyperparameters (e.g., set of features to use or regularization strength). For a given ω, the parameters θ are updated by minimizing in-sample expected loss. The hyperparameters in ω are choosen to minimize out-of-sample expected loss in order to avoid overfitting.

$$\theta_t = \underset{\theta}{argmin} \underset{(X_k, p_{k,t}) \sim f_{\Phi(\theta_{t-1})}}{E} \left[loss\left(p_{k,t}, f_\omega(X_k;\theta)\right)\right] \tag{7}$$

[5] Individual seller has higher or lower patience (or reservation value) than an average seller, thus allowing

[6] We assume that continuous or multi valued features can be represented through binary features. This simplification does not change any of the results.

The expected loss is approximated using a sample of N houses, which is small compared to $2^{\bar{Q}}$ unique sets of house features. Therefore we focus on a class of ML models f_ω that prices J ($\ll 2^{\bar{Q}}$) house clusters, instead of pricing every unique house. The clusters are based on "priced" features Q ($< \bar{Q}$), such that any pair of houses k_1 and k_2 with $x_{q,k_1} = x_{q,k_2}\ \forall\ q \in \{1, \dots, Q\}$ are collapsed into a single cluster. Since ML assigns a single price to all houses in a cluster, effectively the remaining $(\bar{Q} - Q)$ unique features $q \in \{Q + 1, \dots, \bar{Q}\}$ are left "unpriced". Thus model design f_ω is equivalent to choice of hyperparameter Q which governs the model complexity. The model parameters θ are made up of vector of J cluster prices $z_{1\dots J}$. Finally, for clean analytical expressions we choose $loss(p, z) = (p - z)^2$. Under these choices for $(f_\omega, \theta, loss)$, the ML price prediction $z_{k,t}$ is the sample mean of observed sale prices $p_{k',t-1}$ for all houses k' in cluster j_k that were sold at $t - 1$.

$$z_{k,t} = z_{j_k,t} = \frac{\sum_{k'} p_{k',t-1} \times \mathbf{1}(j_{k'} = j_k)}{\sum_{k'} \mathbf{1}(j_{k'} = j_k)} \tag{8}$$

Beside the predictions $z_{k,t}$, a key ML output is an estimate of error in ML prediction $\hat{\sigma}_z^2$. This error comes from three sources – (a) sale prices from last period can not forecast random walk ($e_{k,t}^{rw} = v_{k,t} - v_{k,t-1}$) into the next period, (b) $(\bar{Q} - Q)$ unique features are left unpriced ($v_{k,t-1} - v_{j_k,t-1}$) and (c) small sample mean of observed sale prices ($\bar{v}_{j_k,t-1} - \bar{p}_{j_k,t-1}$) may be different from population mean. We will endogenize ($e_{k,t}^z, \sigma^2(e_{k,t}^z) = \sigma_z^2$) and its empirical approximation ($\hat{e}_{k,t}^z, \sigma^2(\hat{e}_{k,t}^z) = \hat{\sigma}_z^2$) later in the section by substituting analytical expression for $\bar{p}$.

$$e_{k,t}^z = z_{k,t} - v_{k,t} = \underbrace{v_{k,t-1} - v_{k,t}}_{\substack{Random\\Walk}} + \underbrace{v_{j_k,t-1} - v_{k,t-1}}_{\substack{Unpriced\\Features}} + \underbrace{\bar{p}_{j_k,t-1} - \bar{v}_{j_k,t-1}}_{Sample\ Error} \tag{9}$$

A seller (buyer) i constructs the estimate $\tilde{v}_{i,k,t}$ as a combination of own private signal $v_{i,k,t}$ and the ML signal $z_{k,t}$. We assume that individuals have knowledge of noise in their own private signal (σ^2) and the ML model provides empirical estimate of its own noise ($\hat{\sigma}_z^2$). The two signals are weighted based on corresponding degree of noisiness. For example, an accurate ML signal (small σ_z^2) would lead to lower reliance on private valuation.

$$\tilde{v}_{i,k,t} = \alpha v_{i,k,t} + (1-\alpha) z_{k,t} \quad ; \quad \alpha = \frac{\hat{\sigma}_z^2}{\sigma^2 + \hat{\sigma}_z^2} \tag{10}$$

Remember, that the expected sale price $p_{k,t}$ if the house is put up for sale matches up with the mean value estimate of the seller and visiting buyers. The noisiness in individuals own private estimates $e_{i,k,t}$ cancel out because they are constructed independently, but the noisiness in ML signal $e_{k,t}^z$ adds a bias because its correlated across individuals.

$$E\left[p_{k,t}\right] \ = \alpha E\left[v_{i,k,t}\right] + (1-\alpha) E\left[z_{k,t}\right] = v_{k,t} + (1-\alpha) e_{k,t}^z \tag{11}$$

$$p_{k,t} \ = v_{k,t} + (1-\alpha) e_{k,t}^z + \epsilon'_{k,t} \tag{12}$$

In a hypothetical where all buyers and sellers rely entirely on the ML signal ($\alpha = 0$), there would be no dispersion of sale prices $p_{k,t} = E[p_{k,t}]$. All houses would be sold exactly on the ML predicted price i.e. $\epsilon'_{k,t} = 0$. As the individuals rely more on own private independent valuations ($\alpha > 0$), sale prices are more likely to be dispersed i.e. $\epsilon'_{k,t} \neq 0$.

$$\sigma^2\left(\epsilon'_{k,t}\right) = \delta \sigma^2\left(v_{i,k,t}\right) = \delta \alpha^2 \sigma^2 \tag{13}$$

Now we can substitute $p_{k,t}$ in equation (9),

$$e_{k,t}^z = \left(v_{k,t-1} - v_{k,t}\right) + \left(\bar{v}_{j_k,t} - v_{k,t}\right) + \bar{\epsilon'}_{k,t} + (1-\alpha) e_{k,t-1}^z \tag{14}$$

which corresponds to an error variance,

$$\sigma_z^2(Q,\alpha) = E\left[\left(e_{k,t}^z\right)^2\right] = \Lambda(Q,\alpha)/(2\alpha - \alpha^2) \tag{15}$$

where

$$\Lambda(Q,\alpha) = \sigma^2_{rw} + \underbrace{\left(\frac{\bar{Q}-Q}{\bar{Q}}\right)\left(1-\frac{2^Q}{N}\right)\sigma^2(v_{k,t})}_{unpriced\ features} + \underbrace{\frac{2^Q}{N}\delta\alpha^2\sigma^2}_{Sample\ Error} \tag{16}$$

The empirical approximation $\hat{\sigma}^2_z$ is calculated as squared difference of predicted zestimate $z_{j_k,t}$ and sale price $p_{j_k,t}$ averaged over sales observed across T periods.

$$\hat{\sigma}^2_z(Q,\alpha) = \frac{1}{K*T}\Sigma_t\Sigma_{k'}\left(\hat{e}^z_{j_k,t}\right)^2 \cong E\left[\left(\alpha e^z_{k,t}\right)^2\right] = \Lambda(Q,\alpha)*\alpha^2/(2\alpha-\alpha^2) \tag{17}$$

In a hypothetical where ML signal is hidden from the market, the ML signal noise is given by $\sigma^2_z(Q,1) = \Lambda(Q,1)$. Going forward this is refered to as the unconfounded ML signal noise σ^2_z. The correct confounded ML signal noise $(\sigma^2_z)_C$ is higher than the unconfounded ML signal noise σ^2_z by a factor of $(1/(2\alpha-\alpha^2))$. This inflation is the result of autoregressive accumulation of ML signal errors. Since sale price outcomes shift towards ML signal, a "self fulfilling prophecy" leads to an underestimation of ML signal noise $\hat{\sigma}^2_z$ by a factor of α^2.

Proposition 1: If market participants have reliance on ML signal is (1- α) and the unconfounded ML signal noise is $\boldsymbol{\sigma^2_z}$ then the empirically underestimated and the corrected ML signal noises are respectively given by,

$$\boldsymbol{\hat{\sigma}^2_z = \frac{\alpha^2}{2\alpha-\alpha^2}\,\sigma^2_z \quad ; \quad (\sigma^2_z)_C = \frac{1}{2\alpha-\alpha^2}\,\sigma^2_z} \tag{18}$$

The interdependence of $\hat{\sigma}^2_z(Q,\alpha)$ and $\alpha(\hat{\sigma}^2_z)$ form the core of the Feedback Loop. Prior to introduction of ML ($\alpha = 1$), the observations $(X_k, p_{k,t})$ are unconfounded. The ML error estimate $\hat{\sigma}^2_z = \Lambda(Q,1)$ is minimized at $\hat{Q}_0$. The random walk component in Λ can not be modeled and remains static for any choice of Q. The optimal choice $\hat{Q}_0$ finds a balance between the remaining two components. A $Q < \hat{Q}_0$ would increase the sample size $(N2^{-Q})$ and reduce the sample error. However, it lead to a large cluster size, with larger number of unpriced houses features and a large dispersion of values within cluster. A $Q > \hat{Q}_0$ would gain on distinguishing unique houses at the cost of an erratic small sample mean cluster price estimate. Once the ML prediction is available, α reduces below 1. The new ML error estimate $\hat{\sigma}_z^2 = \Lambda(Q, \alpha)$ is minimized at $\hat{Q}_1$. Note that at $\alpha < 1$, the sample error diminishes and the unpriced component starts to dominate Λ. Since the latter is decreasing in Q, we have $\hat{Q}_1 > \hat{Q}_0$. This cycle repeats, resulting in increasing reliance on ML (decreasing α) while the ML gradually overfits (increasing Q). Note that this phenomenon occurs if ML hyperparameter (Q) is iteratively adjusted. For the rest of the section, we assume that this first order problem can be overlooked if the ML design hyperparameters change infrequently. We focus on the more challenging problem of $(\alpha, \hat{\sigma}_z^2)$ convergence under a constant ML design hyperparameter Q.

Consider introduction of ML when individuals originally rely entirely on own valuation ($\alpha = 1$). Let the ML signal noise calculated offline before introduction $\hat{\sigma}_z^2 = \sigma_z^2 = 4\sigma^2$. Substituting $\hat{\sigma}_z^2$ and α in equation (10), individual reliance α on own valuation drops from 1 to 0.8. Now, the ML signal shifts the sale prices, resulting in underestimation of empirical error variance ($\hat{\sigma}_z^2 = 0.66\sigma_z^2$). The lower estimate of ML error further lowers individuals reliance on own valuation ($\alpha = 0.8 \rightarrow 0.72$). This cycle repeats until reliance and ML error converges[7] to $\alpha = 0.66$ and $\hat{\sigma}_z^2 = 0.5\sigma_z^2$ respectively. This feedback cycle is even more acute if actual ML error is low to begin with ($\sigma_z^2 \leq 2\sigma^2$). In this case individuals eventually rely entirely on ML ($\alpha = 1 \rightarrow \cdots \rightarrow 0$)[8].

Proposition 2: Empirical under-estimation of ML signal noise worsens with high reliance on ML. And over-reliance on ML signal grows with diminishing ML signal noise. This feedback cycle repeats until reliance α and ML error estimate $\hat{\boldsymbol{\sigma}}_z^2$ converges to,

$$\left(\boldsymbol{\alpha}, \hat{\boldsymbol{\sigma}}_z^2\right) \xrightarrow{t\rightarrow\infty} \begin{cases} \left(\dfrac{\boldsymbol{\gamma}-\mathbf{2}}{\boldsymbol{\gamma}-\mathbf{1}}, \boldsymbol{\sigma}_z^2(\boldsymbol{\gamma}-\mathbf{2})\right) & \boldsymbol{if\ \gamma > 2} \\ (\mathbf{0},\mathbf{0}) & \boldsymbol{else} \end{cases} \quad ; \quad \boldsymbol{\gamma} = \boldsymbol{\sigma}_z^2/\boldsymbol{\sigma}^2 \tag{19}$$

[7] See Appendix A.1 for discussion on convergence of stable points.

[8] Note that $\hat{\sigma}_z^2$ is calculated over T periods, where 1 period may be a year or longer. In practice, this convergence α → 0 will be thwarted by mitigating factors in the market such as expert observers. We do not model such external factors.

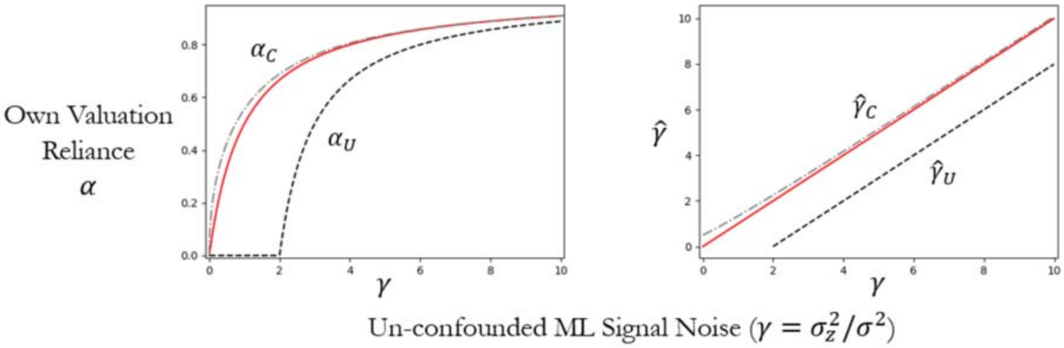


Figure 2: (Left) Reliance (α) for different levels of unconfounded γ ($= \sigma_z^2/\sigma^2$) after convergence. Empirical underestimation results in a low α_U (black dashed line), while a corrected estimation should lead to a high α_C (grey dashed line), relative to an unconfounded setting (red solid line). (Right) ML noise underestimation $\hat{\gamma}_U$ (black dashed line) and corrected estimate $\hat{\gamma}_C$ (grey dashed line), relative to unconfounded γ (red solid line).

