FACULTY
OF SCIENCES

FACULTY OF ECONOMICS AND
BUSINESS ADMINISTRATION

# Addressing socioeconomic challenges with micro-level trace data

GHENT
UNIVERSITY

Milan van den Heuvel

FACULTY OF SCIENCES

FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

# Addressing socioeconomic challenges with micro-level trace data

Supervisor: Prof. Dr. Koen Schoors
Co-supervisor: Prof. Dr. Jan Ryckebusch

Dissertation submitted in fulfillment
of the requirements for the degree of:
Doctor of Science: physics and economic sciences

Ghent University
Faculty of Sciences:
Department of Physics and Astronomy
Faculty of Economics and Business Administration:
Department of Economics
Academic year 2019-2020

GHENT UNIVERSITY

Milan van den Heuvel

# Dankwoord

tl;dr: Thanks to all the people that ever listened to my ramblings, guided me, made me laugh, and helped me think through things!

De laatste vier jaar waren de interessantste en meest intellectueel variërende jaren uit mijn leven. Ik werd gekatapulteerd van interbancaire netwerken in Rusland, naar een virtuele wereld in IJsland, tot diep in de bankensector in ons eigenste landje. Dit was nooit gelukt zonder de nodige begeleiding en hulp van de vele geweldige mensen rondom mij.

Eerst en vooral wil ik Jan en Koen bedanken, die mij toelieten om al vanaf mijn Masterthesis de interdisciplinaire toer op te gaan en later het potentieel in mij zagen om het groepje LAB-M'ers te vervoegen. Jan, bedankt om een geweldige lesgever te zijn. Uw lessen statistische en computationele fysica zullen mij altijd bijblijven en zijn de reden dat complexe systemen mij mateloos blijven interesseren. Ook bedankt voor het vertrouwen en de vrijheid om mij telkens mijn eigen spin te laten geven aan de onderwerpen die we onder de loop namen maar om mij steevast te herinneren aan de natuurkundige grondslagen waarover we spraken. De dagen waarin we samen onze paper volledig hebben herschreven zitten zeker bij de meest leerrijke en productieve dagen van mijn doctoraat. Koen, bedankt om mij wegwijs te maken in de wereld van de economie, om mij het belang van context en verhaal te tonen en voor de vele kansen die je me hebt gegeven om bij te leren. Ik ben naar veel geweldige plekken mogen reizen om ons onderzoek beter te maken, bedankt om dat steeds aan te moedigen.

Naast Jan en Koen had ik eigenlijk ook nog een derde (onofficiële) promotor, een man die na verloop van tijd mijn tred van ver kon herkennen als ik weer eens iets kwam vragen, aftoetsen, of gewoon eens wou praten. Benjamin, bedankt om vanaf de dag dat ik aan mijn masterthesis begon, tot de dag dat ik dit schrijf klaar te staan met je wijze raad, je nuchtere kijk, of gewoon met een paar aanmoedigende woordjes. Ik ben ervan overtuigd dat het vele babysitten op mij goede training was om een geweldige vader te zijn voor Olivia.

Daarnaast wil ik de rest van ons LAB-M team bedanken voor hun vele inspanning om onze projecten tot een goed eind te doen brengen. Kevin, bedankt om altijd tijd te maken om je uitgebreide kennis over programmeren, EVE Online, je unieke interesses, en je kijk op de wereld te delen. Andres, thanks for the many times you helped me process data. Without you, there would have been a couple of master students without a degree ;). Ken, bedankt om mij binnen te loodsen in de bankenwereld, voor het delen van uw levenswijsheid, en de vele toffe lunches in Brussel. Zonder uw goedlachse persoonlijkheid had ik dat project niet overleefd. Tarik, I have never met someone who can structure and push projects forward the way you so effortlessly do. Thanks for trying to teach me to structure my thought before I start rambling on about something, I am still working on it! Marnix, merci om mijn thesisstudent te zijn en om er daarna samen een mooie paper van te maken, ik kijk er naar uit om te zien wat je nog allemaal gaat doen.

Dan was er nog mijn intrek in de faculteit economie, en alle geweldige mensen die me daar geholpen hebben, van het beantwoorden van mijn vele economische vraagjes tot het zorgen voor ontspanning op de vele zotte afterwork feestjes. Bram en Ruben, aka brambo en el blanco, bedankt om op het einde mijn kantoorgenoten te zijn, om mij altijd aan het lachen te brengen, en vooral om mij telkens weer te doen uitkijken om op kantoor te gaan werken. Tom, het was een plezier samen te vloeken over hoeveel werk Big Data Science wel niet was. Ook bedankt om samen met Sümeyra de laatste jaren mijn kantoorburen te zijn en mijn vele vraagjes te beantwoorden, waaronder voornamelijk de vele vragen over de locatie van Koen. Willem, Désirée, Sam, Jolien, Lana, Brecht N., Brecht B., Amy, Jérome, Ruben of ook bekend als de oude kern van de “Economics Partycrew”, bedankt voor de vele geweldige avonden, dat we er nog veel mogen meemaken samen! Verder wil ik nog alle andere collegas, de vaste waarden op de Friday drinks, mijn collegas op het INW (ook wel “De fysica”), mijn (voormalige) huisgenoten Sego en Michiel et al. ;), en mijn vrienden uit Turnhout bedanken voor de ontspanningsmomentjes en om van mijn doctoraat een geweldige ervaring te maken. In het speciaal wil ik Jannes bedanken om mij te fascineren door zijn geweldige machine learning kennis en de vele stimulerende gesprekken. Je hebt me doen inzien waar ik mij binnenkort mee wil gaan bezighouden. Ook wil ik in het bijzonder Martine, Eva, en Sabine nog even in de bloementjes zetten om altijd klaar te staan met de hoognodige hulp voor alle administratie. I also want to thank the jury for being willing to read and comment on this dissertation.

Tot slot wil ik nog graag mijn ouders en zus bedanken. Mijn ouders om mij alle kansen en liefde te geven die ik mij maar kon wensen, voor hun engelengeduld als ik mij verloor in mijn studies en werk, en voor hun onvoorwaardelijke steun gedurende de laatste 27 jaar. Mijn (grote) zus, voor de dure daguitstapjes, de leuke etentjes, en om er altijd te zijn. Ook al wou ik het zelf niet altijd toegeven, die dagen ontspanning waren hoognodig.

*Gent, september 2019*
*Milan van den Heuvel*

# Table of Contents


**Dankwoord** **i**

**1 Introduction** **1**
1.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1
1.1.1 The modern equivalent of the microscope . . . . . . . . . 1
1.1.2 Data: challenges and opportunities . . . . . . . . . . . . . 4
1.1.3 A hunt for data . . . . . . . . . . . . . . . . . . . . . . . 7
1.2 Outline . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

**I The Russian interbank market** **19**

**2 Loan maturity aggregation in interbank lending networks obscures mesoscale structure and economic functions** **21**
2.1 Schematic overview . . . . . . . . . . . . . . . . . . . . . . . . 23
2.2 Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
2.3 Introduction and overview . . . . . . . . . . . . . . . . . . . . . 24
2.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
2.4.0.1 Characterisation of the OGBs . . . . . . . . . . 41

**Appendices** **47**
2.A Maturity layer analysis . . . . . . . . . . . . . . . . . . . . . . . 47
2.A.1 Activity and volume . . . . . . . . . . . . . . . . . . . . 47
2.A.2 Maturity layer activity of banks . . . . . . . . . . . . . . 50
2.A.3 Density . . . . . . . . . . . . . . . . . . . . . . . . . . . 51
2.A.4 Degree distribution . . . . . . . . . . . . . . . . . . . . . 51
2.A.5 Average shortest path length . . . . . . . . . . . . . . . . 56
2.A.6 Clustering coefficient . . . . . . . . . . . . . . . . . . . . 59
2.A.7 Bank size mixing as degree size mixing . . . . . . . . . . 62
2.A.8 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . 63
2.B OG inference with non-contiguous binning . . . . . . . . . . . . 64
2.C Number of active banks per group . . . . . . . . . . . . . . . . . 66
2.D Activity characteristics per OGB . . . . . . . . . . . . . . . . . . 67
2.E Other percentiles for instrength/strength of important banks . . . . 68

2.F Timeline of the announcement crisis . . . . . . . . . . . . . . . . 70
2.G Maturity classes in the ILN literature . . . . . . . . . . . . . . . . 72
2.H Term structure of interest rates and loan volumes in the interbank loan market . . . . . . . . . . . . . . . . . . . . . . . . . . . 73
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 77

**II Belgian client data of a large European bank 85**

**3 The wealth origins of income mobility 87**
3.1 Schematic overview . . . . . . . . . . . . . . . . . . . . . . . . . 89
3.2 Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 90
3.3 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 90
3.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 95
3.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104

**Appendices 107**
3.A Detailed cleaning procedure . . . . . . . . . . . . . . . . . . . . . 107
3.B Summary statistics . . . . . . . . . . . . . . . . . . . . . . . . . 110
3.C Effects of income thresholds . . . . . . . . . . . . . . . . . . . . 113
3.D Extended regression tables . . . . . . . . . . . . . . . . . . . . . 118
3.E Sanity checks . . . . . . . . . . . . . . . . . . . . . . . . . . . . 122
3.F Robustness checks . . . . . . . . . . . . . . . . . . . . . . . . . 125
3.F.1 Trimming . . . . . . . . . . . . . . . . . . . . . . . . . . 125
3.F.2 Income definition . . . . . . . . . . . . . . . . . . . . . . 130
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 135

**4 Liquid wealth heterogeneity, asymmetric consumption dynamics, and myopic loss aversion 141**
4.1 Schematic overview . . . . . . . . . . . . . . . . . . . . . . . . . 143
4.2 abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 144
4.3 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 144
4.4 Data and empirical strategy . . . . . . . . . . . . . . . . . . . . . 148
4.5 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 153
4.6 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 156

**Appendices 157**
4.A Representativeness of the data . . . . . . . . . . . . . . . . . . . 157
4.B Robustness check . . . . . . . . . . . . . . . . . . . . . . . . . . 157
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 159

**III A multi-faceted view on behavior in a virtual world 163**

**5 Revealed behavior from a virtual world: the case of EVE Online 165**
5.1 Schematic overview . . . . . . . . . . . . . . . . . . . . . . . . . 166

5.2 A live feed into socioeconomic behavior . . . . . . . . . . . . . . . 167
5.3 Warping into the EVE Online universe . . . . . . . . . . . . . . . . 168
5.4 Collecting data wearing ice cleats . . . . . . . . . . . . . . . . . . . 170
5.5 Realized, ongoing, and future projects . . . . . . . . . . . . . . . . 171
5.5.1 Realized projects . . . . . . . . . . . . . . . . . . . . . . . . 171
5.5.2 Ongoing projects and the future . . . . . . . . . . . . . . . 172
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 174

**6 Summary and outlook** **177**
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 179

**7 Samenvatting** **181**

# 1
# Introduction

The work presented in this dissertation is part of an emerging specialism within economics. It can be positioned as a subfield within Applied Empirical Economics that leverages the data originating from the recent trend in digitalisation across public and private sectors. Herein, the focus is on employing new types of datasets to provide novel perspectives on enduring questions and to test hypotheses previously only existing in the realm of theory. This shift towards more empirical, data-driven styles of research has been well-documented [1] and was found to be occurring across rather than purely within multiple economic fields [2]. Apart from the types of data used and the new methodological skills necessary to handle and process such data, an epistemological theme stands central in this specialism, complementing that of classic, theory-driven economics [3].