3.3 Economic Implications

Using the framework above, now we can examine if ML introduction makes the participants better off. Instead of fully modeling participant payoff[9], we assert that two market characteristics are key drivers behind average payoff Π – (A) error in participants valuations $(v_{i,k,t} - v_{k,t})$ and (B) error in house sale prices $(p_{i,k,t} - v_{k,t})$.

$$\Pi = \Omega_0 - \Omega_1 \underbrace{E\left[\left(v_{i,k,t} - v_{k,t}\right)^2\right]}_{\substack{Valuation \\ Error}} - \Omega_2 \underbrace{E\left[\left(p_{i,k,t} - v_{k,t}\right)^2\right]}_{\substack{Sale\ Price \\ Error}} \tag{20}$$

The component (A) measures error in valuation of a seller (or buyer) at market entry. A seller with rational expectation of error in their valuations acts cautiously. They are less likely to enter the market, they starts with a conservative high list price, and prefer to spend time learning from buyer offers. Similarly, a cautious buyer is less likely to enter the market, starts with a conservative low offers price, and prefers to spend time learning from variety of seller list prices. This buyer and seller behavior under noisy valuation lowers everyone's expected payoff Π.

[9] In *Appendix A.2,we provide an approach to structurally estimate payoff, to replace the simplified metric used here. The full payoff model cis more apt to answer deeper questions on – type of heterogenous participant that are better off after introduction of ML.*

The component (B) measures errors in eventual sale prices. If the participants on an average have unbiased valuations (even if individual participant is erratic) and they can resolve errors in their private valuations via market interaction. However, if valuations have systematic bias or learning in markets is weak then errors in valuations will propagate to errors in sale prices. Note that sale price errors can go in favor or against a seller (buyer). A risk averse seller is less likely to enter the market if the sale price outcome, even after cautious pricing and waiting, is a gamble.

Proposition 3: Buyers and Sellers are better off after introduction of ML if following conditions on error variance ratios ($\boldsymbol{\gamma = \sigma_z^2/\sigma^2}$) and ($\boldsymbol{\delta = \sigma_\epsilon^2/\sigma^2}$) are satisfied. Both conditions are relaxed if the ML noise underestimation is corrected.

Table 1: Conditions where market outcomes are better off after introduction of ML

Market Outcomes		**Condition where outcomes are better off**	
		Underestimated $\hat{\sigma}_z^2$	Corrected $(\sigma_z^2)_C$
A.	$E\left[\left(v_{i,k,t} - v_{k,t}\right)^2\right]$	$\frac{(\gamma-1)^2}{(\gamma-2)(2\gamma-3)} < 1$	Always
B.	$E\left[\left(p_{i,k,t} - v_{k,t}\right)^2\right]$	$\frac{(\gamma-1)^2}{(\gamma-2)(2\gamma-3)} < \delta$	$\frac{(\gamma+1)^2}{2\gamma^2+5\gamma+1} < \delta$

The former γ is a measure of unconfounded (by feedback) power of ML signal. The latter δ is a measure of market "self-sufficiency" in resolving valuation errors via buyer-seller cross learning.

Figure 3 visually depicts regions of (γ, δ) where introduction of ML improves the market participant payoff. The horizontal line at $\delta = 1$ corresponds to payoff Π component (A). With correctly estimated ML noise, component A is always better off after the introduction of ML. With under estimated ML noise, component A is worse off when private valuations are noisy (high σ^2). In this setting, participants eventually start relying entirely on ML. However such absolute reliance on ML is flawed because empirical ML error approximation is severely underestimated. It deceptively appears preferable to private valuation.

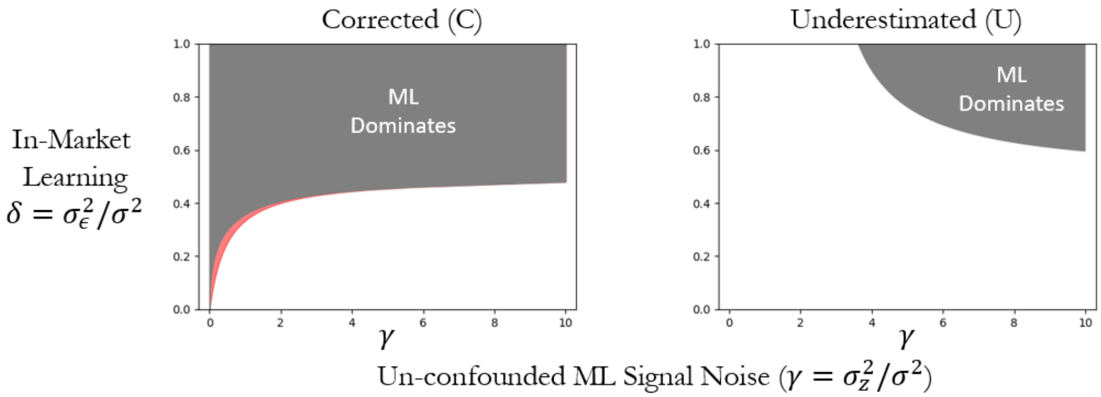


Figure 3: Regions of (γ, δ) where ML introduction makes market participants better off with corrected (Left) as well as under estimated (Right) ML noise. The Left figure also shows a slightly extended (γ, δ) where a hypothetical unconfounded ML would be better off.

A similar tradeoff holds true for payoff Π component (B). Once again with under estimated ML noise, component B is worse off when private valuations are noisy (high σ^2 or low γ). The crucial difference is the role of market "cross learning" or "self-sufficiency" δ. A small δ corresponds to a settings where buyers and sellers, each with noisy valuation, are able to resolve errors by learning via market interactions. If the market is not "self sufficient" or participant bear a large cost of learning (high δ), than importance of entry with good valuation and therefore utility of ML signal is high. Instead if the market is "self sufficient", then introduction of ML doesn't add any value in participants discovering the unbiased sale price. In fact, over reliance on ML only serves to add a correlated bias across all participants. Given this ML error is correlated, unlike private valuation errors that were uncorrelated, it is not resolved by market interaction. Thus ML error under low γ and low δ propagates uncorrected to sale price errors, thus making participants worse off.

Correction of Feedback Loop can be achieved by setting aside ρ proportion of houses where ML signal is hidden. Over T periods ρKT sale prices can be used to estimate unconfounded ML error $\hat{\sigma}_z^2(Q, 1)$ as,

$$\hat{\sigma}_z^2(Q,1) = \frac{1}{\rho K * T} \Sigma_t \Sigma_{k'} r(k;\rho) \left(\hat{e}^z_{j_k,t}\right)^2 \cong E\left[\left(e^z_{k,t}\right)^2\right] = \Lambda(Q,1) \tag{21}$$

The ML estimate $z_{k,t}$ using ρ and $(1-\rho)$ fraction is given by,

$$z_{k,t} = \frac{\sum_{k'} \Big(\lambda\big(1 - r(k;\rho)\big) + (1-\lambda) r(k;\rho)\Big) p_{k',t-1} \ \times \mathbf{1}(j_{k'} = j_k)}{\sum_{k'} \mathbf{1}(j_{k'} = j_k)} \tag{22}$$

where

$$r(k;\rho) \begin{cases} 1 & if\ k\ does\ not\ receive\ z \\ 0 & else \end{cases} \quad ; \quad \lambda = \frac{\hat{\sigma}_z^2(Q,1)}{\hat{\sigma}_z^2(Q,1) + \hat{\sigma}_z^2(Q,\alpha)} = \frac{\Lambda(Q,1)}{\frac{\Lambda(Q,\alpha)}{2\alpha - \alpha^2} + \Lambda(Q,1)} \tag{23}$$

While this or a similar statistical solution are straightforward, the challenge lies in mis-aligned incentives for the ML platform. ML platform revenues directly or indirectly depend on - number of houses where ML signal is available and proximity of ML signal with eventual sale prices. These metric do not incentivize the ML platform to sacrifice visibility of ML signal to ρ proportion of houses or deviate from the default “self fulfilling prophecy” outcome. Resolving these mis-aligned incentives remain an open question for future work.

Theoretical results in this section are driven by three crucial model primitives – (a) First, ML signal influences (ask or offer) actions taken by buyers and sellers in the market, thus indirectly influencing the sale prices. The feedback cycle is irrelevant if ML has little or no influence on sale price. (b) Second, recently observed sale prices of houses with matching features influences ML value estimates. Once again, feedback cycle is broken if ML estimates are relatively static (e.g., they do not attempt to follow the price random walk) and do not rely heavily on recent sales. (c) Third, ML platforms estimate of size of its own errors deviates from actual errors. In the next two section we provide empirical evidence for these two necessary primitives. Further, we also highlight significant variation in strength of the feedback cycle across housing sub markets.

4. Data Description

Table 2: Sample values for hose features, location, listing and price estimate.

Category	Variable	Sample Value
Location	Latitude	29.7 - 42.3 degrees North
	Longitude	71.0 - 95.3 degrees West
	Neighborhood	South Boston ,Carrick, Brighton Heights, etc
	Zip Code	15210, 15212, 15232, etc.
	County	Suffolk, Allegheny, Travis
Features	Floor Size	100 - 10,000 sq. ft.
	Year Built	1799 - 2019
	Last Remodel Year	1799 - 2019
	Bathrooms	0 - 15
	Bedrooms	0 - 15
	Parking	0 - 1000 sq. ft.
	Lot	100 - 10,000 sq. ft.
	Stories	0 - 50
	Solar Potential	0 - 100
	Type	Single Family, Multi Family, Condo, etc.
	Structure Type	Colonial, Victorian, Modern, etc.
	Roof Type	Composition, Shingle, Asphalt, etc.
	Flooring	Hardwood, Carpeted, Tile, etc.
	Patio	Porch, Deck, None, etc.
	Ex-Material	Brick, Wood, Cement, etc.
Listing	List Price	\$10,000 - \$10,000,000
	Days Listed	1 - 365 days
	Sale Price	\$10,000 - \$10,000,000
ML Price	Z (Zestimate)	\$10,000 - \$10,000,000

We use housing market data from Zillow.com, which is an online real estate database company. Zillow provides information on house features (e.g., floor size, year build), house location (e.g., county, zip code, street address), historical as well as current listing information (e.g., list price, sale price) and a price estimate called Zestimate. Zillow describes Zestimate as an “estimate of a home’s market value” and it presents a comparison of Zestimate with actual sale price outcomes as a measure if its accuracy. From this and other publicly available posts [17] from Zillow, we infer from that Zestimate is an ML power prediction of sale price as a function of house features, locations and economic environment. We have access to over 750,000 houses across Austin (Travis County, Texas), Boston (Suffolk County, Massachusetts) and Pittsburgh

(Allegheny County, Pennsylvania). Table 2 provides samples values of features we collect from Zillow.

While house features and location are largely static, listing state and Zestimate change over time. Zillow provides listing and Zestimate history going back up to 10 years. Since Zillow regularly makes minor and major upgrades to Zestimate algorithm, entire 10 year Zestimate trend shifts on a regular basis. For example on 1st Jan 2020, Zillow presents a Zestimate trend from $100,000 on 1st Jan 2010 to $122,000 on 1st Jan 2020 increasing at 2% every year for a house. But following an algorithm update on 1st Feb 2020, Zillow presents a Zestimate trend from $100,000 on 1st Jan 2010 to $148,000 on 1st Jan 2020 increasing at 4% every year for the same house. Thus two Zestimate versions, pre update and post update, assign two different values to the same house on the same date. Our empirical analysis requires us to keep track of such algorithm updates. Therefore we take 25 snapshots of Zillow information at a frequency of approximately two weeks between Feb 2019 and March 2020.

5. Empirical Evidence

In this section we first provide evidence for impact of Zestimate z_i on house sale prices p_i. Second, we show how house sale prices p_j impact the Zestimate z_i of "peer" houses in the neighborhood.

5.1 Impact of Zestimate on House Sale Price

The impact of Zestimate on Sale Price can be determined by randomly assigning similar houses into two groups. First group is provided a Zestimate 5% over true price, while the second group is provided a Zestimate 5% under true price. With both home buyers and sellers having access to these experimental prices, we can determine the difference between average sale prices in the two groups. Naturally, it is not possible to experimentally present incorrect prices in the actual housing market nor is it possible to mimic the entire home search process in lab. But, we know that Zestimate algorithm is frequently upgraded to reduce errors. Thus we can infer historical instances where an erratic Zestimate was presented in the market. The current Zestimate available for a house when it was listed on the market at time t is z_i. The Zestimate available at time $T \gg t$ post algorithm upgrade is $\bar{z}_i$. The error in Zestimate e_i $(e_i = z_i - p_i)$ can be instrumented using z_i^e.

$$z_i = p_i + e_i \quad ; \quad \bar{z}_i = p_i + \bar{e}_i \quad ; \quad z_i^e = \frac{z_i - \bar{z}_i}{\bar{z}_i} \tag{24}$$

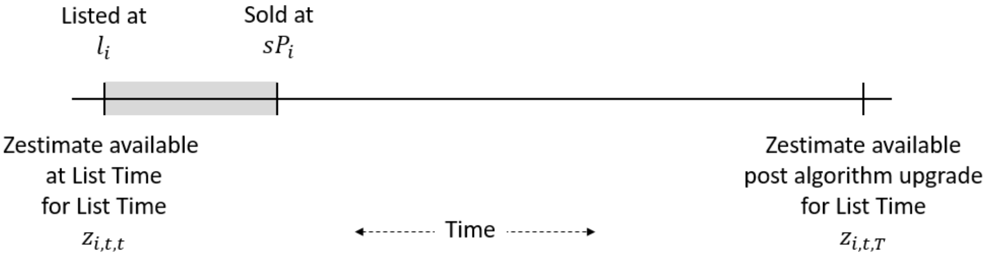


Figure 4: An illustration of list price (l_i), zestimate z_i ($= z_{i,t,t}$) for time t available at list time t and zestimate $\bar{z}_i$ ($= z_{i,t,T}$) for time t available post algorithm update at time T.