In the forthcoming section of this introductory chapter we present a motivation for this specialism, the opportunities and challenges this offers to economic research, and set the stage for the ideas, concepts, and data sources used throughout the dissertation. In Section 1.2 we then sketch the outline of the upcoming chapters.

## 1.1 Introduction

### 1.1.1 The modern equivalent of the microscope

Economists have always strove to capture and understand how people behave in a plethora of socioeconomic situations (e.g. income changes, policy interventions, job search). The desire of economists, interested in individual socioeconomic behavior, to progressively uncover new perspectives and details on enduring questions needed to be matched with the proper resources. As Prof. Alberto Cavallo (Sloan School of Management MIT, Harvard Business School, NBER) put it: *"We [economists] have been using the same datasets over and over again, and since we wanted new answers, we have been developing new econometric techniques to try to transform the data, and get more meaningful information out of them. But it was reaching a point where there is nothing else you can do on that side, and just having a fresh, new dataset brings a whole new perspective, and I think people are starting to realise that, and gradually people are becoming more interested in data collection itself."* [3] With classic surveys-based datasets for instance, it is notoriously difficult to collect consistent measurements over time for the same population [4] (e.g. reporting and recollection biases, motivations for misreporting, etc.). Even those that do are often limited in scale and can only cover a narrow scope to get sufficient response rates. This limits the research possibilities and caused the need for alternative data sources.[1]

Since the digitalisation of the public and private sector, collection of ever more extensive and intensive individual level data became commonplace. This trend gave rise to a new type of dataset. While the literature is littered with varying terms that all refer to such data in one form or another, e.g. Big Data, client data, behavioral data, we choose to use the term *"Trace Data"* here. The term *Trace Data* was chosen to definitively contrast the data we use in this dissertation from often used survey, or self-reported data, and is defined as: *"data originating from the natural usage of (digital) products or services of which the collection does not interfere with the natural flow of behaviour and events in the given context"*. The inherent character and the way *Trace data* is recorded has several key advantages over classic types of (often self-reported) data on the side of the subjects. The data represent objective measurements, hence there is no reporting or recollection bias, no opportunity nor motivation to under- or over-report. There is also no contagion between different variables because of priming or other psychological factors during reporting [5, 6]. These features increase the reliability and ability to compare results across populations. We emphasise here that while such data are recorded in a consistent and unbiased matter, this does not exclude them from typical data limitations such as selection bias, recording errors, etc. Digital data also comes with new issues such as internet bots, spatial clustering, and censoring. For instance, when using client-data from a specialized financial website, the risk exists that the user-base is not representative of the entire population. A recent example is the use of Financial Aggregator Platforms (FAP) to study consumption dynamics. While the collection method solves recollection biases and eliminates any potential mo-

[1] We note that the Nordic countries form an exception and have been collecting and providing administrative data to researchers (and often the general public) for many years.

tivations to misreport in surveys, concerns remain about selection bias [7]. This because it is thought that users of FAPs are generally more financially literate than the population at large.

Realizing the insights economists are after align with their own interest, firms and governments have been granting more and more researchers access to their data. An apt analogy was given by Erik Brynjolfsson, who likened the availability of such data to the invention of the microscope: *"To grasp the potential impact of Big Data, look to the microscope. The microscope, invented four centuries ago, allowed people to see and measure things as never before - at the cellular level. It was a revolution in measurement... Data measurement is the modern equivalent of the microscope. Google searches, Facebook posts and Twitter messages for example make it possible to measure behaviour and sentiment in fine detail and as it happens."* [8].

With the magnitude and detail of these new datasets comes a word of caution. While more granular data is undoubtedly a benefit, a lot of noise often resides in the measurements of such detail. Not unlike physics, it is important to keep in mind the scale relevant for the research question at hand. Just as one would not investigate the motion of celestial bodies with quantum electrodynamics, it would become unnecessarily complex to investigate interactions between countries by using data on individual inhabitants. Adding to this the finitude of data raises another issue. Data Aggregation might hide features and patterns, but selecting a overly specific subsample might spread the data thin and weaken the power of any statistical inference that can be done. Before doing any statistics, a researcher must thus be aware of which granularity is appropriate/most informative. This will be illustrated in Chapter 2.

Following [9], we note four features of *Trace data*: (1) data is often available in (near) real-time, (2) data is available at a larger scale, (3) data is available on previously unmeasured activities, and (4) data come with less structure. Our three sets of *Trace Data* are on banks in the Russian interbank loan market, Belgian clients of a large European bank, and players in a massive virtual world. With the exception of the Russian interbank data (Chapter 2), which has a regulatory origin and is on bank behavior, these datasets can be thought of as a by-product of everyday actions and behavior of individuals. Like a lot of *Trace datasets*, they are not generated with a specific research question in mind but rather originate from normal usage of a service or product, for instance from using a bank account (Chapter 3 and 4), or participating in a virtual world (Chapter 5). Such data has also been coined 'data in the wild' by Einav and Levin [3] and require a different approach to formulating research question as we will explain below.

Like most economic research currently, we approach research questions retrospectively, hence our research does not utilize feature (1). However, all datasets presented here could be updated in (near) real-time. In Chapter 5 we present some of the opportunities this would bring about.

### 1.1.2 Data: challenges and opportunities

The features that *Trace data* possess have led to many new opportunities and challenges for economists.

Many examples of novel types of questions and research designs can be found that exploit the different aspects of the data. Having the ability to regularly update the data's time frame (on a daily or an even shorter basis) has facilitated the construction of more precise economic indicators [10]. The larger scale of the data, often even covering a large portion of the population, has allowed researchers to perform detailed explorations of the cross-sectional and temporal variance [11, 12], even for very specific sub-populations [13–15]. Lastly, the availability of previously difficult to observe activities (e.g. geolocation data, item-specific consumption behavior, etc.) has brought about new theories and insights [16, 17], previously impossible to test.

New opportunities also bring about challenges. Since the private sector is weary of making their data public, simply obtaining access is a non-trivial challenge involving cumbersome meetings and elaborate contract negotiations. The data used throughout the chapters, detailed below, is no exception and is subject to NDA agreements and strict access contracts. Once obtained, as per feature (4), the data often comes with less or even no structure. Often a file dump from several sources, in different formats, with different legacy structures are provided that do not follow any standard formatting.

While economists have always been sophisticated data users, the sheer size of these data and the computational complexities that come with them sometimes hit the limits of the statistical packages traditionally used in econom(metr)ics. For instance, the STATA software package, well-known among economists, currently requires loading the complete dataset into memory to do operations or analytics, something that is often infeasible when working with *Trace data*. Recent computational developments such as NoSQL databases, distributed computing pipelines, and online or iterative approaches to calculations, are becoming more and more a part of the standard economic toolbox. Hal R. Varian, professor at UC Berkley and chief economist at Google says in his paper *"Big Data: New Tricks for Econometrics"* [18]: *"In this essay, I will describe a few of these tools for manipulating and analyzing big data. I believe that these methods have a lot to offer and should be more widely known and used by economists. In fact, my standard advice to graduate students these days is go to the computer science department and take a class in machine learning."*. As a result, it will become a valuable skill for future economists to be comfortable with these new data science tools to clean, standardize, and explore ever-larger amounts of unstructured data.

We would also like to mention the issue of privacy, which received a lot of attention in Europe with the new GDPR directive [19][2]. One of the challenges with these detailed datasets is that, with growing size and detail, they become more difficult to anonymize. This puts the privacy of the individuals contained in the data at risk. For instance de Montjoye *et al.* [20] showed for a dataset of 1.1 million individuals, for which they had just three months of credit card data, that just *four datapoints* was enough to identify a person within the crowd. The exact same was shown to be true for hourly location data like that from mobile phone records [21]. Researchers can thus access and perform research on these data (within the confounds of the company) but are often prohibited from sharing any of the detailed data in their publications. This has led to an ongoing discussion about journal's data policies and the effects this will have on replicability of studies using such proprietary data sources. Constructing privacy-preserving methods, which is part of the "differential privacy" literature, has received much attention because of it (see e.g. [22]). There is also the question of ethics behind combining and analysing such personal data. Especially in the way results are being utilized or capitalized upon (e.g. The Cambridge Analytica scandal [23]).

Apart from the computational skills, this data-driven paradigm-shift has extended the classic epistomological framing of economics. It does so by not solely relying on theoretical insights to steer the research, which then becomes mostly deductive, but rather working in complementary phases of induction and deduction [24]. Herein theories and hypotheses are constructed and refined by interacting with the data, often without a preconceived notion of what patterns to expect. The researcher often starts with just a general direction of interest. Then, by studying the existing patterns in the data, research questions, theories, and hypotheses can be constructed to test against the data. These results can again be used to generate or refine questions, theories, and hypotheses. This iterative process can continue till the researcher has found sufficient evidence for a theory explaining the patterns in the data. For instance, in Chapter 3, we were first interested in general economic behavior of individuals. It was only after getting to know the data that we saw the potential to investigate the subject of income mobility. Then by visualising income changes per wealth category, we found a remarkable difference in income performances between those with and without wealth. This eventually led to the research presented in Chapter 3. A prototypical structure of the workflow in this dissertation is drawn in Fig.1.1. This transformation is also not exclusive to economics but has been disseminating through all fields of social sciences. It was put forth in a well-known Google paper by Halevy *et al.* [25], "The unreasonable effectiveness of data" (the title of which is a parody on the seminal

[2]GDPR was put into law to protect the data generated by companies about their users/customers. It, among other things, prohibits companies from using or selling their customers' data without explicit consent.

paper by Wigner on "The unreasonable effectiveness of mathematics in the natural sciences" [26]) which reads: "... sciences that involve human beings rather than elementary particles have proven more resistant to elegant mathematics. ... we should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data.".