We create two groups, one that received positive error (z_i^e >+1%) and other that received negative error (z_i^e <-1%). While we do not exactly know the true price of a house, a (less erratic) post algorithm upgrade Zestimate $\bar{z}_i$ and expansive set of house features $\vec{X}_i$ (such as Floor Area, Number of Bedrooms, Year of construction etc) contain information on the underlying true price. Therefore two groups propensity score matched (PSM) on post algorithm upgrade Zestimate $\bar{z}_i$ and house features $\vec{X}_i$ are expected to have similar true prices. We can measure the difference in average sale prices across the two groups. Figure 5,6 and Table 3 report effectiveness of PSM[10].

$$pScore_i(\phi) = \frac{1}{1 + \exp\left(-[\bar{z}_i, \vec{X}_i] * \boldsymbol{\phi}\right)} \tag{25}$$

The first group received an average Zestimate error z_i^e =7.1% while the second group received an average Zestimate error z_i^e =-7.1%. Table 4 Model 1 reports a difference in sale price of 2.22% among the two groups. This suggests that Zestimate error gap of 14.2% is shrunk to a to

[10] *Appendix Table 12, 13 reports standardized bias for every covariate before and after PSM. It also reports the descriptive statistics for key covariates in the matched samples.*

a sale price gap of 2.22% because buyers and sellers only rely partially on Zestimate for pricing. Model 3 suggest that the difference in sale prices is as high as 7.5% among small $50,000 (or 500 square feet) houses that are likely to be more standardized and may witness greater reliance of buyers and sellers on Zestimate. The difference is only 1.4% among large $500,000 houses. If buyers and sellers relied entirely on Zestimate for pricing, we would see a sale price difference of 14.2% between the two groups. Using the observed sale price difference, we infer degree of reliance on Zestimate as ranging from 0.1 (14.2% → 1.4%) to 0.5 (14.2% → 7.5%). Going forward we use the range 0.1-0.5 to emphasize differences in reliance on Zestimate across housing sub markets.

$$log(p_i) = \beta_0 + \beta_1\, log(\bar{z}_i) + \boldsymbol{\beta_2}\, \underbrace{(z_i^e > 0)}_{Treatment} + \beta_3 pScore_i \tag{26}$$

Table 3: Treatment and Control statistics before and after PSM. A Rubin's R between 0.5-2.0 and a Rubin's B less than 25 indicate a good match.

Statistic	Before Matching	After Matching
Mean % Std. Bias	10.21%	0.93%
Rubin's R	0.99	1.01
Rubin's B	71.82%	0.30%

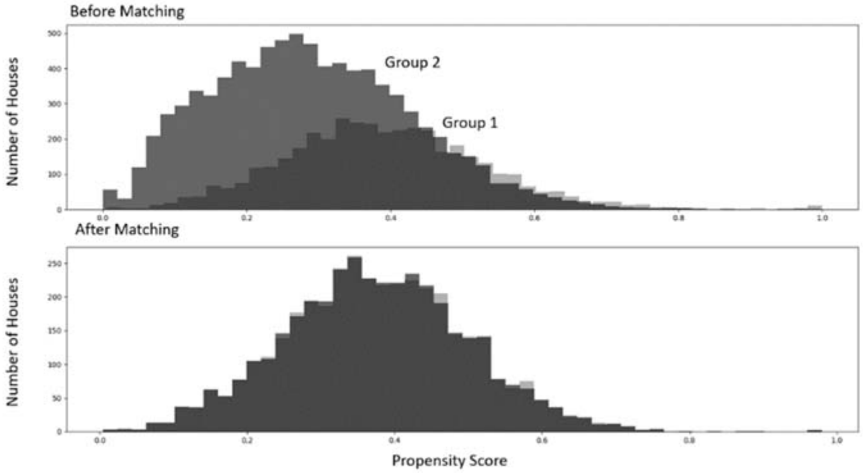


Figure 5: Histogram of propensity score for treatment and control groups before and after PSM.

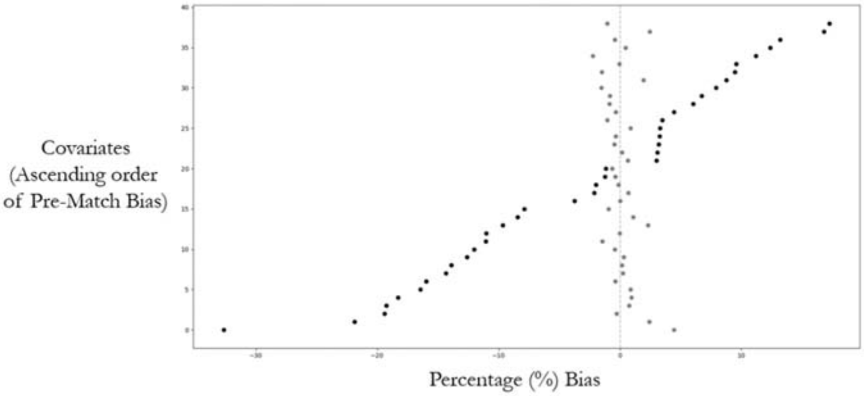


Figure 6: Percentage Standardized Bias for matching covariates before (dark) and after (light) PSM.

Next we use observed list price and time to sale to uncover Zestimate impact on buyers and sellers independently. Table 5 Model 1 reports a difference in initial list price of 1.4% among the two groups. Since list price is decided by the seller in isolation, this confirms Zestimate impact on seller. If Zestimate had no impact on buyers, houses that are listed higher under influence of positive Zestimate error should be harder to sell. Controlling for Seller's initial list price, Model 2 reports a difference in sale price of 1.2% and Model 3 report a shorter time to sale by 4 days. Both of these demonstrates Zestimate influence on buyer's willingness to pay.

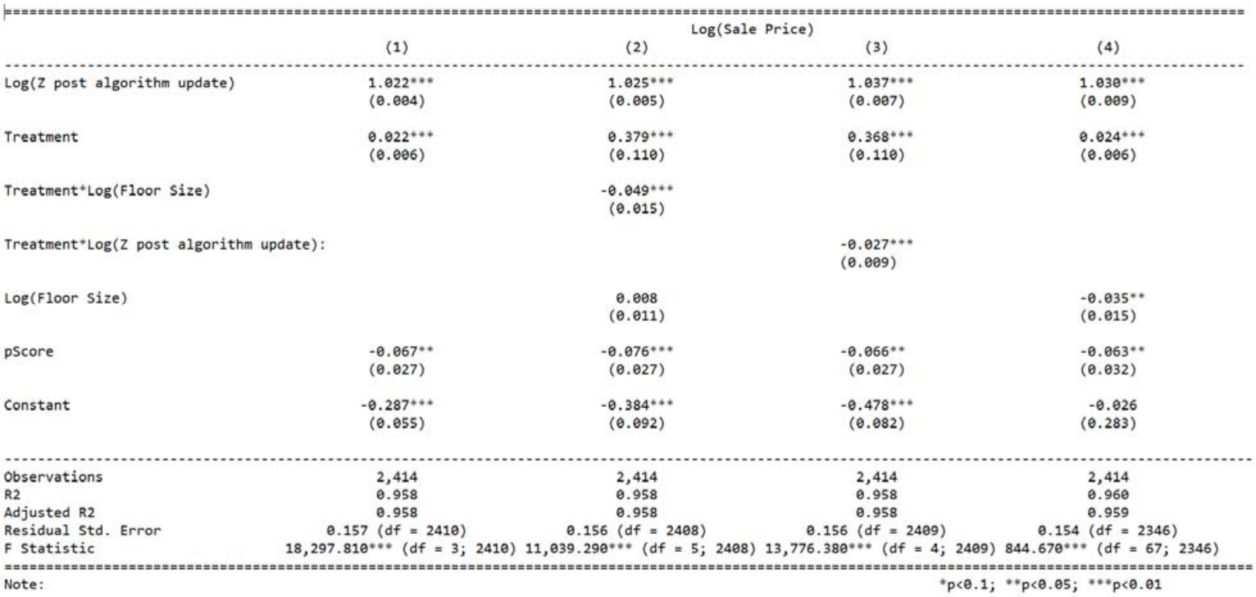
$$log(l_i) = \beta_0 + \beta_1\, log(\bar{z}_i) + \boldsymbol{\beta_2} \underbrace{(z_i^e > 0)}_{Treatment} + \beta_3 pScore_i \tag{27}$$

$$log(p_i) = \beta_0 + \beta_1\, log(\bar{z}_i) + \boldsymbol{\beta_2} \underbrace{(z_i^e > 0)}_{Treatment} + \beta_3 pScore_i + \beta_4 \underbrace{\left(\frac{l_i - \bar{z}_i}{\bar{z}_i}\right)}_{Markup} \tag{28}$$

5.2 Impact of House Sale Price on Zestimate

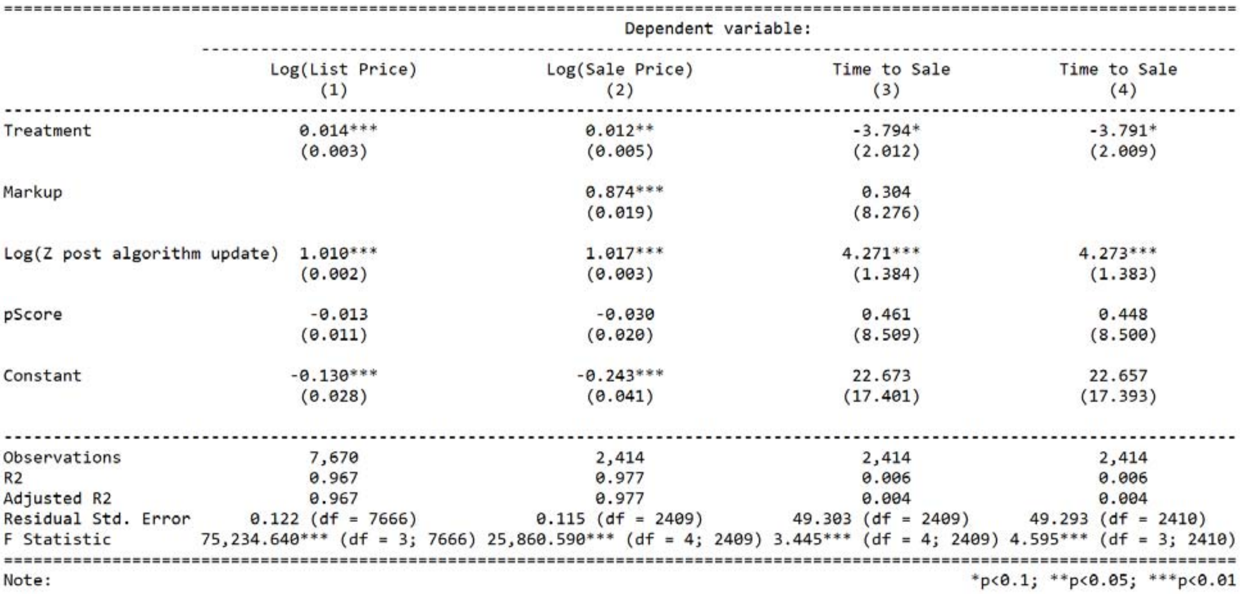
The Zestimate algorithm is proprietary to Zillow and thus opaque to us. Zillow describes the Zestimate as composed of both expert driven economic modeling and data driven predictive ML. For the Feedback Loop phenomenon to be relevant, it is critical that the latter, data driven predictive ML, is a significant driver for Zestimate. We know that Zillow reports a set of 4 or 5 house sales alongside Zestimate of every house. These houses, referred to as peer houses hereon, are in geographical vicinity and have similar features as the focal house. In Appendix A.3,

Table 4: (Model 1)Primary specification for impact of Zestimate error treatment on sale price (equation 26). (Model 2) Interaction of Zestimate error treatment with house size. (Model 3) Interaction of Zestimate error treatment with house price. (Model 4) The treatment effect after adding explicit regression controls for house features $\vec{X}_i$.



```
==========================================================================================================================
                                                                   Log(Sale Price)
                                     (1)                        (2)                         (3)                     (4)
--------------------------------------------------------------------------------------------------------------------------
Log(Z post algorithm update)       1.022***                   1.025***                    1.037***                1.030***
                                   (0.004)                    (0.005)                     (0.007)                 (0.009)

Treatment                          0.022***                   0.379***                    0.368***                0.024***
                                   (0.006)                    (0.110)                     (0.110)                 (0.006)

Treatment*Log(Floor Size)                                    -0.049***
                                                              (0.015)

Treatment*Log(Z post algorithm update):                                                  -0.027***
                                                                                          (0.009)

Log(Floor Size)                                               0.008                                              -0.035**
                                                              (0.011)                                             (0.015)

pScore                             -0.067**                  -0.076***                   -0.066**                -0.063**
                                   (0.027)                    (0.027)                     (0.027)                 (0.032)

Constant                          -0.287***                  -0.384***                   -0.478***                -0.026
                                   (0.055)                    (0.092)                     (0.082)                 (0.283)

--------------------------------------------------------------------------------------------------------------------------
Observations                        2,414                      2,414                       2,414                   2,414
R2                                  0.958                      0.958                       0.958                   0.960
Adjusted R2                         0.958                      0.958                       0.958                   0.959
Residual Std. Error           0.157 (df = 2410)          0.156 (df = 2408)           0.156 (df = 2409)       0.154 (df = 2346)
F Statistic             18,297.810*** (df = 3; 2410) 11,039.290*** (df = 5; 2408) 13,776.380*** (df = 4; 2409) 844.670*** (df = 67; 2346)
==========================================================================================================================
Note:                                                                                       *p<0.1; **p<0.05; ***p<0.01
```

Table 5: (Model 1) Impact of Zestimate error treatment on list price (equation 27). (Model 2) Impact of Zestimate error treatment on Sale Price, controlling for list price (equation 28). (Model 3 and 4) Impact of Zestimate error treatment on Time to Sale (in days).



```
=============================================================================================================
                                                      Dependent variable:
                         ------------------------------------------------------------------------------------
                              Log(List Price)            Log(Sale Price)           Time to Sale       Time to Sale
                                    (1)                        (2)                      (3)                (4)
-------------------------------------------------------------------------------------------------------------
Treatment                        0.014***                   0.012**                  -3.794*            -3.791*
                                 (0.003)                    (0.005)                  (2.012)            (2.009)

Markup                                                      0.874***                  0.304
                                                            (0.019)                  (8.276)

Log(Z post algorithm update)  1.010***                      1.017***                  4.271***           4.273***
                                 (0.002)                    (0.003)                  (1.384)            (1.383)

pScore                            -0.013                     -0.030                   0.461              0.448
                                 (0.011)                    (0.020)                  (8.509)            (8.500)

Constant                        -0.130***                  -0.243***                 22.673             22.657
                                 (0.028)                    (0.041)                 (17.401)           (17.393)

-------------------------------------------------------------------------------------------------------------
Observations                      7,670                      2,414                    2,414              2,414
R2                                0.967                      0.977                    0.006              0.006
Adjusted R2                       0.967                      0.977                    0.004              0.004
Residual Std. Error         0.122 (df = 7666)          0.115 (df = 2409)        49.303 (df = 2409)  49.293 (df = 2410)
F Statistic           75,234.640*** (df = 3; 7666) 25,860.590*** (df = 4; 2409) 3.445*** (df = 4; 2409) 4.595*** (df = 3; 2410)
=============================================================================================================
Note:                                                                          *p<0.1; **p<0.05; ***p<0.01
```

we reverse engineer the choice of these peer sales and determine a precise criteria with a high confidence – (i) sales within past 12 months, (ii) distance <2 km from focal house and (iii) floor size that is between half to double of the focal house. Figure 7 and Table 6 illustrates an example house and its five peer sales. We hypothesize that a simple average sale price of peer houses $\bar{p}_j$ is a dominant driver of Zestimate z_i. If established, it would greatly simplify the mechanism behind Zestimate calculation and help to illustrate the presence of Feedback Loops.