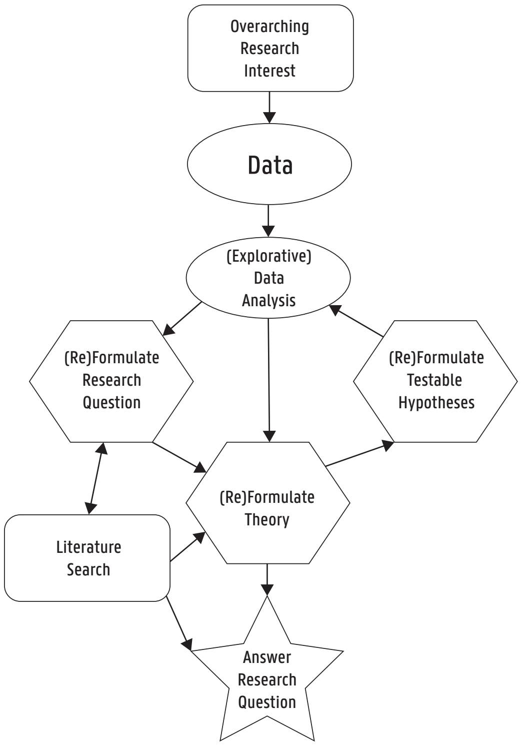



*Figure 1.1: Simplified schematic of the typical workflow of the data-driven research projects in this dissertation.*

Even though most of this section has underlined the transformative power data has and will continue to have on the field of (empirical) economics, I want to underline the importance of a solid economic framework. Extracting value from complex datasets on a purely experimental basis, by randomly looking for any type of pattern in the data, will rarely lead to insightful conclusions. As said by F. Mazzocchi: *"The data-driven approach constitutes a novel tool for scientific research. Yet this does not imply that it will supersede cognitive and methodological procedures, which have been refined during centuries of philosophical and scientific thought. There is no "end of theory" but only new opportunities. Framing the issue of Big Data in terms of oppositions, that is, deduction versus induction,* *hypothesis-driven versus data-driven or human versus machine, misses the point that both strategies are necessary and can complement each other. As others have argued, the inductive and deductive phases should be seen as an iterative cycle of knowledge acquisition."* [24] Economic theory thus provides the researcher with the organizing frameworks to better understand which mechanisms could be at play, which hypotheses to test, and which events pose natural experiments in the data. Herein we also agree with Einav *et al.* [9] who said: "Although the point is not usually emphasized, there is a sense that the richer the data, the more important it becomes to have an organizing theory to make any progress."

### 1.1.3 A hunt for data

Gathering and gaining access to sources of *Trace data* is not a trivial task. Here we describe the datasets used in this dissertation and explain, in short, the history and reasoning behind obtaining each dataset. All the datsets used in this dissertation can be classified as being of the *Trace data* type and thus carry the benefits and challenges as mentioned above.

#### Russian interbank lending

This dataset contains the full populations of contracts from the Russian domestic unsecured interbank lending market and is of the regulatory nature. It was provided by Schoors and Karas [27] and has been painstakingly assembled from public and private reporting sources. Originally it was collected with banking and finance questions in mind (e.g. importance of liquidity [28]) but was eventually also used for questions on its network features [29]. Due to the strict governmental regulatory oversight in banking, interbank literature has a long history of being very data-driven. The novelty of our data lies in the fact that it contains every issued loan together with their individual characteristics, in contrast to traditionally available monthly to yearly aggregated interbank exposures. This new level of detail enables a more precise investigation of the generative process behind loan issuance, what network structures this generates, and how this would influence risk. While leaving the latter for future research, we focus on the former in Chapter 2. This research thus situates itself in the class that leverages the extent and detail of the data to perform a detailed exploration of the cross-sectional variance.

The dataset covers the $55$ month period from January $2000$ until October $2004$. This period contains a total of $2.4$ million loans, each annotated with its lender, borrower, month of issuance, loan size, interest rate, and maturity class. The structure of the data follow the reporting standards of the CBR [3]. Banks report loan

[3]While these loans do get reported by the banks involved, this dataset is available for the other banks to possibly check. Given the large risks and penalties (up to loosing their banking licence) for misreporting, this data can be categorized as *Trace data*.

issuances to the legislator on the first day of each month. The maturity classes that are included are: overnight (less than one day, (<1d), 2-7 days (2-7d), 8-30 days (8-30d), 31-90 days (31-90d), 91-180 days (91-180d), 0.5-1 year (0.5-1y), 1-3 years (1-3y), more than 3 years (>3y).

Given the specificity of the data, the cleaning and contextualizing required a large amount of expert economic knowledge on top of mastery of the Russian language. The time frame of the data is also littered with important economic events that directly affected the Russian banking system (e.g. the "1998 Russian default", the trust crises in the second half of 2003 and the summer of 2004). Knowledge of these events and their implications was very important to correctly design the research and interpret the results.

This dataset is a great example of the fact that, as mentioned before, interaction with the data can lead to new research questions, even after years of working on it. It also shows how expert knowledge and economic framing are what make the data useful.

#### Financial data on a European bank's Belgian clients

This dataset contains several million Belgian clients from a large European bank and covers their detailed individual financial information over an eleven year period (2006-2016). The contents range from personal characteristics, financial portfolio, to every monetary transaction going in and out of the clients' accounts. As mentioned above, this *Trace data* is a by-product of the services provided by a private firm (i.e. financial services by a bank). It was obtained through a collaboration with the bank led by Ken Bastiaensen, Benjamin Vandermarliere, and Koen Schoors. Its novelty lies in the scarce availability of such data, especially because of the proportion of the Belgian population it covers (Belgian had around 11.5 million inhabitants in 2018). Its contents such as transaction specific consumption and financial wealth enables novel perspectives on enduring question such as consumption dynamics, saving dynamics, and inequality. Data for such questions were (and often still are) mostly available from surveys, which only allows for aggregated views and not transaction specific as is possible here and allows for full control over the construction of variables (e.g. income and consumption). The main limitation of the data is that we only capture one bank and thus miss part of the financial behavior of people who are active at several banks. We leverage the dataset in Chapter 3 and 4 to investigate cross-sectional variance of specific subsets of the population and study the mechanisms behind income mobility (Ch. 3) and consumption dynamics (Ch. 4).

With this dataset, we started out having a general direction of interest, namely the motivation to investigate the financial behavior of individuals. Even before the data was fully available, a preliminary evaluation was performed into the viability and potential of several lines of research. Then, after about half a year of contract negotiations, bureaucracy for access codes and infrastructure, working with their data team to complete the data, merging several legacy systems, and dealing with an environment which is not as flexible as is commonplace in academia, we finally had access to the extremely valuable dataset. Given, because of privacy concerns, we only had access to perform research on it from within the bank, on laptops owned by the bank, via which we connected to a secure server on a non-disclosed location of the bank. This is not unlike the conditions Raj Chetty and his team faced when working on American IRS data [3] and are becoming more and more common in research.

The final dataset originated from two main sources which themselves were compiled from several other sources. The first source contained the personal characteristics of the individuals such as date of birth, civil state, gender, postal code, and nationality. This source is updated at least once per calendar year, per individual, not necessarily synchronized along individuals. The other source contained every single action that involved one or more of their accounts, updated continuously with every new action. Herein each transaction contained a flag regarding the action type, including whether the transaction comes from a point of sale in Belgium, a cash withdrawal, a payment via credit card, etc. From these $248$ codes, we constructed relevant variables in order to adopt multiple definitions of individual income and consumption.

Here again economic knowledge was paramount in the construction of the final dataset. We started by performing an Exploratory Data Analysis (EDA) during which we found several interesting patterns in the data. With these patterns in mind, another literature search was performed which amounted to the formation of several novel research questions impossible to foresee before being familiar with the data. Chapters 3 and 4 are on two of these research questions.

### EVE Online: a virtual world

A virtual world is defined as: *"a computer-based online community environment that is designed and shared by individuals so that they can interact in a custom-built, simulated world."*. While these worlds originally greatly limited what players were able to do in them, they have more recently become complex and adaptive. Strict game boundaries implemented by the designers have made way for the possibility that users directly influence and drive what happens to and in the virtual world. For instance, earlier virtual worlds often had items that were sold by computer characters at a fixed price with unlimited supply, which is of course very unrealistic. Recent worlds have however become more realistic, allowing users to mine raw materials, manufacture items, and set their own selling price. EVE Online is a prime example of such a complex virtual world.

As described in [30–32], Eve Online (EVE) is an open-ended Massively Multiplayer Online Game set in a science fiction universe created by Icelandic company

CCP Games (CCP) in 2003. More than $500,000$ players compete for resources, territory, and power while engaging in a variety of professions and activities including mining, manufacturing, trading, diplomacy, leadership, exploration, and warfare, both versus the environment and against other players. As a sandbox game [33], EVE provides its players with a virtual world and the tools to explore, but players have the freedom to choose what, when, and how to approach the available content, including how they gather resources, what the goal is they want to attain, how they attain this goal, how they interact with other players, and on which activities they spent their time. Items and resources in the game are bought and sold with an in-game currency called the Inter-Stellar Kredit (ISK). Players earn ISK as a reward for engaging in activities such as playing against the environment (e.g. defeating non-player characters), selling resources gained through mining, selling goods made through industry, offering services like courier contracts or protection to other players, or by paying real-world money. Because ISK has both in-game and real-world value, and because losses in EVE are permanent, players tend to be risk averse, and this fosters realistic socioeconomic behavior.

The dataset contains detailed information on all aspects of life for players within EVE Online. This ranges from where they go, who they interact with, what kind of relationships they have with other players, to what they pay for which items, and how they manage their supply chains. Here again the data originates as a by-product from the services provided by the company. Access was granted because of a collaboration with CCP established thanks to large contributions by Kevin Hoefman, Koen Schoors, and Jan Ryckebusch.

This availability of complete information on previously impossible or hard to measure behavior and characteristics opens up possibilities for novel questions on individual behavior and group organisation/coordination. An important issue for businesses is for instance pricing. Because the characteristics of items (e.g. watches, houses) and their values to people are not explicit, it has been notoriously difficult to come up with pricing models. In a virtual world however, there is complete and transparent information, which allows to explore such pricing dynamics. This example was studied by Hoefman *et al.* [30] using EVE Online data (see Chapter 5). The main strength of this data, namely its digital nature and the ability to record everything because of it, is also the main limitation. Being a virtual world, it becomes indispensable to understand what behavior could be representative for that in the real-world and what is an artifact of the virtual nature of the world. As a result, great care was taken in the projects outlined in Chapter 5 to focus on virtual behavior which is guided by the same (economic) motivation as in real-life. To the extent possible, a comparison to real-world data was also made.

Apart from some of the largest digital multinationals (e.g. Amazon, Alphabet, Apple) most companies only know of one specific aspect of their customer's life. For instance the bank with which we collaborated only has information on the financial aspect of a person's life. One could imagine that in the future, information about any and all aspects of an individual's life will be available somewhere through interaction with digital products and services. The ethics of such a possibility, if it is going to be centralized, and who will have ownership remains a very important open question. One take on this is the vision of Tim Berners-Lee, one of the founders of the world wide web. He and his team at MIT are working on what they think should be the new internet. It's called SOLID (Social Linked Data) [34] and with it they want to hand back ownership of data to the data subjects instead of keeping it in the hands of the multinationals. Every individual would have their own "SOLID Pod" which collects the data they have generated with their activity. People could then freely choose who has access to what parts of their data for which uses. Those willing could then for instance share a multi-faceted view of their life for research, something that could potentially be a treasure trove for the social sciences. In contrast with this vision, the Chinese government is planning to centralise a plethora of data on its citizens to observe and nudge their behavior (e.g. a social-credit system).

While a real-world multi-faceted view on individuals still lies in the future, these virtual worlds, in which detailed multi-faceted data on individual's activity is available, provide a novel opportunity for research. Literature using virtual worlds as online-laboratories has been fast-growing with over $35,000$ papers[4] using them. However, less than 3% of these are in an economics-related field[5]. EVE Online is currently one of the purest examples of a complex virtual world.

## 1.2 Outline

This PhD dissertation details four research projects on different socioeconomic questions. While the subjects may appear eclectic, they are all part of a growing data-driven specialism that has extended the classic epistemological framing in economics. We split up the projects along the data they employ. The first part covers an analysis of the Russian interbank lending market. Part two covers two contributions pertaining to the societal problems of inequality and marginal propensity to consume. The third and last part details the use of a virtual world (EVE) as a novel testbed for socioeconomic theories. To give an overview, a schematic in line with Fig. 1.1 will be given at the beginning of each chapter. Chapter 6 concludes with a summary and presents an outlook on future research within this specialism and its effects on the field of economics.

[4] Obtained from running a search on the SCOPUS database on papers using the word "Virtual World" in their title, abstract, or as a keyword.

[5] SCOPUS contained a little over 450 papers containing both "Virtual World" and "Economics" anywhere in their title, abstract, or keywords

## Part I: The Russian interbank market

**Chapter 2: Loan maturity aggregation in interbank lending networks obscures mesoscale structure and economic functions** While having more details in data is often a benefit, this chapter shows that extra granularity only matters if it is informative for the research question. On the one hand, too much aggregation hides features and patterns in the data. On the other hand, introducing too many details on which the data is split (e.g. by looking only at individuals of a certain age, living in a certain place, and having a certain height) may spread the data thin and weakens the power of any statistical inference that can be done. An important question to ask before doing any statistics is thus which granularity is appropriate/most informative.

This chapter is co-authored with Marnix Van Soom, Koen Schoors, and Jan Ryckebusch and has been revised and resubmitted to Scientific Reports under the title *"Loan maturity aggregation in interbank lending networks obscures mesoscale structure and economic functions"*.

We employ the complete interbank loan contract dataset from Russia, containing more maturity granularity than previously available, to investigate whether common data granularity limitations in the interbank literature obscures important structure and economic functions. The maturity information can be utilized to separate the loans into different layers according to their maturity. We apply the layered stochastic block model of Peixoto [35] and other tools from network science to the time series of bilateral loans. We find that disregarding maturity information, or completely collapsing the maturity layers into one layer, consistently obscures mesoscale structure. The optimal maturity granularity lies between completely collapsing and completely separating the layers and depends on the development phase of the interbank market, with a more developed market requiring more maturity layers for optimal description. Closer inspection of the inferred maturity bins associated with the optimal maturity granularity reveals specific economic functions, from liquidity intermediation to financing. Collapsing a network with multiple underlying maturity layers, common in economic research, is therefore not only an incomplete representation of the interbank lending network's mesoscale structure, but also conceals existing economic functions.

## Part II: Belgian client data of a large European bank

In the next two chapters, we combine de-identified, client-level financial and demographic data on Belgian clients of a large European bank. We utilize the magnitude of the dataset to zoom in on a particularly interesting subset of the population, namely career starters. By carefully selecting a subset of the population, we are able to minimize confounding factors, increase comparability, and better focus on the research question at hand.

**Chapter 3: The wealth origins of income mobility.** In the third chapter, zooming in on career starters enabled us to explore possible mechanisms contributing to wealth inequality perpetuation. To this end, we analysed drivers of early career performance in Belgium, one of the most equal countries in the world. We selected career starters because the start of a career stands as a major life event to study how and when wealth inequality interacts with income mobility.

This chapter was co-authored with Tarik Roukny, Benjamin Vandermarliere, Jan Ryckebusch, and Koen Schoors under the title *"The wealth origins of income mobility"*.

The large economic gap between rich and poor continues to have a detrimental effect on societies. In this chapter, we propose and evaluate three possible mechanisms - social capital, innate ability, and job search conditions - behind the observed positive relationship between wealth inequality and income immobility. We find higher earning performances for individuals with higher financial wealth at the start of their careers. Neither social capital nor innate ability are able to fully explain this finding. However, evidence is found for a transmission channel via job-search conditions. Our result suggests that wealth mitigates budget constraints in job searches and leads to more optimal human capital allocations for those who can afford it, contributing to the perpetuation of economic inequality.

**Chapter 4: Liquid wealth heterogeneity, asymmetric consumption dynamics, and myopic loss aversion** In the fourth chapter, we leverage the comparability in certain financial aspects of career starters in Belgium. Specifically, that they have low (to no) debt and hold most (to all) of their wealth as liquid wealth. These similarities and the fact that we have transaction-level time-series data on income, consumption, and wealth allows us to exploit the remaining heterogeneity in consumption responses to income changes in more detail than has been possible in literature.

The idea is to study how people cope with changes in income and how this is reflected in their consumption. If an individual experiences a decline in income, does she changes her consumption by the same amount or only partially? If only partially, which financial characteristics determine this proportion? In literature, which has heavily relied on survey data, it was found that beyond personal characteristics, liquidity plays a large role in determining how individuals change their consumption in response to an income change. Multiple models have been proposed but none that cover all the different findings in literature. To this end, we empirically study the effect of liquid wealth on consumption dynamics in the absence of both illiquid wealth and debt.

We find an asymmetric consumption response to anticipated income changes, with a stronger response to income increases than to decreases. This asymmetry in consumption responses originates from the asymmetric consumption smoothing effect of liquid wealth. This indicates that people try to resist reducing consumption more than they resist increasing consumption. Rational models of consumption are unable to fully explain the results. These results are consistent, however, with the predictions of the behavioural model of myopic loss aversion which originates from a combination of present-bias or myopia (preference to spend now) and loss aversion (reluctance to lowering consumption). This suggests that there may be a link between a country's population structure and the effectiveness of its macroeconomic policies.

This chapter was co-authored with Benjamin Vandermarliere, and Koen Schoors and submitted to *the journal of economic behavior and organization* under the title *"Liquid wealth heterogeneity and the asymmetry in consumption dynamics"*.

## Part III: A multi-faceted view on behavior in a virtual world

**Chapter 5: Revealed behavior in a virtual world: the case of EVE Online** In this last chapter, I present the potential of virtual worlds as a research tool. Due to the scarcity of access to detailed *Trace data* from industry and the often limited scope of such data, virtual worlds present a unique and novel opportunity to gain a multi-faceted view on individual-level behavior. These data can be used to investigate enduring socioeconomic questions if the motivations of the individuals inside the virtual world align with the motivations in the real world. For this reason, the most interesting among the virtual worlds are those of the "sandbox" type. This type of approach to constructing virtual worlds provides players with an environment and the tools to explore it but allows them to choose what, when, and how they approach the available content. Allowing players to make their own decisions and drive the experience of the virtual world fosters realistic economic behavior. Arguable the most advanced and complex open-ended sandbox game is EVE Online.

This chapter details the world of EVE Online and explains what design choices drive the realism of in-game behavior. I then describe the contributions I personally made for our team to acquire large amounts of data from EVE Online, enabling the construction of a multi-faceted view on the behavior of (groups of) players inside EVE. I further describe three published papers that utilized the collected data. Two of these examine and expand on the theory of structural balance [31, 32] and the third expands on the theory of hedonic pricing [30].

I end with a short description of work in progress and an outlook on future possibilities.

## Computational setup

Because of the high computational requirements to perform the work presented here, several high performance computing resources were used. Chapter 2 was conducted using the computational resources (Stevin Supercomputer Infrastructure) and services provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government - department EWI. In total, over $3,000,000$ hours of single-core CPU time were used to complete the calculations in this chapter. Chapter 3 and 4 were primarily conducted using one of the European bank's secured shared servers. This server had 64 cores and 512 GB of RAM, of which often the majority was needed to process the data. For Chapter 5, the extraction was done using CCP's hadoop cluster and MSQL server. The processing of the data was done on our group's own server which has 32 cores and 256 GB of RAM.

# References

[1] Daniel S. Hamermesh. *Six Decades of Top Economics Publishing: Who and How?* Journal of Economic Literature, 51(1):162–72, March 2013.

[2] Joshua Angrist, Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu. *Economic Research Evolves: Fields and Styles*. American Economic Review, 107(5):293–97, May 2017.

[3] Linnet Taylor, Ralph Schroeder, and Eric Meyer. *Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?* Big Data & Society, 1(2):2053951714536877, 2014.

[4] Jeffrey C. Moore, Linda L. Stinson, and Edward J. Welniak. *Income Measurement Error in Surveys: A Review*. Journal of Official Statistics, 16, 01 2000.

[5] Seth Stephens-Davidowitz. *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. William Morrow & Co, 2018.

[6] Daniel Kahneman. *Thinking, fast and slow*. New York : Farrar, Straus and Giroux, 2011.

[7] Scott R. Baker. *Debt and the Response to Household Income Shocks: Validation and Application of Linked Financial Account Data*. Journal of Political Economy, 126(4):1504–1557, 2018.

[8] S. Lohr. *The age of Big Data*. New York Times, 11 Feb 2012.

[9] Liran Einav and Jonathan Levin. *Economics in the age of big data*. Science, 346(6210):1243089, 2014.

[10] Alberto Cavallo. *Scraped Data and Sticky Prices*. The Review of Economics and Statistics, 100(1):105–119, 2018.

[11] Raj Chetty, John N Friedman, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. *The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility*. Working Paper 25147, National Bureau of Economic Research, October 2018.

[12] Michael Gelman, Shachar Kariv, Matthew D. Shapiro, Dan Silverman, and Steven Tadelis. *How individuals respond to a liquidity shock: Evidence from the 2013 government shutdown*. Journal of Public Economics, 2018.

[13] Thomas Piketty and Emmanuel Saez. *Inequality in the long run*. Science, 344(6186):838–843, 2014.

[14] Raj Chetty, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang. *The fading American dream: Trends in absolute income mobility since 1940*. Science, 356(6336):398–406, 2017.

[15] Raj Chetty, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. *Race and Economic Opportunity in the United States: An Intergenerational Perspective*. Working Paper 24441, National Bureau of Economic Research, March 2018.

[16] Vivek Kumar Singh, Burcin Bozkaya, and Alex Pentland. *Money Walks: Implicit Mobility Behavior and Financial Well-Being*. PLOS ONE, 10(8):1–17, 08 2015.