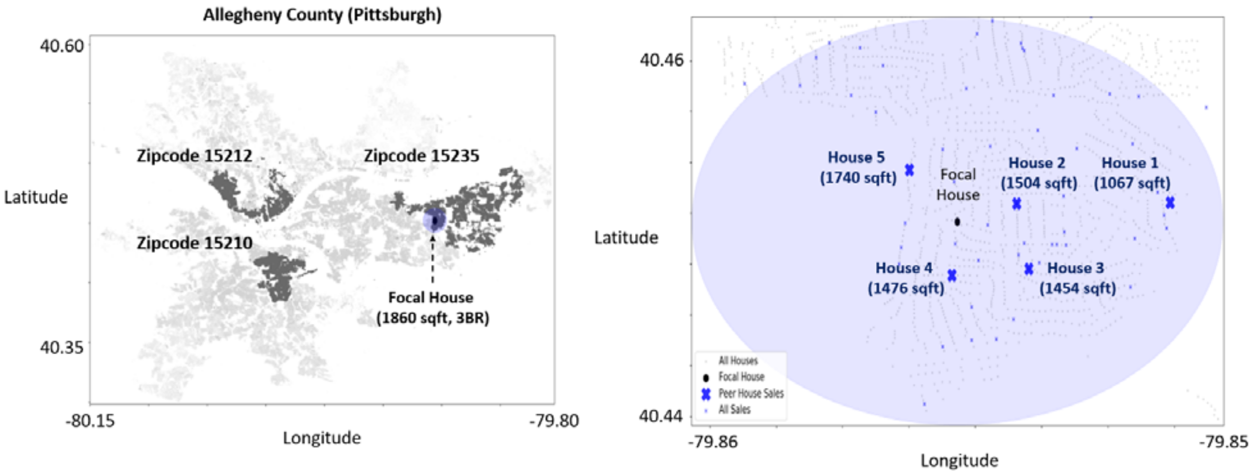


*Figure 7: (Left) All houses in Allegheny County. Houses from three zip codes are used for evaluating impact of peer sales on Zestimate. A 3 bedroom 1660 square feet house in Zip Code 15235 is depicted as an example. (Right) A zoom in view of 1.5 km radius around the focal house shows many sales in last 12 months (small blue x). Five peer houses with similar floor size are shown as large blue **X**.*

Table 6: Floor size, sale price and sale time for each of these five peer houses.

Peer House	Floor Size (sqft)	Sale Price	Sale Month	Peers of Peer Houses
1	1067	$160,575	04/2019	a, **e**, f, **3**, **4**
2	1504	$161,000	04/2019	b, c, d, **3**
3	1454	$160,500	03/2019	c, d, **e**, f, g
4	1476	$149,900	03/2019	f, h, i, **3**
5	1740	$160,000	12/2018	Not available in our sample

One approach to establish the role of $\bar{p}_j$ would be to, consider a pair of adjacent similar houses - A and B. At time t_1, both houses have same Zestimate, same peer house sets ($J = \{1,2,3,4\}$) and therefore same $\bar{p}_j$. At time t_2, the peer set for one house remains static while the peer set changes for the other house. Given the geographical adjacency and similarity in features, any difference in Zestimate $(z_{B,t_2} - z_{A,t_2})$ at time t_2 arises from change in the peer set $(J_B = \{1,2,3,4\} \rightarrow \{1,3,4,5\})$. Thus we would be able to isolate the role of average sale price of peer houses $\bar{p}_j$ as a dominant driver of Zestimate z_i.

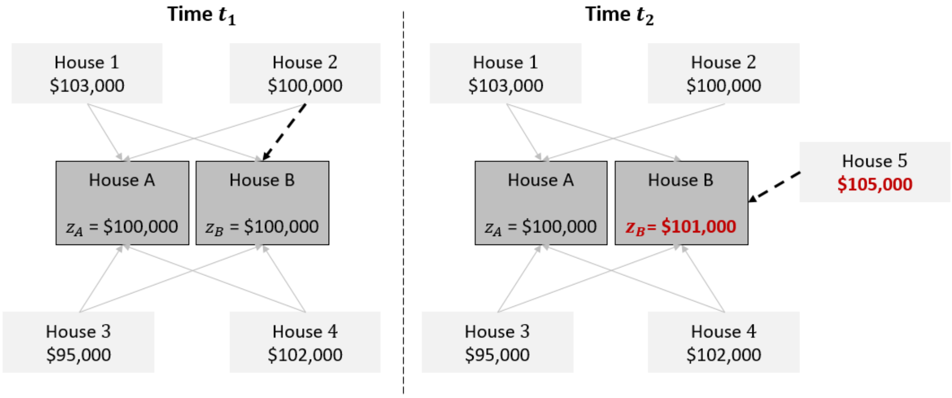


Figure 8: A hypothetical example where at time t_1, House A and B have the same set of peer sales {1,2,3,4}, thus the same "average peer sale price" and Zestimate. At time t_2, House 5 replaces House in the peer set of House B. The "average peer sale price" for House B increases by $1,250 and its Zestimate increases by $1,000.

$$pScore_i(\phi) = \frac{1}{1 + exp\left(-\left[\left(\bar{p}_j\right)_{i,t_1}, z_{i,t_1}, \theta_{i,t_1}, \Delta\theta_i, \vec{X}_i\right] * \boldsymbol{\phi}\right)} \quad ; \quad \left(\bar{p}_j\right)_{i,t} = \left(\sum_{j \in peers\ of\ i\ at\ t} p_j\right) \tag{29}$$

We operationalize this approach by creating – control group with no change in peer set and treatment group with change in peer set corresponding to 0-5% increase in $\bar{p}_j$ with an average 2.3% increase. Next, we propensity score match (PSM) the control and treatment group units on – peer set at t_1, Zestimate at t_1, house features $\vec{X}_i$ and other potential drivers of Zestimate change (θ) such as tax estimate. The treatment effect of $\Delta\bar{p}_j$ can now be measured as β_2 in equation (29). An average 2.3% increase in $\bar{p}_j$ corresponds to a 1.9% increase in Zestimate. We repeat the same steps with two alternative treatment groups - 5-10% increase in $\bar{p}_j$ and 10-15% increase in $\bar{p}_j$. Table 7 reports the corresponding results. Overall, a 1% increase in $\bar{p}_j$ corresponds to 0.66-0.83% increase in Zestimate.

$$\Delta z_i = \left(z_{i,t_2} - z_{i,t_1}\right) = \beta_0 + \beta_1 \Delta\theta_i + \boldsymbol{\beta_2}\underbrace{(\Delta\bar{p}_j > 0)}_{Treatment} + \beta_3 pScore_i \qquad (29)$$

Table 7: Treatment effects using three different treatment groups with varying level of treatment intensity or "average peer sale price" change.

Treatment Group	Range of Treatment intensity change	Mean Treatment Intensity	Treatment Effect (β_2)	Treatment Effect/Intensity
I	0 – 5 %	2.3%	1.9%	0.83
II	5 – 10 %	6.7%	4.4%	0.66
III	10 – 15%	11.5%	8.6%	0.75

In order to further reinforce the role of $\bar{p}_j$ in driving Zestimate z_i, we also look at out of sample explanatory power. We find that this simple metric $\bar{p}_j$ explains close to 96% variation in the Zestimate. While we are aware that the actual Zestimate model may be significantly more sophisticated, the extremely high out of sample explanatory power ($R^2 \approx 0.96$) suggest that a big chunk of the full model is driven by information contained in peer sale average. In Appendix, we present more details as well as alternative models that incorporate – weighted average of peer sale, unsold peer listings, focal house features, geographical fixed effects, tax information etc. In summary, we find that these contain much less Zestimate explanatory power on their own and add very little or no explanatory power when used alongside the simple peer sale average metric.

5.3 Feedback Loop

Empirical evidence in this section suggest that market participants have a reliance of 0.1 to 0.5 on Zestimate. For example, if Zestimate of a house has an error of +10%, sale prices is likely to shift up by 1 to 5%. Consider such a house that sells 5% over its true price. Empirically we observe that a single new house sale acts as a peer sale for 25 houses on average, but it may act as peer sale for up to 500 houses. Assuming a house has four peers, the average peer price $\bar{p}_j$ for up to 500 houses increases by 1.25%. Since every 1% increase in $\bar{p}_j$ corresponds to 0.66-0.83% increase in Zestimate, this is equivalent to a Zestimate increase of 0.82 to 1% (0.66*1.25 to 0.83*1.25%) for up to 500 houses. This would be a significant artificial price inflation originating from a Zestimate error on a single house. This Feedback is intensified if market participants over rely on

Zestimate. In Section 2, we describe under estimation of ML errors as a key driver of over reliance. If Zestimate has large errors, but Zillow underestimates these errors and projects high level of confidence, the error amplification via the confounded feedback is more intense.

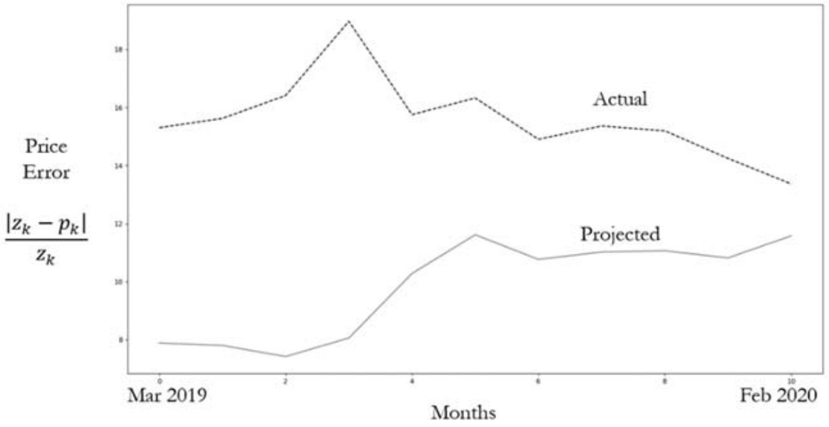


Figure 9: Actual monthly average error between Zestimate and sale price for houses sold between March 2019 and Feb 2020. Projected error refers to monthly average of Zillow's estimate of likely error for the same houses before the houses go on the market.

Figure 9 provides monthly average of error between Zestimate available before a house goes on the market and its eventual sale price. This "actual" error is inclusive of any confounding since Zestimate is likely to pull the sale prices closer to itself. One would expect this "actual" error to be the lower bound on the unconfounded error. Figure 9 also provides Zillow's own estimate of likely error in its Zestimate for the same set of houses before the houses go on the market. On a cursory look, Zestimate error appears to be significantly under estimated by Zillow. But its worth noting that Zillow doesn't provide the exact formula behind its error estimate. For example, it simply says that sale price of a house is likely to be within $\pm$7% of Zestimate $\$200,000$. Figure 9 simply plots the confidence interval (e.g., 7%) averaged over all houses that sold in a given month. Nevertheless, its important to note that Zillow likely performed an upgrade of its confidence interval calculations in July-August 2019. Its projected confidence went down from $\pm$8% to $\pm$11% on an average[11]. Since its "actual" error have been declining, its very likely that confidence interval or equivalently projection of its Zestimate error in the first of half of 2019 were

[11] *Zillow also reported [17] median errors for houses in Boston and Pittsburgh as 7.4% and 12% respectively. These numbers also appear to be optimistic or underestimated for what is empirically observed.*

underestimated. Since Zillow's algorithm is opaque, these empirical observation are purely a conjecture.

6. Conclusion

We expect ML price with its costless accessibility, perceived impartiality and high reported accuracy to have significant influence on housing market participants. An average 15% reliance , and up to 50% reliance in some sub markets, on ML price empirically validates this hypothesis. We also expect ML to learn accurate pricing from modeling large amounts of historical data. But in contrast to this conventional wisdom, we find that ML pricing can be myopic and "self fulfilling". ML price of a house is largely a derivative of sale prices of 4-5 peer houses. A single sale can contribute 10% in ML price for up to 500 peer houses. Further, ML platform may systematically under estimates its own error. Using these primitives we propose a theoretical model of ML Feedback Loop. We formulate equilibrium level of ML error estimate, reliance on ML and resulting expected payoff for the participants. We identify condition where participants can be worse off after introduction of ML. A low complexity ML model is more robust under the confounded Feedback Loop. A "self-sufficient" market can efficiently resolve disagreement in participant private valuations, but it is more likely to be worse off after introduction of ML.

We anticipate two significant lines of enquiry in follow up to this paper. First, a solution to the Feedback Loop that deals with the statistical challenge as well as the problem of mis-aligned ML platform incentives to sacrifice revenue in favor of any correction procedure. Second, a full structural estimation of utility and payoff for heterogeneous market participants. In particular, distinguishing payoffs for participants heterogeneous in ability to accurately price and patience to spend time in the market can be particularly informative. Prior to introduction of ML, the housing market is characterized by – long time to sale, large participation costs, and high variance in valuations. This benefits participants that are willing to patiently spend time in the market but keeps out participants with high ability to accurately price. The latter category of participants prefer to utilize their superior pricing ability at transacting higher level financial instruments e.g., REITs, securitized home loans. Introduction of ML changes market characteristics to – shorter time to sale, lower perceived participation costs, and high bias in valuation. This encourages entry of participants with high ability to accurately price. While one expects ML pricing to level the playing field across participants with different ability to price, in fact it may increase the disparity. Naturally, this is a conjecture that requires future work.

References

[1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

[2] Anglin, P. M., & Arnott, R. (1991). Residential real estate brokerage as a principal-agent problem. The Journal of Real Estate Finance and Economics, 4(2), 99-125.