[17] Marta C. Gonzlez, Csar A. Hidalgo, and Albert-Lszl Barabsi. *Understanding individual human mobility patterns*. Nature, 453:779, June 2008.

[18] Hal R. Varian. *Big Data: New Tricks for Econometrics*. Journal of Economic Perspectives, 28(2):3–28, May 2014.

[19] European Commission. *Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation) . COM(2012) 11 final. Brussels : European Commission*, 2012.

[20] Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex "Sandy" Pentland. *Unique in the shopping mall: On the reidentifiability of credit card metadata*. Science, 347(6221):536–539, 2015.

[21] Yves-Alexandre de Montjoye, Csar A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. *Unique in the Crowd: The privacy bounds of human mobility*. Scientific Reports, 3:1376, March 2013.

[22] Tatiana Komarova, Denis Nekipelov, and Evgeny Yakovlev. *Estimation of Treatment Effects from Combined Data: Identification versus Data Security*, pages 279–308. University of Chicago Press, April 2015.

[23] *The Cambridge Analytica Files*.

[24] Fulvio Mazzocchi. *Could Big Data be the end of theory in science?* EMBO reports, 16(10):1250–1255, 2015.

[25] A. Halevy, P. Norvig, and F. Pereira. *The Unreasonable Effectiveness of Data*. IEEE Intelligent Systems, 24(2):8–12, March 2009.

[26] Eugene P. Wigner. *The unreasonable effectiveness of mathematics in the natural sciences. Richard courant lecture in mathematical sciences delivered at New York University, May 11, 1959*. Communications on Pure and Applied Mathematics, 13(1):1–14, 1960.

[27] Alexei Karas and Koen J. L. Schoors. *A Guide to Russian Banks Data*. SSRN Electronic Journal, 2010.

[28] A. Karas, K. Schoors, and G. Lanine. *Liquidity matters: Evidence from the Russian interbank market*. Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium 08/520, Ghent University, Faculty of Economics and Business Administration, June 2008.

[29] Benjamin Vandermarliere, Alexei Karas, Jan Ryckebusch, and Koen Schoors. *Beyond the power law: Uncovering stylized facts in interbank networks*. Physica A: Statistical Mechanics and its Applications, 428:443 – 457, 2015.

[30] Kevin Hoefman, Aaron Bramson, Koen Schoors, and Jan Ryckebusch. *The impact of functional and social value on the price of goods*. PLOS ONE, 13(11):1–12, 11 2018.

[31] Andres M. Belaza, Kevin Hoefman, Jan Ryckebusch, Aaron Bramson, Milan van den Heuvel, and Koen Schoors. *Statistical physics of balance theory*. PLOS ONE, 12(8):1–19, 08 2017.

[32] Andres M. Belaza, Jan Ryckebusch, Aaron Bramson, Corneel Casert, Kevin Hoefman, Koen Schoors, Milan van den Heuvel, and Benjamin Vandermarliere. *Social stability and extended social balance: Quantifying the role of inactive links in social networks*. Physica A: Statistical Mechanics and its Applications, 518:270 – 284, 2019.

[33] Kurt Squire. *Open-Ended Video Games: A Model for Developing Learning for the Interactive Age*. In The Ecology of Games: Connecting Youth, Games, and Learning, pages 167–198. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning, 2008.

[34] Andrei Vlad Sambra, Essam Mansour, Sandro Hawke, Maged Zereba, Nicola Greco, Abdurrahman Ghanem, Dmitri Zagidulin, Ashraf Aboulnaga, and Tim Berners-Lee. *Solid : A Platform for Decentralized Social Applications Based on Linked Data*, http://emansour.com/research/meccano/solid_protocols.pdf (Accessed 17/01/2019), 2016.

[35] Tiago P. Peixoto. *Inferring the mesoscale structure of layered, edge-valued, and time-varying networks*. Phys. Rev. E, 92:042807, Oct 2015.

# Part I

# The Russian interbank market

# 2

# Loan maturity aggregation in interbank lending networks obscures mesoscale structure and economic functions

co-authored with: Marnix Van Soom, Jan Ryckebusch, and Koen Schoors.

## 2.1 Schematic overview

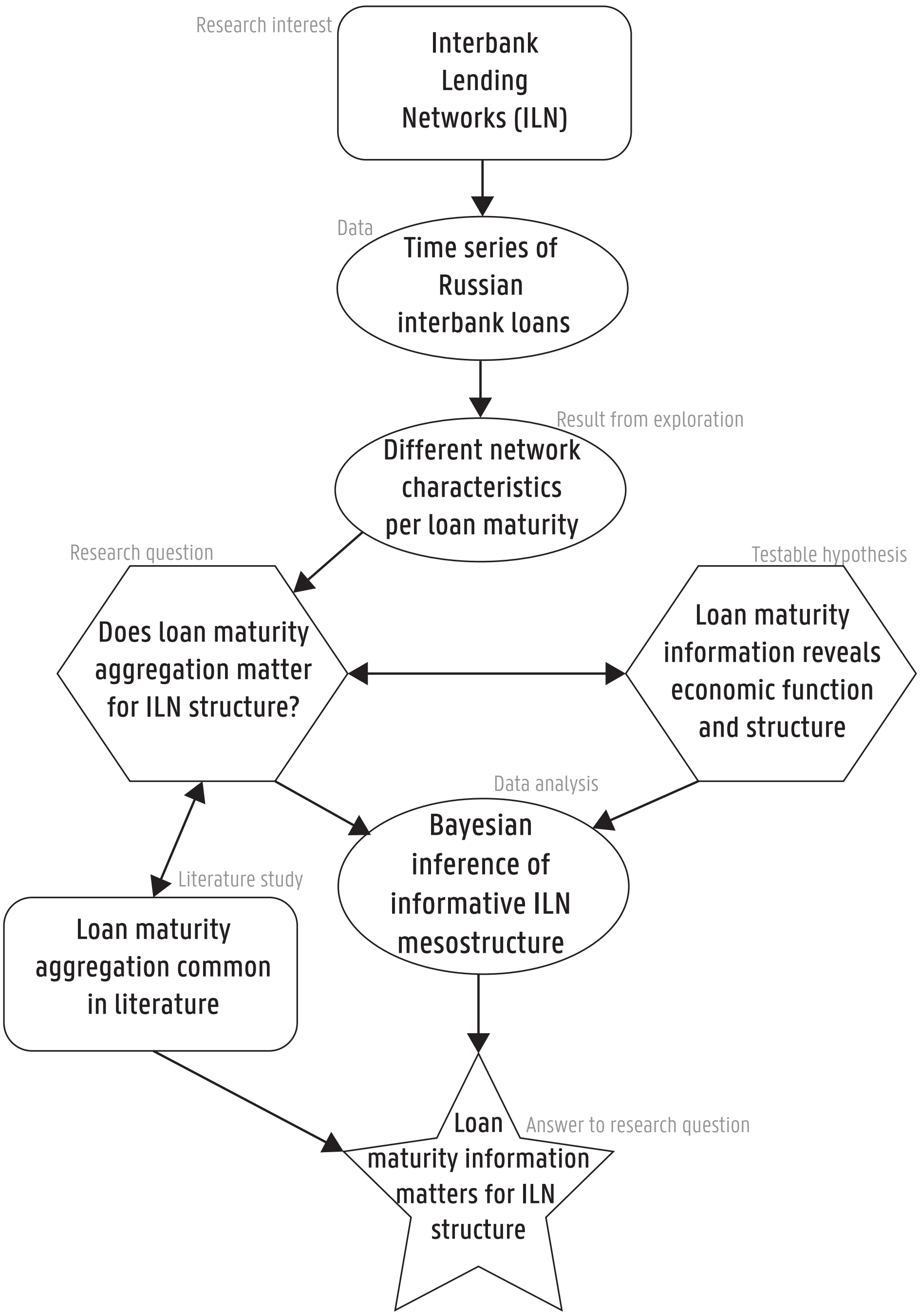



*Figure 2.1: Schematic overview of the workflow of the project on Russian Interbank Networks.*

## 2.2 Abstract

Since the 2007-2009 financial crisis, substantial academic effort has been dedicated to improving our understanding of interbank lending networks (ILNs). Because of data limitations (or by choice), the literature largely lacks multiple loan maturities. We employ a complete interbank loan contract dataset to investigate whether maturity details are informative of the network structure. Applying the layered stochastic block model of Peixoto (2015) and other tools from network science on a time series of unsecured bilateral loans with multiple maturity layers in the Russian ILN, we find that collapsing all such layers consistently obscures mesoscale structure. The optimal maturity granularity lies between completely collapsing and completely separating the maturity layers and depends on the development phase of the interbank market, with a more developed market requiring more layers for optimal description. Closer inspection of the inferred maturity bins associated with the optimal maturity granularity reveals specific economic functions, from liquidity intermediation to financing. Collapsing a network with multiple underlying maturity layers, or only focusing on one such layer, common in interbank network research, is therefore not only an incomplete representation of the ILNs mesoscale structure, but also conceals existing economic functions. This holds important insights and opportunities for theoretical and empirical studies on interbank market functioning, contagion, stability, and on the desirable level of regulatory data disclosure.

## 2.3 Introduction and overview

Interbank lending networks (ILNs) are complex network models of the interbank money markets, often called the plumbing of modern financial systems [1]. Banks make interbank loans on such markets to accommodate daily liquidity imbalances and manage their duration gap, defined as the maturity disparity between bank assets and liabilities [2]. For example, a bank holding more cash than desired may profitably lend this cash to other banks in need of cash [3].

The financial crisis of 2007-2009 has brought the interbank money markets in the public eye because their "drying up" (failure to lend cash to banks in need) was a major channel of financial contagion [4]. Since then, substantial academic effort has been dedicated towards improving our understanding of these markets. It turns out that representing the interbank money market as a network is a simple and powerful abstraction [5], that is arguably more realistic than modelling it as one representative bank, as is customary in traditional macro-finance [6]. Network analysis is therefore now one of the standard tools of financial stability experts worldwide, i.a. at the IMF [7] and ECB [8].