[3] Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., ... & Snelson, E. (2013). Counterfactual reasoning and learning systems: The example of computational advertising. The Journal of Machine Learning Research, 14(1), 3207-3260.

[4] Chaney, A. J., Stewart, B. M., & Engelhardt, B. E. (2018, September). How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In Proceedings of the 12th ACM Conference on Recommender Systems (pp. 224-232).

[5] Cheng, I. H., Raina, S., & Xiong, W. (2014). Wall Street and the housing bubble. American Economic Review, 104(9), 2797-2829.

[6] Goodman Jr, J. L., & Ittner, J. B. (1992). The accuracy of home owners' estimates of house value. Journal of housing economics, 2(4), 339-357.

[7] Han, L., & Strange, W. C. (2015). The microstructure of housing markets: Search, bargaining, and brokerage. In Handbook of regional and urban economics (Vol. 5, pp. 813-886). Elsevier.

[8] Ihlanfeldt, K. R., & Martinez-Vazquez, J. (1986). Alternative value estimates of owner-occupied housing: evidence on sample selection bias and systematic errors. Journal of Urban Economics, 20(3), 356-369.

[9] Kiel, K. A., & Zabel, J. E. (1999). The accuracy of owner-provided house values: The 1978–1991 American Housing Survey. Real Estate Economics, 27(2), 263-298.

[10] Linneman, P. (1986). An empirical test of the efficiency of the housing market. Journal of Urban Economics, 20(2), 140-154.

[11] Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. arXiv preprint arXiv:2002.06673.

[12] Redfin.com. 16 September 2020. About the Redfin Estimate: Home Value Estimator. https://www.redfin.com/redfin-estimate

[13] Schmit, S., & Riquelme, C. (2018, March). Human interaction with recommendation systems. In International Conference on Artificial Intelligence and Statistics (pp. 862-870).

[14] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. In Advances in neural information processing systems (pp. 2503-2511).

[15] Sinha, A., Gleich, D. F., & Ramani, K. (2016). Deconvolving feedback loops in recommender systems. In Advances in neural information processing systems (pp. 3243-3251).

[16] Wager, S., Chamandy, N., Muralidharan, O., & Najmi, A. (2014). Feedback detection for live predictors. In Advances in Neural Information Processing Systems (pp. 3428-3436).

[17] Zillow.com. 16 September 2020. What is a Zestimate: Zillow’s Zestimate Accuracy. https://www.zillow.com/zestimate/

Appendix

A.1 Performative Stability

We follow Perdomo et al (2020) to identify convergence properties of the ML Feedback Loop in the housing market. The Machine Learning model $\theta \equiv [z_{1,\dots,J}, \sigma_z^2]$ induces a joint distribution $\mathcal{D}$ over covariates[12] and outcome variable $y \equiv (j_k, p_k)$. The model θ is evaluated over resulting distribution $\mathcal{D}(\theta)$ via a loss function $l(y;\theta)$. A model θ_{PS} is performatively stable if it minimizes risk over the distribution $\mathcal{D}(\theta_{PS})$ imposed by itself. In practice, ML model is iteratively updated ($\theta_t \rightarrow \theta_{t+1}$) with repeated empirical risk minimization (RERM) using n_t samples drawn from $\mathcal{D}(\theta_t)$ at iteration t. We want to understand convergence to performative stable θ_{PS} using RERM. We need to loss $l(y;\theta)$ to be "nicely" behaved and the distribution $\mathcal{D}(\theta)$ induced by the ML model θ to be sufficiently insensitive to changes in the ML model θ. More formally we check whether $l(y;\theta)$ is β-jointly smooth and γ-strongly convex, and whether $\mathcal{D}(\theta)$ is ϵ-sensitive with $\epsilon < \gamma/2\beta$.

$$l(y;\theta) = l(j_k, p_k; z_{1,\dots,J}, \sigma_z^2) = \left(p_k - z_{j_k}\right)^2 + \left(\left(p_k - \widetilde{z_{J_k}}\right)^2 - \sigma_z^2\right)^2$$

$$\theta_{PS} = \underset{\theta}{\operatorname{argmin}} \underset{y \sim \mathcal{D}(\theta)}{\mathrm{E}} l(y;\theta)$$

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmin}} \underset{y \sim \mathcal{D}^{n_t}(\theta_t)}{\mathrm{E}} l(y;\theta)$$

β-jointly smooth enforces a strong form of continuity such that $l(y;\theta)$ changes slowly both in y and θ. This is satisfied for all $\beta \geq 2$.

$$\left\|\nabla_\theta l(y;\theta) - \nabla_\theta l(y;\theta')\right\|_2 \leq \beta\left\|\theta - \theta'\right\|_2 \quad \forall\, \theta, \theta' \in \Theta$$
$$\Rightarrow 4\left(\left(z_{j_k} - z_{j_k}{}'\right)^2 + (\sigma_z^2 - \sigma'^2_z)^2\right) \leq \beta^2\left((z_1 - z_1{}')^2 + \cdots + \left(z_J - z_J{}'\right)^2 + (\sigma_z^2 - \sigma'^2_z)^2\right)$$

[12] The probability distribution over covariates is independent of ML model in our case. We made an implicit assumption that the ML model imposes a shift in sale price, but does not impose any change in sample of houses that go on the market in any given period. In practice, houses with high accuracy ML model may be more likely to go on the market.

$\Rightarrow \beta \geq 2$

$\left||\nabla_\theta l(y;\theta) - \nabla_\theta l(y';\theta)|\right|_2 \leq \beta\left||y - y'|\right|_2 \quad \forall\, y, y' \in \mathrm{Y}$
$\Rightarrow 4(\, p_k - p_k{}')^2 \;\leq \beta^2(p_k - p_k{}')^2$
$\Rightarrow \beta \geq 2$

γ-strongly convexity extends strict convexity of $l(y;\theta)$. This is satisfied for all $\gamma \leq 2$.

$l(y;\theta) \geq l(y;\theta') + \nabla_\theta l(y;\theta')^T(\theta - \theta') + (\gamma/2)\left||\theta - \theta'|\right|_2^2$
$\Rightarrow \left(p_k - z_{j_k}\right)^2 \geq \left(p_k - z_{j_k}{}'\right)^2 + 2\left(p_k - z_{j_k}{}'\right)\left(z_{j_k} - z_{j_k}{}'\right) + (\gamma/2)\left(z_{j_k} - z_{j_k}{}'\right)^2$
$\Rightarrow \left(p_k - z_{j_k}\right)^2 \geq \left(p_k - z_{j_k}\right)^2 + (\gamma/2 - 1)\left(z_{j_k} - z_{j_k}{}'\right)^2$
$\Rightarrow 0 \geq \gamma/2 - 1$
$\Rightarrow \gamma \leq 2$

ϵ-sensitivity checks if the earth mover distance (Wasserstein-1) between probability distributions $\mathcal{D}$ induced by two different models θ and θ' is sufficiently small. Specifically, ϵ must be less than $\gamma/2\beta$. Using maximum and minimum values of γ and β respectively, $\gamma/2\beta \leq 0.5$. Thus we need $\epsilon < 0.5$, which is satisfied for $(1-\alpha) \leq \epsilon$ and $\sigma_z \leq \epsilon$. This corresponds low reliance and small errors in an ML predictions respectively.

$W_1\left(\mathcal{D}(j_k, p_k;\theta), \mathcal{D}(j_k, p_k;\theta')\right) \leq \epsilon\left||\theta - \theta'|\right|_2$
where $\mathcal{D}(j_k, p_k;\theta) = (1/J)\Sigma_j\phi\left(p_k - \alpha v_k - (1-\alpha)z_j, \sigma_\epsilon\right)$
$\Rightarrow \left||(\alpha - \alpha')v_k + (1-\alpha)z_j - (1-\alpha')z'_j|\right|_2 \leq \epsilon\left||[z_{1\ldots J}, \sigma_z^2] - [z'_{1\ldots J}, \sigma'^2_z]|\right|_2$
$\Rightarrow \left||(1-\alpha)(z_j - z'_j)|\right|_2 \leq \epsilon\left||z_{1\ldots J} - z'_{1\ldots J}|\right|_2$ and $\left||(\alpha - \alpha')(v_k - z_j)|\right|_2 \leq \epsilon\left||\sigma_z^2 - \sigma'^2_z|\right|_2$
$\Rightarrow (1-\alpha) \leq \epsilon$ and $\sigma_z \leq \epsilon$

Under “nicely” behaved loss $l(y;\theta)$ and sufficiently insensitive distribution $\mathcal{D}(\theta)$ RERM initialized at θ_0 converges within radius δ of stable point θ_{PS} using n_t samples at iteration t with probability $1-p$.

$$\left||\theta_t - \theta_{PS}|\right|_2 \leq \delta$$

where

$$n_t = \mathcal{O}\left(\frac{1}{(\epsilon\delta)^m}\log(t/p)\right) \;\; ; \;\; t \geq \left(1 - \frac{2\epsilon\beta}{\gamma}\right)^{-1} \log\left(\frac{\left\|\theta_0 - \theta_{PS}\right\|_2}{\delta}\right)$$

A.2 Payoff Calculation

In Section 2, we motivated valuation error and sale price error as drivers of participants payoff Π. Now we provide an approach to empirically estimate payoff as a function of heterogeneous participant characteristics and ML signal. This will answer – what type of participant are better off after introduction of ML, and under different ML settings.

A house listed for sale on the market receives offers from prospective buyers every week. The best offer for house k in week τ is $(\mu_k + e_{k,\tau})$. Here $e_{k,\tau} \sim N(0, \sigma_B)$ reflects noise in arrival of offers and heterogeneity in buyer preferences. A seller k starts with a belief $N(\mu_k^0, \sigma_S^0)$ at week $\tau = 0$ about the best offer $(\mu_k + e_{k,\tau})$ they will receive if their house listed for sale on the market. After spending τ weeks in the market, the seller updates their belief to $N(\mu_k^\tau, \sigma_S^\tau)$. The more weeks they spend in the market the closer their belief gets to actual average offer $\mu_k^\tau \to \mu_k$ and $\sigma_S^\tau \to \sigma_B$. Hypothetically, if all buyers and sellers had perfectly matching valuations $(\sigma_S^\tau = \sigma_B = 0, \mu_k^\tau = \mu_k)$ then every house would be listed at $l_{k,0} = \mu_k$ and sell instantaneously. In practice, noisy valuations mean that market participants act cautiously to protect themselves from erratic decisions. For example, sellers (buyers) enter the market at a high list price (low offer price), they prefer to cautiously lower the list price once they observe low offers from buyers. Thus they incur the cost of keeping the house on market for longer in order to resolve their valuation noise.

Offer Parameter Estimates: For every house on market, we observe house characteristics, list price and whether or not the week τ ends in a sale completion $(X_k, l_{k,\tau}, sale_{k,\tau})$, until the sequence terminates in a sale or exit from the market. Observed sale or no sale outcome $sale_{k,\tau}$ at week τ reveals the underlying offer distribution parameterized by $\hat{\mu}_k$ and $(\hat{\sigma}_B)_j$. We assume $\hat{\sigma}_B$ to be a constant in a submarket j e.g., 1000-2000 sqft single family homes in Shadyside neighborhood of Allegheny county. Note that individual buyers valuations and willingness to offer may evolve as they spend time in the market. But, we can not discern behavior and dynamics of an individual buyer. Therefore we only model offers drawn from a stationary distribution that is a combination of a crowd of buyers at different stage of their search process.

Listing Choice Parameter Estimates: The seller chooses list price $l_{k,\tau}$ to maximize their lifetime value $V(S) = V\left(\left\{\mu_k^\tau, \sigma_S^\tau, l_{k,\tau}\right\}; \sigma^B, rI_k, oMC\right)$ which is a function of state $\left\{\mu_k^\tau, \sigma_S^\tau, l_{k,\tau}\right\}$, variance in offers ($\sigma^B$), fixed costs (e.g., outside option of rental income rI_k, on market cost oMC), and discount factor. Observed sequence $l_{k,\tau}$ reveals - seller beliefs $N(\hat{\mu}_k^0, \hat{\sigma}_S^0)$, costs ($\widehat{rI}_k, \widehat{oMC}$), lifetime value $\hat{V}$ and optimal listing policy $CCP(a/S) = \widehat{CCP}\left(l_{k,\tau+1}/\left\{\mu_k^\tau, \sigma_S^\tau, l_{k,\tau}\right\}\right)$.

Buyer and Seller Sensitivity to ML Signal: Introduction of ML signal (z, σ_z) has two main effects,

a. $\mu_k(z), \mu_k^0(z)$: ML price shifts both buyer offers $\mu_k(z)$ and seller belief $\mu_k^0(z)$.

b. $\sigma_S^\tau(\sigma_z), \sigma_B(\sigma_z)$: A small ML noise σ_z lowers seller uncertainty σ_S^0 and variance in buyer offers σ^B. The latter indirectly helps the seller quickly alleviate their uncertainty ($\sigma_S^\tau \rightarrow \sigma_B$).

Using the estimates $\hat{\mu}_k, \hat{\mu}_k^0, (\hat{\sigma}_S^\tau)_j, (\hat{\sigma}_B)_j$ we can identify the sensitivity to ML signals (z, σ_z).

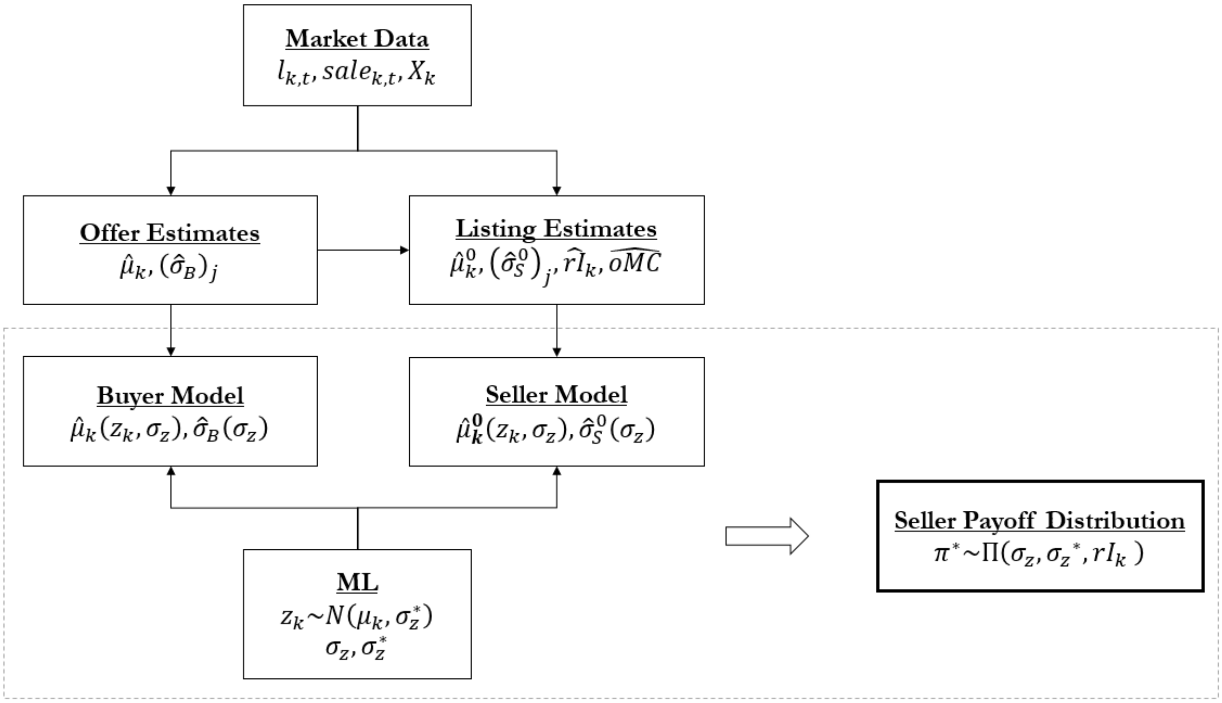

Market Data
$l_{k,t}, sale_{k,t}, X_k$
Offer Estimates
$\hat{\mu}_k, (\hat{\sigma}_B)_j$
Listing Estimates
$\hat{\mu}_k^0, \left(\hat{\sigma}_S^0\right)_j, \widehat{rI}_k, \widehat{oMC}$
Buyer Model
$\hat{\mu}_k(z_k, \sigma_z), \hat{\sigma}_B(\sigma_z)$
Seller Model
$\hat{\mu}_{\boldsymbol{k}}^{\boldsymbol{0}}(z_k, \sigma_z), \hat{\sigma}_S^0(\sigma_z)$
ML
$z_k \sim N(\mu_k, \sigma_z^*)$
σ_z, σ_z^*
Seller Payoff Distribution
$\pi^* \sim \Pi(\sigma_z, {\sigma_z}^*, rI_k\ )$

Figure 10:Market Data is used to estimate Offer Parameters and then Listing Choice Parameters. Then Buyer ($\hat{\mu}_k, \hat{\sigma}_B$) and Seller ($\hat{\mu}_{\boldsymbol{k}}^{\mathbf{0}}$, $\hat{\sigma}_S^0$) parameter sensitivity with ML Signal (z_k, σ_z) is identified. Finally seller's payoff distribution ($\pi^* \sim \Pi(\sigma_z, {\sigma_z}^*, rI_k\ )$) is approximated using monte carlo simulations.

Payoff: The seller enters the market with an expected lifetime value $V(S)$ and follows policy $CCP(a/S)$ to maximize this payoff. But both V computation and CCP policy rely on ML signal noise σ_z presented by the ML platform. Therefore V and CCP are erroneous if the true unconfounded ML signal noise $\sigma_z^* > \sigma_z$. We (econometrician) want to infer the actual seller payoffs as a function of ML noise underestimation (${\sigma_z}^* - \sigma_z$). The inference can be performed using structural estimates described earlier and Monte Carlo simulations over stochastic measures as follows,

[1] $\mu_k \sim N\big(\mu_k, \sigma(\mu_k)\big)$; $z \sim N(\mu_k; {\sigma_z}^*)$

[2] $E\left[\hat{\mu}_k^{(0)}\right] = E\left[\hat{\mu}_k^{(0)}(\mu_k, z)\right]$; $\hat{\mu}_k = \hat{\mu}_k(\mu_k, z)$

[3] $\hat{\sigma}_S^\tau = \hat{\sigma}_S^\tau(\sigma_z)$; $\hat{\sigma}_B = \hat{\sigma}_B(\sigma_z)$

[4] $\hat{\mu}_k^0 \sim N\left(E\left[\hat{\mu}_k^{(0)}\right], \hat{\sigma}_S^\tau(\sigma_z = 0)\right)$

[5] $\pi^* \sim marketSim(\widehat{CCP}, \hat{\mu}^{(0)}, \hat{\mu}, \sigma_S^{(0)}, \sigma_B, rI_k)$

Note that in market simulation to calculate payoff outcomes π^*, we use seller's policy $CCP(a/S)$ which is optimal for his computation of lifetime value V, even if it is suboptimal with respect to true lifetime payoff know to us (econometrician).

The inference yields a probability density function over possible payoff outcomes $\pi \sim \Pi(\sigma_z, {\sigma_z}^*, rI)$. This payoff density function captures both payoff expectation $E[\pi]$ and variance $Var(\pi)$. We expect introduction of ML price to have two main effects - increases $E[\pi]$ via lowered (σ_0^S, σ^B) and a stretched out $Var(\pi)$ due to large σ_z^* but underestimated σ_z. A large $Var(\pi)$ is akin to the seller gambling on a coin toss, instead of a deterministic payoff. This would be undesirable to a risk averse seller. We can use payoff density functions Π under counterfactual ($\sigma_z, {\sigma_z}^*, rI$) settings to answer following questions – (i) How much does a seller gain after introduction of ML? (ii) Under what conditions ($\sigma_z < \sigma_z{}^*$) is a seller worse off after introduction of ML? (iii) How do the gains differ for patient (high outside rental income) and impatient (low outside rental income) seller? (iv) How do the gains differ for risk neutral ($E[\pi]$) and risk averse ($E[\pi] - A * Var(\pi)$) seller? (v) How do the gains differ for seller who has accurate valuation ($\mu_k^{(0)} \approx \mu_k$, e.g. using high quality agents) and sellers who has noisy valuation.

A.3 Proxy Peer Sales Model

Zillow reports a set of 4 or 5 peer house sales alongside Zestimate of every house. We attempt to reverse engineer the choice of these peer sales. We use $isPeer_{i,j,t} = \{0,1\}$ to represent if house sale j is tagged as a peer sale of house i at time snapshot t. Given the characteristics of house i and j, we want to predict if the pair would be tagged as peers. We analyze 1346 houses ($i \in [1,1346]$) in Allegheny county, each observed over 7 time snapshots. In Allegheny county, we have approximately 1400 house sales ($j \in [1,1400]$) in Allegheny county that form candidate peer set for every house. Out of 1400 candidates only 4-5 houses are selected as peer sales for each house.

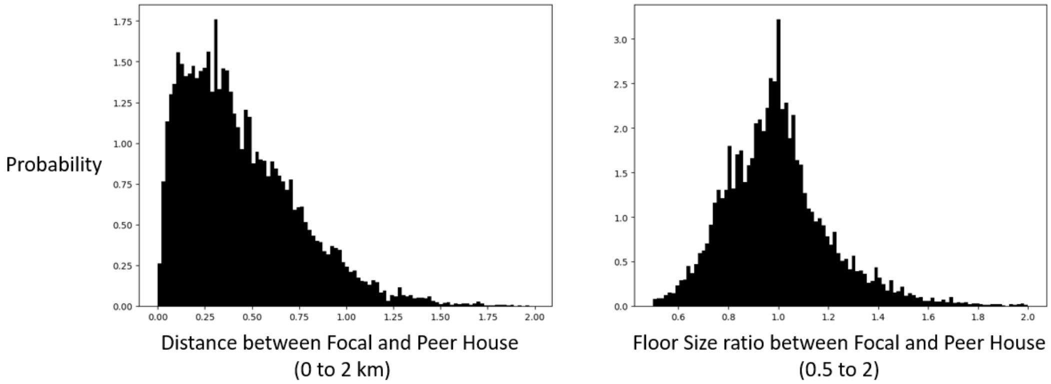


Figure 11: Probability distribution of (Left) geographical distance and (Right) floor size ratio between focal and peer sales. The distribution are empirically calculated using Zillow's reported peer sales for houses in Zip Codes (15210, 15212, 15235) between March and August 2019.

We infer from simple descriptive analysis that a house has a very high likelihood of being tagged as a peer ($isPeer_{i,j,t}$ = 1) if – (i) sold within past 12 months, (ii) it is within 2 km of the focal house and (iii) it has a floor size that is no less then half and no more than double of the focal house.

$$P\left(isPeer_{i,j,t} = 1\right) = \left(t - saleTime_j < 12months\right) \times \left(Dist_{i,j} < 2kms\right)$$
$$\times \left(0.5 < \frac{floorSize_j}{floorSize_i} < 2\right) \times \frac{1}{1 + exp\left(-y_{ijt}\right)}$$

where

$$y_{ijt} \; = \beta_0 + \beta_1 Dist_{i,j} + \beta_2\left(ZipCode_j = ZipCode_i\right) + \beta_3 abs\left(log\left(\frac{floorSize_j}{floorSize_i}\right)\right)$$
$$+\beta_4 abs\left(bedrooms_j - bedrooms_i\right) + \beta_5 abs\left(bathrooms_j - bathrooms_i\right)$$

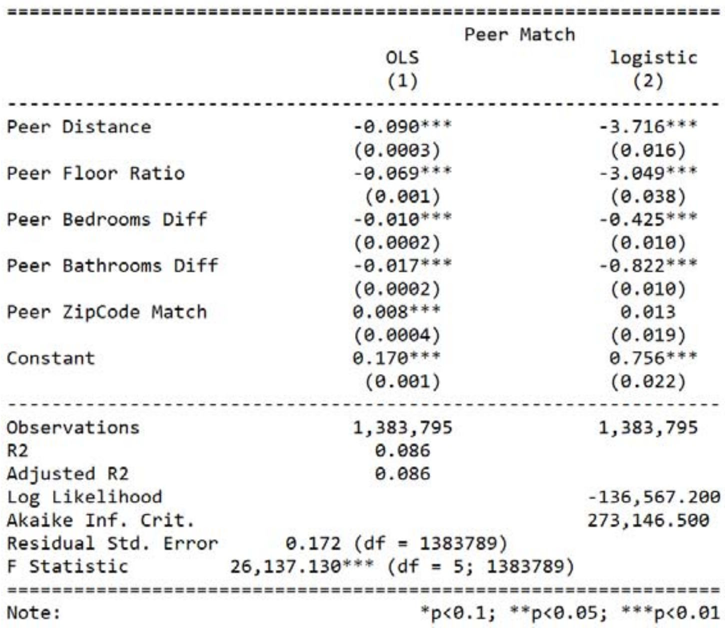
To further predict $isPeer_{i,j,t}$ among houses that satisfy the three criterion above, we use a simple logistic regression model using – distance between house i and j, binary indicator whether the two houses are in the same zipcode, ratio of floor sizes, absolute difference of number of bedrooms and bathrooms. As expected, Table 9 reports that houses in close geographical vicinity (distance and zip code) and similar house features (size, bedrooms, bathrooms) are more likely to be peers. An accuracy of 98.7% and an F1 score of 0.32 (compared with F1 score of 0.02 for a random model) suggest that our predictive model is able to successfully pick out peers of a house.

Table 8: (Top) Confusion Matrix and classification evaluation metrics (TPR, FPR, Accuracy, F1 Score) for the peer prediction model. (Bottom) The logistic regression components y_{ijt} within the overall predictive model.

Confusion Matrix		Actual	
		Peer	Not a Peer
Predicted	Peer	4	16
	Not a Peer	1	1379

True Positive Rate	80%
False Positive Rate	1.14%
Precision	20%
Accuracy	98.7%
F1 Score	0.32

Table 9: Model of whether a house, which satisfies sale time within 12 months, distance within 2 km and floor area between 0.5 to 2 times of focal house, is a peer to the focal house.



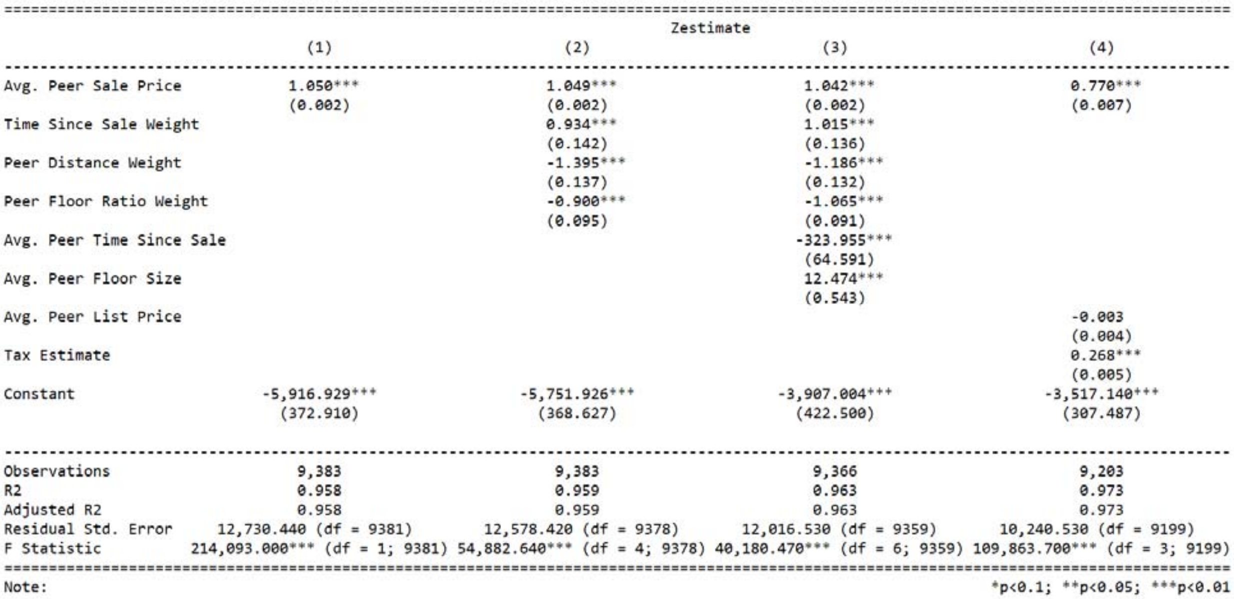
```
==============================================================
                                        Peer Match
                                 OLS                logistic
                                 (1)                  (2)
--------------------------------------------------------------
Peer Distance                 -0.090***            -3.716***
                              (0.0003)             (0.016)
Peer Floor Ratio              -0.069***            -3.049***
                               (0.001)             (0.038)
Peer Bedrooms Diff            -0.010***            -0.425***
                              (0.0002)             (0.010)
Peer Bathrooms Diff           -0.017***            -0.822***
                              (0.0002)             (0.010)
Peer ZipCode Match            0.008***              0.013
                              (0.0004)             (0.019)
Constant                      0.170***             0.756***
                               (0.001)             (0.022)
--------------------------------------------------------------
Observations                  1,383,795            1,383,795
R2                              0.086
Adjusted R2                     0.086
Log Likelihood                                  -136,567.200
Akaike Inf. Crit.                                273,146.500
Residual Std. Error      0.172 (df = 1383789)
F Statistic         26,137.130*** (df = 5; 1383789)
==============================================================
Note:                              *p<0.1; **p<0.05; ***p<0.01
```