A thorough understanding of the interbank money market is vital to prevent systemic meltdowns [9, 10], implying a need for ever more realistic models of ILNs. We contribute by studying an often unavailable and sometimes disregarded aspect of ILNs, i.e. the different loan maturities. The maturity of a loan, i.e. the period after which the loan must be repaid, is an important instrument for banks to organise their lending and borrowing activity in function of risk minimisation [2, 11, 12]. The vast majority of empirical financial network papers study the overnight interbank market [6]. This seems to partly originate from the widely held view that overnight lending makes up the majority of interbank exposure, a view that stands in contrast with recent results showing the average contract length in the German interbank market to be well over one year [2]. Sometimes however, empirical ILN literature is unable to include different loan maturities because of data limitations (see [13] for a recent overview of available interbank data). As a consequence agent-based models of ILNs that often either neglect loan maturities or limit it to a modelling detail [14, 15], even though maturity choices reflect bank risk strategies. For example, stress testing models would benefit from exposure data enriched with maturity information because different risk strategies can lead to different network structures, which in turn have different contagion processes. Excluding maturity layers can then lead to an underestimation of the likelihood of contagion [16]. Enabled by a particularly granular dataset, a panel of all lending contracts issued in the unsecured Russian ILN [17], we investigate qualitatively what kind of information is lost by not differentiating between the loan maturities. We further try to estimate the pitfalls of this common practice. The approach we take consists of explicitly modelling the Russian ILN by layered stochastic block models (SBMs). This allows us to determine to what extent the loan maturities are informative of the Russian ILN's *mesoscale*, i.e. the higher-level organisation of the banks into bank groups.

The dataset used in this work consists of 57 monthly ILNs constructed from the complete panel of contracts on the Russian domestic unsecured interbank lending market, one for each month in the period from January 2000 to October 2004, except for January 2003. There are a total of 2.4 million loans, each annotated with its lender, borrower, month of issuance, loan size and *maturity class*. The dataset is unusual in the sense that it is an loan issuance network rather than an exposure network, the latter being the typical ILN representation found in the literature [13]. The exposure network associated with a given issuance network can be derived by aggregating the loan sizes in the issuance network over time. The data originates directly from the bank reports to the Central Bank of Russia (CBR) and came pre-assigned into eight contiguous maturity classes, so that every loan can be assigned to a particular maturity class (e.g. 2-7 days). Fig. 2.2 shows the monthly number of active banks, loans issued, and outstanding loans per maturity class, together with a list of the eight maturity classes (further details on the dataset can be found in the Methods section). We look at the monthly ILNs as layered networks such that

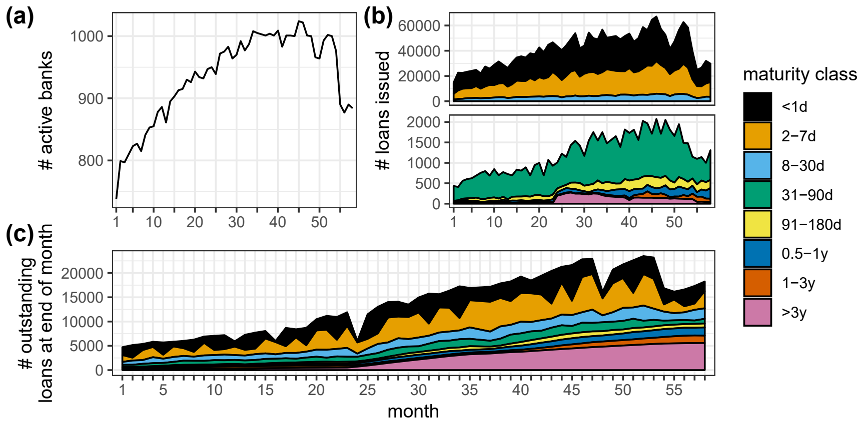



*Figure 2.2: Temporal evolution of the lending activity in the Russian interbank lending network. Month 1 corresponds with January 2000.* ***(a)*** *The number of active banks on the lending market per month. A bank is active whenever it is the originator or the beneficiary of at least one new interbank loan in the given month.* ***(b)*** *The number of loans issued per month and per maturity. Note the different scales on the vertical axis of the two panels: The majority of issued loans have "short" ($\leq$ 30 days) maturities.* ***(c)*** *The number of outstanding loans, i.e. loans open on the last working day of each month. From this perspective the loans with longer maturities now play an equally important role as the short ones. Note that we have included loans issued before January 2000 for this panel (see Methods).*

each maturity class corresponds to one *maturity layer*. By putting contiguous maturity layers into bins, they can be coarse-grained to achieve various levels of *layer granularity*, ranging from complete differentiation (eight bins, one layer per bin) to complete aggregation (all layers merged into one bin, the collapsed network).

First, we provide descriptive statistics of the monthly ILNs in a comparative framework. We characterise the network topology of each maturity layer separately using the typical network measures from ILN literature (listed in the left-most column of Table 2.1) and compare the results across the maturity layers and across the literature. Broadly speaking and integrating over time, we find layer non-homogeneity for most topology measures. This means that, while stylised facts found in the literature were also found in some of the maturity layers, the topology measures do not take on similar values for all maturity layers simultaneously. Layer non-homogeneity points to the fact that complete aggregation or single layer focus involves the loss of the topological diversity present in the maturity layers, although some layers do share similar lending patterns in specific months (by lending pattern we mean a topological pattern in an ILN). For example the three most dense layers – the "short" layers, i.e. all maturities below 30 days – display a strong similarity throughout the full time period. The short layers' topological similarity, together with their dominant share in the number of loans issued, could suggest that complete aggregation might not be harmful after all, irrespective of what the found layer non-homogeneity suggests. The question then becomes whether the information lost by complete aggregation is relevant for ILN structure and economic function. Because financial stability plays a central role in literature and policy, we are interested in the money markets' mesoscale organisation, which can effectuate the propagation of instability and risk [18], rather than individual banks' lending strategies. Relevant information is thus any set of information that allows for characterisation of the mesoscale structure rather than other, more specific, more "noisy", information about local lending patterns (e.g. clustering).

In order to characterize the effect of aggregating maturity layers on the ILN mesoscale, we explicitly model the monthly ILNs by layered SBMs. SBMs can infer statistically significant group structure in networks, without making informative prior assumptions about the type of mesostructure itself. SBMs model the mesoscale of a network by assuming that the nodes in a network behave "group-like" rather than on individual account. *Layered* SBMs generalize SBMs for layered networks by allowing the group structure to have a different topological pattern on each layer. As a concrete example, imagine a network with two layers representing the mating (layer 1) and conflict (layer 2) interactions in a population of deer (the nodes). One possible layered SBM of this network is the division of the deer into two groups, male and female, so that observed occurrences of mating and fighting between two deer are explained only by their sexes. Note that such a simple model – which does not take things like social status etc. into account to explain the observed interactions – could well suffice to infer the deer's sexes if these were unknown.

Layered SBMs formulated in a Bayesian setting can be extended to the *coarse-grained* layered SBM in order to infer the appropriate level of layer granularity (the optimal granularity, OG) along with the group structure. The bins in the OG correspond to lending patterns between the bank groups that differ from each other in a statistically significant way. We explain this in detail in the next section and give a simple illustration in Fig.2.3, but the essence of how an OG can be inferred can already be understood by a simple regularisation argument. The complexity and modelling power of a coarse-grained layered SBM is determined primarily by the number of groups and the degree of layer granularity, as these parameters simultaneously define the "resolution" available to model the observed layered network. To prevent overfitting (i.e. modelling noise), any increase in model complexity should be warranted by enough statistical evidence in the data. Thus the OG may be determined in general by any regularisation principle to balance model complexity and quality of fit; we use Bayesian model selection for this. One specific advantage of this approach is that there is formally no difference between inferring the number of groups – and in fact the groups themselves – and the OG; both are determined as part of a single inference for each monthly ILN.

The inferred monthly OGs are displayed in Fig. 2.4(a). Each OG consists of a set of bins numbered from short to long maturity by the OG bin index (OGB index). We mention here two findings easily deduced from Fig. 2.4(a). First, we find that the OG always lies between complete differentiation and complete aggregation. This means that on the one hand the eight available maturity classes are partly redundant and that the lending patterns may be described more effectively by merging maturity layers into bins, as is indeed the case for the short layers mentioned before. On the other hand complete aggregation apparently discards important information needed to model the monthly ILN's mesostructure: The lending patterns between the bank groups depend significantly on the maturity classes of the loans. The lending patterns between the bank groups depend significantly on the maturity classes of the loans. This also illustrates the more general added benefit of statistically inferring the optimal maturity bins, instead of setting them manually based on some heuristic or prior: if one would set bin widths too narrowly, the data might be too thinly spread to detect any structure at all. Conversely, if one would set bin widths too widely, the different structures, and their economic functions, might be obfuscated. Second, the monthly *number of bins* in the OG correlate roughly with the known two phases in the Russian money market's development: early development (roughly before month 35) and emerging maturity (from month 35 onward) (see Methods). This points to a natural ordering in number and complexity of lending patterns emerging at different phases of market development.

To interpret the lending patterns, we take a closer look at the monthly OG bins at the network and individual bank level. (Note that this goes beyond the previously established network measures of the individual maturity layers, as these are now merged according to whether they form a statistically significant lending pattern.) The notable result here is that the bins may be characterised by a simple aspect of the most important banks' individual lending behaviour: At monthly time scales the important banks in "short" bins tend to both lend and borrow equal amounts of cash (indicative of financial intermediation), while the important banks in the "long" bins tend to either lend or borrow (indicative of financing). For the development phases, this suggest that patterns of financial intermediation are present at all phases of development, while patterns more in line with financing only appear at later phases.

Our coarse-grained layered SBM of the ILN thus uncovers a correlation between statistically significant lending patterns between groups of banks, the economic functions of important banks, and the maturity classes of the loans involved, showing that maturity information matters for the understanding of ILNs.

## 2.4 Results

The adopted notation is in line with the one of [19, 20]. An ILN at month $t$ is denoted as $\{G_l\}$, with $G_l$ the network for one of the eight maturity layers: $G_{<\text{1d}}$, $G_{\text{2-7d}}$, $G_{\text{8-30d}}$, $G_{\text{31-90d}}$, $G_{\text{91-180d}}$, $G_{\text{0.5-1y}}$, $G_{\text{1-3y}}$, $G_{>\text{3y}}$. As we analyze each month independently we do not attach a time label to $\{G_l\}$. The $G_l$ are directed weighted multigraphs. In addition to the asymmetrical adjacency matrix $A^l_{ij} \in \mathbb{N}_0$ of layer $l$, the $G_l$ possesses the unordered edge covariates $x^l_{ijk}$ ($k \in [1, A^l_{ij}]$ for $A^l_{ij} > 0$). Thereby, each $k$th parallel edge between banks $i$ and $j$ represents a loan lent from bank $i$ to bank $j$ with a maturity in maturity class $l$ and of size $x^l_{ijk}$. The *collapsed network* $G_c$ corresponds to complete maturity aggregation. Its adjacency matrix $A_{ij} = \sum_l A^l_{ij}$ and the edge covariates $x_{ijk}$ are constructed by flattening (i.e. collapsing) the $x^l_{ijk}$ along the $k, l$ axes. We denote a level of granularity by the maturity bin set $\{\ell\}$ where $\ell$ specifies a set of merged layers. For example, the OG of month 58 (see the last month in Fig. 2.4(a)) corresponds to $\{\ell\} = \{\{<\text{1d}, \text{2-7d}\}, \{\text{8-30d}\}, \{\text{31-90d}\}, \{\text{91-180d}, \text{0.5-1y}, \text{1-3y}, >\text{3y}\}\}$. The ILN $\{G_\ell\}$ representing the level of granularity specified by $\{\ell\}$ is constructed from $\{G_l\}$ by merging maturity layers according to $\{\ell\}$.