A.4 Proxy Zestimate Model

$$z_{i,t} = \beta_0 + \beta_1 \left(\overline{p}_j\right)_{i,t} + \epsilon_{i,t}$$

We propose a very simple and interpretable proxy metric for approximating the Zestimate of a house as - an average of peer sale prices $\bar{p}_j$. Table 10 Model 1 reports that this simple proxy metric explains close to 96% variation in the actual Zestimate. While we are aware that the actual Zestimate model may be significantly more sophisticated, the extremely high out of sample explanatory power ($R^2 \approx 0.96$) suggest that a big chunk of the full model is driven by information contained in peer sales. We iterate over various alternative predictors – (i) weighted (instead of unweighted) average of peer sale prices, (ii) House Features X_i, (iii) Location (Neighborhood, ZipCode, County) fixed effects, (iv) Time fixed effects, (v) Tax Estimate. We first describe the various peer sale weighting strategies and then the remaining predictors.

Table 10: Model of Zestimate using average peer sale price, weighted peer sale prices, average peer floor size, average peer list price and county tax estimate.



```
=================================================================================================================================
                                                                 Zestimate
                                  (1)                            (2)                            (3)                           (4)
---------------------------------------------------------------------------------------------------------------------------------
Avg. Peer Sale Price            1.050***                       1.049***                       1.042***                      0.770***
                                (0.002)                        (0.002)                        (0.002)                       (0.007)
Time Since Sale Weight                                         0.934***                       1.015***
                                                               (0.142)                        (0.136)
Peer Distance Weight                                           -1.395***                      -1.186***
                                                               (0.137)                        (0.132)
Peer Floor Ratio Weight                                        -0.900***                      -1.065***
                                                               (0.095)                        (0.091)
Avg. Peer Time Since Sale                                                                    -323.955***
                                                                                              (64.591)
Avg. Peer Floor Size                                                                          12.474***
                                                                                              (0.543)
Avg. Peer List Price                                                                                                        -0.003
                                                                                                                            (0.004)
Tax Estimate                                                                                                                0.268***
                                                                                                                            (0.005)
Constant                      -5,916.929***                  -5,751.926***                  -3,907.004***                 -3,517.140***
                                (372.910)                      (368.627)                      (422.500)                     (307.487)

---------------------------------------------------------------------------------------------------------------------------------
Observations                      9,383                          9,383                          9,366                         9,203
R2                                0.958                          0.959                          0.963                         0.973
Adjusted R2                       0.958                          0.959                          0.963                         0.973
Residual Std. Error        12,730.440 (df = 9381)        12,578.420 (df = 9378)        12,016.530 (df = 9359)        10,240.530 (df = 9199)
F Statistic           214,093.000*** (df = 1; 9381) 54,882.640*** (df = 4; 9378) 40,180.470*** (df = 6; 9359) 109,863.700*** (df = 3; 9199)
=================================================================================================================================
Note:                                                                                              *p<0.1; **p<0.05; ***p<0.01
```

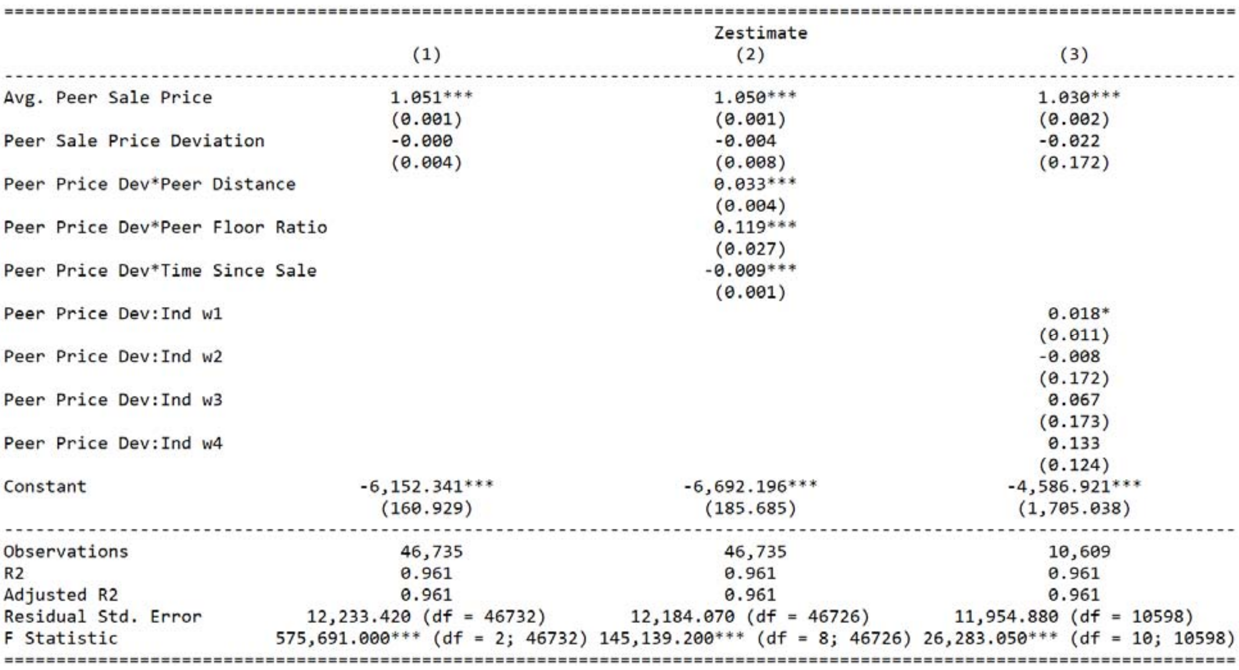
$$z_{i,t} = \beta_0 + \beta_1 \left(\overline{p}_j\right)_{i,t} + \beta_2 \left(p_k - \left(\overline{p}_j\right)_{i,t}\right) * \overrightarrow{w_{\iota k}} + \xi_{i,k,t}$$

We test if deviation of a peer sale price p_k from average peer sale price $\bar{p}_j$ weighted by $\overrightarrow{w_{\iota k}}$ has a significant correlation with Zestimate $z_{i,t}$. We investigate following peer weighting,

- Peer Distance $w_{ik}^{(1)} = Dist_{ik}$: A house closer to the focal house should have a higher weight in determining focal house Zestimate.
- Peer Floor Size Ratio $w_{ik}^{(2)} = abs(log(floorSize_k/floorSize_i))$: A house with floor size closer to the focal house should have a higher weight in determining focal house Zestimate.
- Time since sale: $w_{ikt}^{(3)} = (t - saleTime_k)$: A house sold more recently should have a higher weight in determining focal house Zestimate.
- Degree of Peer Independence : A house which contains more independent (and less correlated information) should have a higher weight in determining focal house Zestimate. Consider a focal house i, and its peers given by $S(i)$. House $k \in S(i)$ is one of these peer houses. Houses $S(i) - k$ are the remaining peer houses. $S(k)$ are peer houses that determined Zestimate of house k when it was listed in the market. Similarly, $S(S(i) - k)$ are peer houses that determined Zestimate of houses $S(i) - k$ when they was listed in the market.

- $w_{ik}^{(4)} = \Sigma_{\mathrm{j} \in S(i)} Dist_{jk} / |S(i)|$: A house k that is farther away from all other peers ($S(i) - k$) of the focal house i adds more independent information. It should have a higher weight in determining focal house Zestimate.
- $w_{ik}^{(5)}$: A house k which is peer for other peers of the focal house $k \in S(S(i) - k)$ adds less independent information. It should have a lower weight in determining focal house Zestimate.
- $w_{ik}^{(6)}$: A house for whom other peers of focal house acted as peers ($S(i) - k \cap S(k) \neq \phi$) adds less independent information. It should have a lower weight in determining focal house Zestimate.
- $w_{ik}^{(7)}$: A house whose peers intersect with peers of other peers of focal house ($S(S(i) - k) \cap S(k) \neq \phi$) adds less independent information and it should have a lower weight in determining focal house Zestimate

Table 11: Model of Zestimate using different weighted peer sale prices.



```
=====================================================================================================
                                                          Zestimate
                                      (1)                    (2)                      (3)
-----------------------------------------------------------------------------------------------------
Avg. Peer Sale Price               1.051***               1.050***                 1.030***
                                   (0.001)                (0.001)                  (0.002)
Peer Sale Price Deviation          -0.000                 -0.004                   -0.022
                                   (0.004)                (0.008)                  (0.172)
Peer Price Dev*Peer Distance                              0.033***
                                                          (0.004)
Peer Price Dev*Peer Floor Ratio                           0.119***
                                                          (0.027)
Peer Price Dev*Time Since Sale                           -0.009***
                                                          (0.001)
Peer Price Dev:Ind w1                                                               0.018*
                                                                                   (0.011)
Peer Price Dev:Ind w2                                                              -0.008
                                                                                   (0.172)
Peer Price Dev:Ind w3                                                               0.067
                                                                                   (0.173)
Peer Price Dev:Ind w4                                                               0.133
                                                                                   (0.124)
Constant                        -6,152.341***          -6,692.196***            -4,586.921***
                                  (160.929)              (185.685)                (1,705.038)
-----------------------------------------------------------------------------------------------------
Observations                       46,735                 46,735                   10,609
R2                                  0.961                  0.961                    0.961
Adjusted R2                         0.961                  0.961                    0.961
Residual Std. Error         12,233.420 (df = 46732)  12,184.070 (df = 46726)  11,954.880 (df = 10598)
F Statistic            575,691.000*** (df = 2; 46732) 145,139.200*** (df = 8; 46726) 26,283.050*** (df = 10; 10598)
=====================================================================================================
```

We find the $w_{ikt}^{(1)}, w_{ikt}^{(2)}, w_{ikt}^{(3)}$ to be statistically significant. A peer sale 2 km from focal house is weighted 6.8% more then an adjacent peer sale. A peer sale with double or half the floor size of the focal house is weighted 8.6% more then a peer sale of same size. A peer sale 12 months from present is weighted 11% less then a peer sale in the current month. We use $W_{ikt} = \left[w_{ikt}^{(1)}, w_{ikt}^{(2)}, w_{ikt}^{(3)}\right]$ in a Kernel regression to predict Zestimate of the focal house.

$$\hat{z}_{i,t} = \frac{\sum_k p_k * K(x_k - x_i; \delta)}{\sum_k K(x_k - x_i; \delta)} \qquad ; \qquad K(x_k - x_i; \delta) = exp(-\delta * W_{ikt})$$

$$\hat{\delta} = [0.32, 1.6, 0.042]$$

Table 11:Alternative Models and features to predict Zestimate

Features	Out of Sample R^2	
	Linear Model	Support Vector Model
Average Peer Sale Price only	**96.15**	95.97
+ Kernel Weights	96.59	96.61
+ All other covariates	96.79	97.34

We find that additional features (House Features, Location fixed effects, Time fixed effects, and Tax Estimate do no contain significant Zestimate explanatory power on their own. Table 11 shows that all these features and weighted peer sale price add very little to the explanatory power when used alongside the simple average peer sale price $\bar{p}_j$ metric.