### Descriptive statistics of the Russian ILN in a comparative framework

We start by characterising the maturity layers and the collapsed form of the Russian ILN in terms of monthly (and occasionally yearly) time series of several ILN measures typically used in the literature. Layer analysis of layered ILNs has been performed for the interbank money markets of several countries, e.g. Mexico [21] and the UK [22]. The layers in those works, however, do (or can) typically not differentiate between maturity classes [6, 13]. The maturity classes per layer of those that did are shown in Appendix G. The work by Bargigli et al. [1], which is related to this work, separates overnight loans, loans up to 1y and loans >1y in a three layer end-of-year exposure network representation of the Italian ILN. They find that the structures of the three maturity layers are not representative of each other. The measures that we analyse are: density, degree distribution, clustering coefficients, average shortest path length, degree mixing (as a proxy for bank size mixing), loan activity, and loan size. The results are compared to the stylised facts found in literature. A summary of this analysis is given in Table 2.1, while details of the analysis can be found in the Supplementary Material, Appendix 2.A.

The stylised facts describing prototypical ILNs are often deduced from exposure networks with only one maturity class. Looking at the maturity layers separately, we also find layer non-homogeneity for all measures in Table 2.1 except for the distribution of degree and transaction volumes. This means that variations of the ILN measures across the maturity layers are observed. While the stylised

| ILN measure | Value | Layer homogeneity | Selected studies |
|---|---|---|---|
| density | **sparse** | No | [1, 6, 23, 25–27] |
| degree distribution (in and out degrees) | **heavy-tailed** | Yes | [1, 6, 23–25, 28] |
| topological structure | scale-free / core-periphery | | [1, 29, 30]/ [6, 25–28] |
| clustering coefficients | low / **high** | No | [23]/ [1, 6, 25, 30] |
| average shortest path length | **small** / "small world" | No | [25, 27]/ [6, 23, 30] |
| bank size mixing | **disassortative** | No | [6, 21, 23, 25, 27, 29] |
| distribution of transaction volumes | **heavy-tailed** | Yes | [23, 24] |

*Table 2.1: Stylized network properties of interbank lending networks (ILNs) according to selected studies. For some ILN measures conflicting values are reported, in which case we separate them by a backslash. For example, low clustering coefficients are reported by [23]. The values for the ILN measures in bold apply to the collapsed Russian ILN (see [24] and Appendix 2.A). With the layer homogeneity we indicate whether the quoted ILN measures apply to all maturity layers of the Russian ILN.*

features are valid for the short maturity layers ($G_{<1\mathrm{d}}$, $G_{2\text{-}7\mathrm{d}}$, $G_{8\text{-}30\mathrm{d}}$) and for the collapsed network ($G_c$), they become progressively invalid with growing loan maturities. As a matter of fact, we find that the stylised facts of the collapsed Russian ILN do not hold over all maturity layers. Given the similarities in the stylised facts across countries, we anticipate a similar behaviour for the ILNs of other countries. The short layers $G_{<1\mathrm{d}}$, $G_{2\text{-}7\mathrm{d}}$, $G_{8\text{-}30\mathrm{d}}$ contain 97% of all issued loans. Upon merging those we more or less retrieve the collapsed Russian ILN with all the loan issuances. The longer maturity loans represent a few percent of the issuance but they are of sizeable economic relevance due to their specific turnover and their weight in the outstanding loans (see Fig. 2.2(c)). During times of turmoil the short-term loans are often not renewed, but the long-term ones remain on the books till maturity, making them important for the stability of the ILN. This is reflected in the interest rate spreads shown in Appendix 2.H.

Core-periphery (CP) structure has been observed in many real-world networks [31] and in ILNs [5, 6]. The seminal work of Craig and von Peter [28] introduced an economic foundation for its occurrence in ILNs, i.e. the elementary function of economic intermediation performed by banks. CP structure breaks with traditional theoretical banking literature where the interbank money market is modelled as a centralised exchange in which banks smooth out liquidity shocks. In contrast to the centralised exchange model, an ILN with CP structure gives rise to a sparse network. Thereby, a group of densely connected "core" banks perform the economic function of financial intermediation between numerous smaller, sparsely connected "periphery" banks. Formally, an ILN has CP structure if the lending patterns can be fully explained by grouping the banks into either "core banks" or "periphery banks". In the ideal situation, the bilateral relations between the banks define the bank group memberships (i.e. whether a bank is core or periphery) by the following set of rules: (i) core banks lend to each other; (ii) periphery banks do not lend to each other; (iii) core banks lend to periphery banks; (iv) core banks borrow from periphery banks. For real-world ILNs with imperfect CP structure, several algorithms have been proposed (e.g. [28, 32]) to detect CP structure and to determine the group memberships. We believe however that the proper way to establish CP structure in networks is by Bayesian inference of SBMs as these are ideally suited to parametrise the CP two-group structure and the four rules mentioned above, rather than minimising an objective function that might detect spurious CP structure, as shown by some recent literature [1, 33, 34]. Inferring CP structure by Bayesian inference of SBMs has been proposed in [35] and applied to the Italian e-MID ILN in [36, 37] where depending on the time scale and SBM model extension, a bipartite or CP structure was found. We find some indications supporting a CP structure in the Russian ILN for $G_c$ and $G_{<1\text{d}}$, $G_{2\text{-}7\text{d}}$, $G_{8\text{-}30\text{d}}$. As explained in Appendix 2.A these indications stem from the heavy-tailed degree distributions, disassortative degree mixing and the small average shortest path length.

## Modelling the Russian ILN with the coarse-grained layered SBM

The idea of banks behaving in groups with respect to lending and borrowing because of trading relationships in the interbank money market has been posited in various forms in the literature [6, 12, 26, 28, 38–42], though not often explicitly in the form of SBMs [5, 36, 37]. Such group structure, also called network mesoscale structure, is abundant in real-life complex networks [43], notably social networks [44]. In the Russian interbank money market, there are several reasons to anticipate that the group structure is important: relationships are a way to solve problems with asymmetric information, a pervasive problem in the Russian banking sector and economy at large [45]; fragmentation of the Russian financial market due to the country's size (i.e. eight time-zones); the presence of institutions controlled by the state to various degrees [46]. The most important advantages of using SBMs to detect the group or mesoscale structure are [47]: (i) Theoretical guarantees against overfitting; (ii) They can be extended easily when formulated in a Bayesian setting; (iii) The ability to describe a wide variety of lending patterns (e.g. ILNs modelled by community structure, bipartite structure, CP structure or Erdős-Rényi graphs). As mentioned before, the SBM flavor we use to model the monthly ILNs is the coarse-grained layered SBM, which extends the layered SBM, itself an extension of the standard SBM. With "coarse-grained layered SBM" we refer to an extension to the layered SBM developed in [19] which allows one to infer the OG along with the bank groups and other SBM parameters. We shortly introduce the SBM, the layered SBM and the rationale behind its coarse-grained extension in a qualitative setting before discussing the coarse-grained layered SBM for the Russian ILN. A comprehensive discussion about SBMs in a Bayesian setting can be found in [47] and the coarse-grained layered SBM used in this work is presented in [19, 20].

**Introduction to the SBM, the layered SBM and the coarse-grained layered SBM**

SBMs are canonical models to study clustering and perform community detection [48, 49]. SBMs model topological patterns by assuming that the nodes "behave group-like" rather than on individual account, i.e. for a network $G$ with $N$ nodes modelled by $B \leq N$ groups, one assumes that the amount of connections between any two nodes $1 \leq i, j \leq N$ depends only on their group memberships $1 \leq b_i, b_j \leq B$, where $b_i$ is the group assignment of the $i$th node. When formulated in a Bayesian setting, the basic goal of SBMs is to determine the posterior probability distribution of all possible group assignments $\{b_i\}$ (where $B = \max_i b_i$) given the observed network $G$, a quantity written as $p(\{b_i\}|G)$. Because this is intractable for networks with more than a few nodes and edges, one is typically content with the maximum a posteriori probability (MAP) estimate, i.e. $\operatorname{argmax}_{\{b_i\}} p(\{b_i\}|G)$, to which one refers to as "the fit" to the observed network $G$. Maximising the posterior $p(\{b_i\}|G)$ in search for the MAP estimate equivalently minimises the information-theoretic *description length* (DL) of the data $G$, i.e. $\Sigma = -\log p(G, \{b_i\}) = \mathcal{S} + \mathcal{L}$ with $\mathcal{S} = -\log p(G|\{b_i\})$ and $\mathcal{L} = -\log p(\{b_i\})$. Choosing the base of the log to be 2, $\mathcal{S}$ is the number of bits needed to describe the data ($G$) given the model parameters $\{b_i\}$ and $\mathcal{L}$ is the number of bits necessary to describe the model parameters. In other words, the best fit to the data is the one that *compresses* it most, i.e. yields the shortest DL. This is the minimum description length principle (MDL).

Though MDL as a regularisation device is fully equivalent to Bayesian model selection [50, 51], we invoke the MDL principle in our qualitative discussion as it provides an arguably more intuitive explanation to how SBMs formulated in a Bayesian setting achieve robustness against overfitting when the model and the prior probabilities accurately represent our (lack of) knowledge [47, 52]. In the case of SBMs the primary parameter that controls the model's complexity is the number of groups $B$. Increasing $B$ improves the maximum likelihood fit $p(G|\{b_i\})$ monotonically, as new groups become available to account for any possibly insignificant deviation from the group's behaviour. More complicated models (larger $B$) are only preferred if there is sufficient evidence available in the data to compensate the extra degrees of freedom. This is achieved in the Bayesian formalism by specifying a prior $p(\{b_i\})$ and subsequent model selection based on integrated likelihoods and statistical significance. From the MDL view this robustness against overfitting is achieved in the following manner: If $B$ becomes large, it decreases $\mathcal{S}$ but increases $\mathcal{L}$. The latter functions as a "penalty" that disfavours overly complex models [53]. The optimal choice of $B$ minimises the DL $\Sigma$, which induces a proper balance between $\mathcal{S}$ and $\mathcal{L}$. In other words, the optimal choice of $B$ and $\{b_i\}$ corresponds with the model that compresses the data most.