Tables and Figures

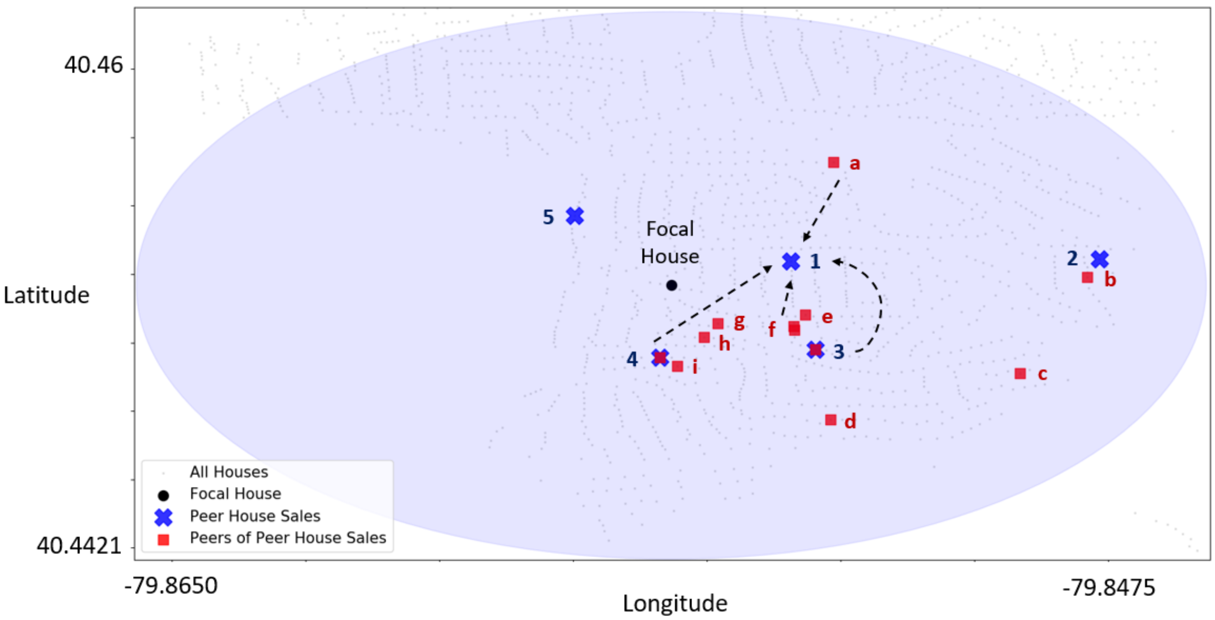


Figure 12: (Top) A zoom in view of 1.5 km radius around the focal house. Each of the five peer houses (1,2,3,4,5 blue **X***) had a Zestimate when they were on market. Those Zestimate were determine by their own set of peers (a,b,…,h,i red ■).*

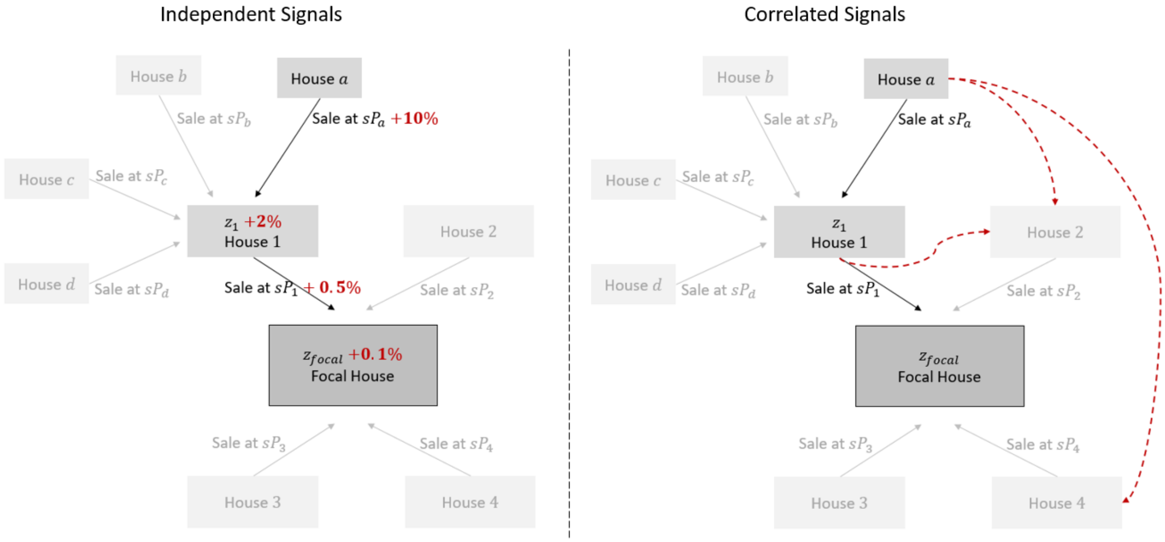


Figure 13: (Left) Zestimate for the focal house depends on sale price of peers house 1, 2, … etc. The sale price of peer house 1 depends on the zestimate received by house 1. This zestimate in turn depends on sale price of its peers house a, b, … etc. (Right) Zestimate for the focal house depends on sale price of peers house 1, 2, … etc. But house 1 and house 2 sale prices are not independent signals. House 1 drives Zestimate of House 2 and House a drives Zestimates of both house 1 and house 2.

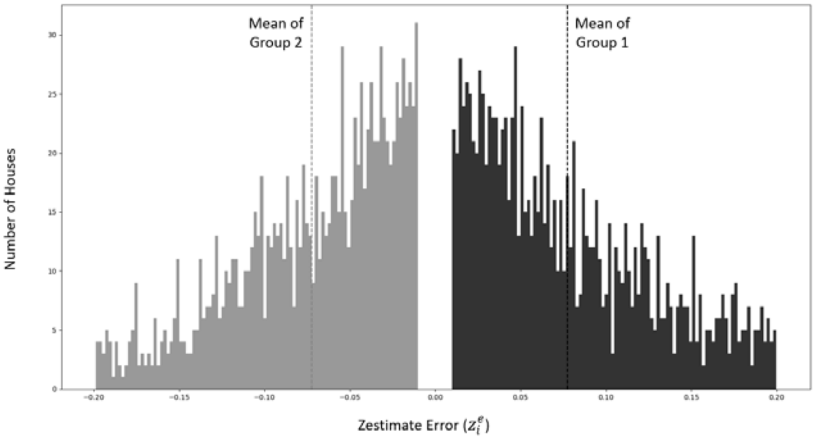


Figure 14: Two groups classified on positive error ($z_i^e > 1\%$) and negative error ($z_i^e < -1\%$) for propensity score model of Zestimate impact on List and Sale Prices.

Table 12: Percentage Standardized Bias for all matching covariates before and after matching for propensity score model of Zestimate impact on List and Sale Prices

Covariate	**Post Bias**	**Pre Bias**
Z (post algorithm update)	-0.01	-11.01
Abs(Z Error)	4.46	-32.65
Last Remodel Year	2.3	-9.65
Solar Potential	-0.14	-1.96
Lot Size	0.03	-3.74
Bedrooms	0.24	-14.34
Month of Listing	-1.49	9.48
Number of stories	-0.35	4.45
Parking	2.41	-21.85
Bathrooms	0.87	-16.45
Year Built	0.74	-19.24
Floor Size	0.93	-18.27
Zipcode (15237)	0.87	3.32
Zipcode (78704)	-0.95	-7.87
Zipcode (78745)	0.64	3.01
County (Travis)	-0.26	-19.4
County (Allegheny County)	-2.25	11.21
County (Suffolk)	2.47	16.8
Nbrhood (South Boston)	1.95	8.76
Nbrhood (Other)	0.16	3.08
RoofType (Composition)	-0.37	-15.98
RoofType (Shake/Shingle)	-1.53	7.91
RoofType (Asphalt)	-0.48	3.2
StructureType (Unreported)	0.32	-12.6
StructureType (Other)	-0.06	9.57
StructureType (Colonial)	-0.86	6.04
Flooring (Other)	1.1	-8.43
Flooring (Unreported)	-0.42	13.18
Flooring (Hardwood)	-1.04	3.47
Patio (Unreported)	-1.05	17.26
Patio (Porch, Patio)	-1.44	-11.05
Patio (Other)	0.68	-2.12
ExMaterial (Brick)	-0.83	6.74
ExMaterial (Unreported)	-0.4	-1.22
ExMaterial (Other)	-0.66	-1.16
Type (Single Family)	0.18	-13.89
Type (Condo)	0.47	12.38
Type (Multi Family)	-0.34	3.28

Table 13: Descriptive Statistics for key house features, Zestimate, List and Sale Prices for propensity score model of Zestimate impact on List and Sale Prices

Variable	Mean	Std. Dev	Min	Max
Z (pre algorithm update)	433737.6	393379.4	25065	8244491
Z (post algorithm update)	433321.7	382211.1	21226	7911868
Z Error %	0.109	0.09	-20	20
Z Confidence Interval %	17.3	10.34	10	130
Sale Price	387602.5	376852.3	13000	4950000
List Price	438729.8	390021	4800	6675000
Markup %	1.08	0.12	-9.55	22.5
Time to Sale	27.8	47.94	1	350
Floor Size	1768.6	831.3	293	6000
Year Built	1961	37.56	1799	2019
Bathrooms	2.3	0.96	1	14
Bedrooms	3.1	1.2	1	12
Parking	156.6	220.55	0	995
Lot	4396.4	5113.88	293	9000
Stories	1.9	2.39	0	25
Last Remodel Year	1973.2	39.91	0	2019
Solar Potential	73	25.47	0	95.66

Figure 14: Four groups classified on no change, 0-5% change, 5-10% change and 10-15% change in average peer sale price for propensity score model for impact of average peer sale price change on Zestimate.

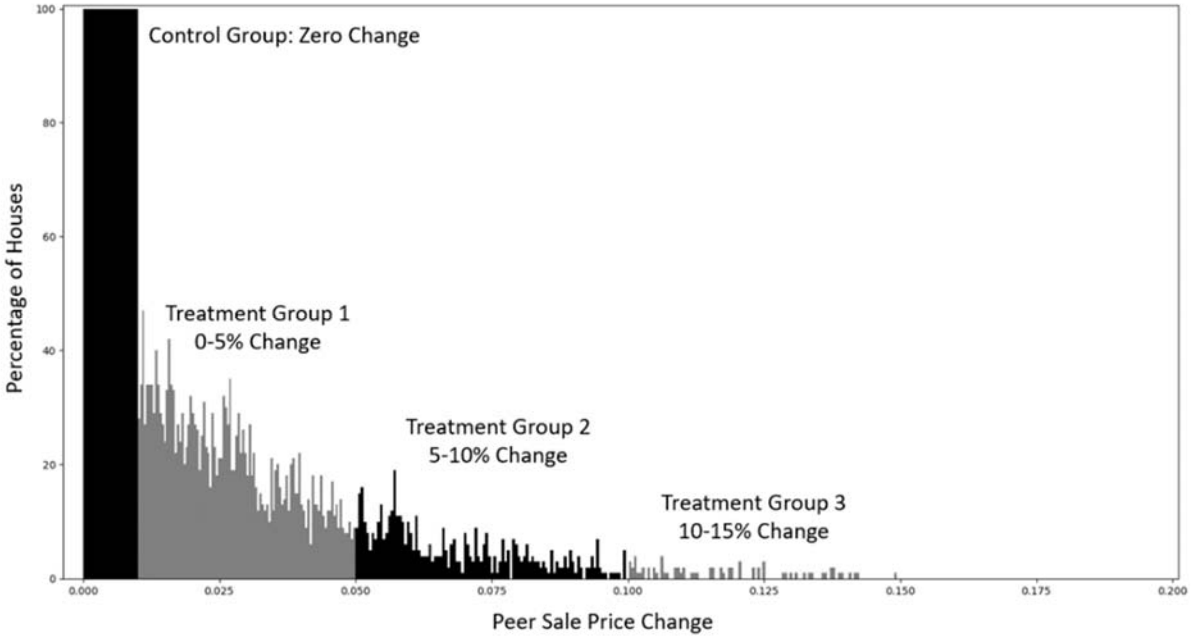


Table 14: Descriptive Statistics for key house features, Zestimate, List and Sale Prices for propensity score model for impact of average peer sale price change on Zestimate.

Variable	Mean	Std. Dev	Min	Max
Z (t1)	155351.5667	61937.24	14716	653757
Z (t2)	155461.1689	62407.52	14716	657858
Z change %	0.004	3.88	-45	54.1
Peer Sale Price (t1)	153689.7778	57805.42	13704	481000
Peer Sale Price (t2)	153742.7283	58189.89	13704	481000
Peer Sale Price change %	-0.04	5.38	-59.2	97.1
Peer List Price (t1)	155413.3021	66661.28	21459.9	617000
Peer List Price (t2)	155219.6595	66456.15	21459.9	617000
Peer List Price change %	-0.136	9.78	-42.2	42
Tax Estimate (t1)	151884.0816	65303.19	10427	649119
Tax Estimate (t2)	152919.8321	65688.55	10427	649119
Tax Estimate change %	-0.67	0.72	-1.2	4
Floor Size	1583.139891	460.4352	828	3840

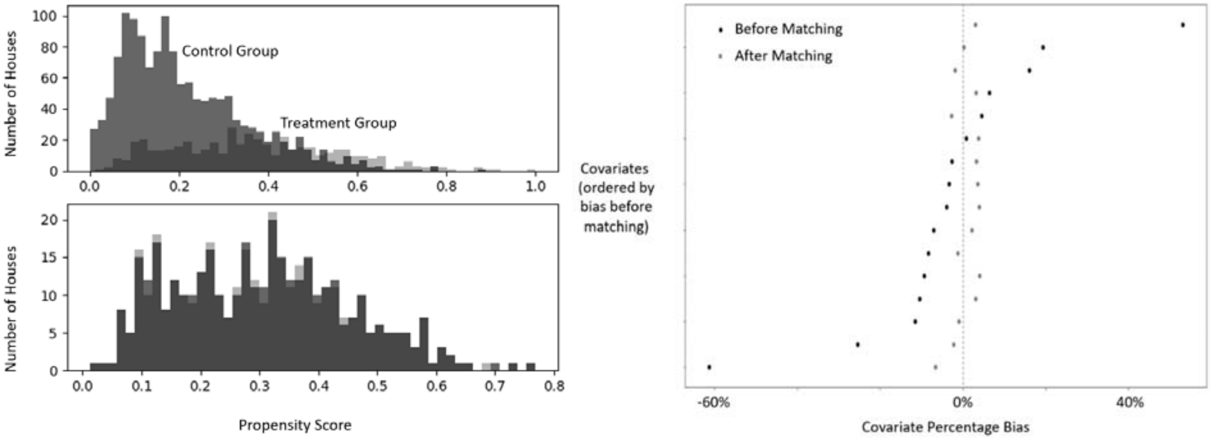


Figure 15: (Left) Histogram of propensity score for treatment and control groups before and after PSM. (Right) Percentage Standardized Bias for matching covariates before (dark) and after (light) PSM for impact of average peer sale price change on Zestimate

Table 15: Treatment and Control statistics before and after PSM. A Rubin's R between 0.5-2.0 and a Rubin's B less than 25 indicate a good match.

Statistic	Before Matching	After Matching
Mean % Std. Bias	15.24%	2.88%
Rubin's R	1.68	1.00
Rubin's B	93.44%	0.14%

Table 16: Percentage Standardized Bias for all matching covariates before and after matching.

Covariate	**Post Bias**	**Pre Bias**
Z (t1)	2.16	-7.09
Avg. Peer Sale Price (t1)	3.99	-9.36
Tax Estimate (t1)	3.59	-3.35
Change in Tax Estimate	3.02	53.11
Avg. Peer List Price (t1)	3.17	6.34
Change avg. in Peer List Price	0.16	19.29
Floor Size	3.27	-2.66
Neighborhood (others)	-2.77	4.53
Neighborhood (BrightonHeights)	3.73	0.8
Neighborhood (Carrick)	-1.25	-8.4
Time FE (6)	3.94	-3.97
Time FE (3)	-2.24	-25.46
Time FE (7)	-6.71	-61.37
Zipcode (15235)	-1.96	16.05
Zipcode (15212)	3.1	-10.45
Zipcode (15210)	-1.02	-11.58