Layered SBMs are extensions of SBMs that additionally allow the group behaviour to depend on the network layers. In this work we use a specific layered SBM known as the independent layers SBM for all monthly ILNs. The independent layers SBM assumes that a layered network $\{G_l\}$ can be modelled as one group structure which exhibits a topological pattern in each layer. In other words, each layer $G_l$ is modelled by an independent SBM constrained by the fact that the group memberships of the nodes $\{b_i\}$ are the same across all layers. Thus the model complexity of a layered SBM is now additionally controlled by the number of layers $L$ present in the observed layered network $\{G_l\}$ ($1 \leq l \leq L$), next to the number of groups $B$. The increased model power relative to the standard SBM again raises the question of overfitting: given the group structure $\{b_i\}$ of the layered network $\{G_l\}$, is it necessary to posit $L$ different topological patterns for each layer $G_l$, or can some layers be explained equally well by just one topological pattern and hence be merged? The layered SBM itself cannot provide an answer to this question, as $L$ is determined by $\{G_l\}$ and is thus simply a fixed component of the model complexity.

The coarse-grained layered SBM extends the layered SBM by assuming that the observed "high resolution" layered network $\{G_l\}$ may be explained by an underlying "lower resolution" layered network $\{G_\ell\}$ consisting of the merged $G_l$ according to $\{\ell\}$, a set of layer bins which specifies the level of granularity of $\{G_\ell\}$. The optimal level of granularity (OG) is inferred simultaneously with the number of groups $B$ and the group structure $\{b_i\}$ by searching for the layered SBM that most compresses the *lower resolution* layered network $\{G_\ell\}$, while taking into account the inevitable information loss incurred due to the lower degree of granularity (i.e. decrease in quality of fit). When the OG for a given layered network $\{G_l\}$ is complete aggregation, this indicates that the layer divisions in $\{G_l\}$ do not correlate with the mesoscale of its associated collapsed form $\{G_c\}$. By contrast, an OG that is different from complete aggregation points to a mesoscale structure that is too complicated to be understood at the collapsed network level. The OG merges layers such that the layer bins in $\{\ell\}$ "acquire meaning" so that at the bin interfaces the topological patterns between the groups change in a statistically significant way.

**The coarse-grained layered SBM for the Russian ILN**

We approach the Russian ILN as a time series of monthly ILNs and model each month separately with the coarse-grained layered SBM introduced in the previous section. Note that this choice does not make any assumptions on nor excludes the existence of temporal correlations throughout the months. The specific flavor of the underlying layered SBM that we use is the microcanonical independent layers weighted DCSBM, which takes the following features (above the expected group structure captured by the standard SBM) into account:

- Edge directedness and the possibility of parallel edges, i.e. multiple loans can be made between two banks in a given month.
- Heavy-tailed degree distributions (see Appendix 2.A). This is captured by the degree-corrected SBM (DCSBM) [54].
- Heavy-tailed loan size distributions [24]. This is captured by extending the DCSBM to the weighted DCSBM [20] where the sizes of the loans between bank groups are modelled by log-normal distributions. In this way each ordered pair of bank groups has a lending pattern modelled as consisting of loans whose size's magnitude has a characteristic scale [52].
- Maturity classes, modelled as network layers. This is captured by the independent layer SBM. The lending pattern in each layer is modelled by an independent weighted DCSBM with the constraint that the bank groups are identical in each layer.

Because of computational limitations, we do not explicitly take into account variables such as balance sheets, bank ownership, and interest rates. We deem, however, that this does not severely impact the ability of our model to capture the structural variability in the network. Indeed, structural features originating from the omitted variables can be effectively captured by the model. First, while balance sheets are not explicitly taken into account, we include degree distribution, which correlates to bank size, and loan size distribution (see Appendix 2.A). Second, if bank ownership drives significantly different lending- and borrowing behaviour, these patterns can still be captured by the SBMs without explicitly including the information. It would be interesting to study if and how bank ownership correlates to the function banks fulfil in the different maturity layers. This falls outside the scope of this work. A study of the term structure of interest rates is included in Appendix 2.H. In line with the expectation theory of interest rates [55], the yield-curve is upward sloping for longer maturities. Longer loan maturities lead to higher interest rates through averaging expected future short-term rates and adding a premium for liquidity- and default-risk. The introduction of maturity layers effectively captures the fact that differing risk patterns manifest themselves through differing loan structures in the layers. Recent work on the generative processes of risk in ILNs has shown that information such as interest rate and Credit Default Swap (CDS) spreads [56, 57] play a central role in contagion dynamics. We note that securitised products, CDO, and CDS play a very marginal role in the Russian ILN in our period under study. In this way the Russian ILN provides a natural lab where the control parameter of risk derivatives can be effectively set to zero. In the forthcoming discussion on possible future research directions those opportunities will be highlighted. Local structures such as dense subgraphs fall beyond current SBMs' potential to capture mesoscale structure [47]. In the Methods section we motivate the choice to treat the maturity layers of the monthly ILNs as independent.

The coarse-grained layered SBM augments the parameter set of the layered SBM $\{\theta\}$ by the maturity bin set specifying the level of granularity $\{\ell\}$. Thus its parameters are denoted as $(\{\theta\}, \{\ell\})$. The coarse-grained layered SBM of a monthly ILN $\{G_l\}$ is a generative model given by [19]

$$p(\{G_l\}, \{\theta\}, \{\ell\}) = p(\{G_l\}|\{\theta\}, \{\ell\}) \times p(\{\theta\}) \times p(\{\ell\}). \tag{2.1}$$

Expressions for the model likelihood $p(\{G_l\}|\{\theta\}, \{\ell\}) \propto p(\{G_\ell\}|\{\theta\})$ and prior probabilities $p(\{\theta\})$ and $p(\{\ell\})$ can be found in [19, 20] where $p(\{G_\ell\}|\{\theta\})$ and $p(\{\theta\})$ are defined by the layered SBM and $p(\{G_l\}|\{\theta\}, \{\ell\})$ and $p(\{\ell\})$ are defined by the coarse-grained extension. As the underlying maturity classes are inherently ordered, the uninformative prior probability for the maturity bin set $p(\{\ell\})$ is determined by the constraint that only contiguous layers may be binned. In Appendix 2.B we infer the OGs under the more general non-contiguous binning assumption (with a different prior $p(\{\ell\})$) and find qualitatively similar results as in Fig. 2.4.

The posterior probability of the parameters $\{\theta\}, \{\ell\}$ is proportional to Eq. 2.1:

$$p(\{\theta\}, \{\ell\}|\{G_l\}) = \frac{p(\{G_l\}, \{\theta\}, \{\ell\})}{p(\{G_l\})}, \tag{2.2}$$

where $p(\{G_l\})$ is independent of $\{\theta\}$ and $\{\ell\}$. We may infer the bank groups $\{b_i\} \in \{\theta\}$ and the OG as the maximum a posteriori probability (MAP) estimate by searching for the mode of Eq. 2.2 (or equivalently Eq. 2.1) with the inference algorithm explained in the Methods. In addition, we can compare the *posterior odds ratio* (POR) $\Lambda$ between two coarse-grained layered SBMs $\mathcal{M}_a, \mathcal{M}_b$ representing two different levels of granularity $\{\ell\}_a, \{\ell\}_b$ by evaluating the ratio:

$$\Lambda = \frac{p(\{\theta\}_a, \{\ell\}_a, \mathcal{M}_a|\{G_l\})}{p(\{\theta\}_b, \{\ell\}_b, \mathcal{M}_b|\{G_l\})} = \frac{p(\{\theta\}_a, \{\ell\}_a|\{G_l\})}{p(\{\theta\}_b, \{\ell\}_b|\{G_l\})} = \frac{p(\{G_l\}, \{\theta\}_a, \{\ell\}_a)}{p(\{G_l\}, \{\theta\}_b, \{\ell\}_b)}, \tag{2.3}$$

where the constant $p(\{G_l\})$ and the prior beliefs $p(\mathcal{M}_a), p(\mathcal{M}_b)$ for the coarse-grained layerered SBMs $\mathcal{M}_a$ and $\mathcal{M}_b$ have cancelled out, as we had no prior preference with regard to the degree of granularity (i.e. $p(\mathcal{M}_a) = p(\mathcal{M}_b)$). Values of $\Lambda > 1$ indicate that according to the data, $\{\ell\}_a$ is preferred over $\{\ell\}_b$ with a degree of statistical significance given by the magnitude of $\Lambda$ [20]. The model selection implicit in the POR can be illustrated as follows. Given that $\Sigma_a$ ($\Sigma_b$) denotes the DL of $\{G_l\}$ according to model $\mathcal{M}_a$ ($\mathcal{M}_b$) the above equation implies that $\log \Lambda = \Sigma_b - \Sigma_a$ and one recovers the MDL principle. Indeed, the preferred model is the one that achieves the most optimal compression of the data. Accordingly, the POR is a model selection criterion that operates similarly to alternate information-based "goodness-of-fit" criteria such as BIC [58] and AIC [59]. The POR criterion, however, is "exact" for the coarse-grained layered SBMs at hand while BIC and AIC rely on specific assumptions about the asymptotic shape of the model likelihood which are known to be invalid for the SBM [47, 60].

We use $\log_{10} \Lambda$ to determine confidence levels for rejecting complete differentiation and complete aggregation as OG for the monthly ILNs. This is achieved by setting in Eq. 2.3 $\{\ell\}_b$ to the OG inferred from the algorithm and $\{\ell\}_a$ to either complete differentiation or complete aggregation. Note that these confidence levels of rejection give rise to values $\log_{10} \Lambda \leq 0$. Large negative values of $\log_{10} \Lambda$ point to strong evidence for rejecting complete differentiation and/or complete aggregation.

**Example for a small layered network**

As an illustration, we determine the OG for a small network of 50 nodes with three generic interactions (A, B and C) drawn in bundles (Fig. 2.3(a)). Instead of specifying the individual interactions, one can describe the network in a more parsimonious way by specifying the "wiring patterns" between groups of nodes. The nodes are grouped in circles, squares and triangles, so as to encode a specific high-level description of the network. For example, interaction type A does not occur between two squares and between two triangles. Interaction A gives rise to circle-triangle and circle-square interactions. There are no circle-circle interactions of the C type, and circle-circle interactions of the types A and B are sparse. This is the kind of higher-level organisation into groups of nodes that SBMs can infer.

The results of the fits with the coarse-grained layered SBM for three levels of granularity are displayed in Fig. 2.3(b). Comparing the DL $\Sigma$ for the three fits, the preferred model is the OG $\{\{\mathrm{A}, \mathrm{B}\}, \mathrm{C}\}$ with $\Sigma = 1687$ bits $\approx 211$ bytes. The network in Fig. 2.3(a) can be saved to disk in graph-tool's [61] native binary format as a file with a size of approximately 3,400 bytes, while the OG coarse-grained layered SBM can compress this (e.g. using arithmetic coding) down to $\Sigma \approx 211$ bytes (excluding the bytes needed for storage of practicalities [62] such as the file header). We use the POR $\Lambda$ of Eq. 2.3 to determine the confidence levels. The model $\{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ that stands for complete differentiation is rejected with $\log_{10} \Lambda \approx -80$, indicating that it is an overly complicated model of the

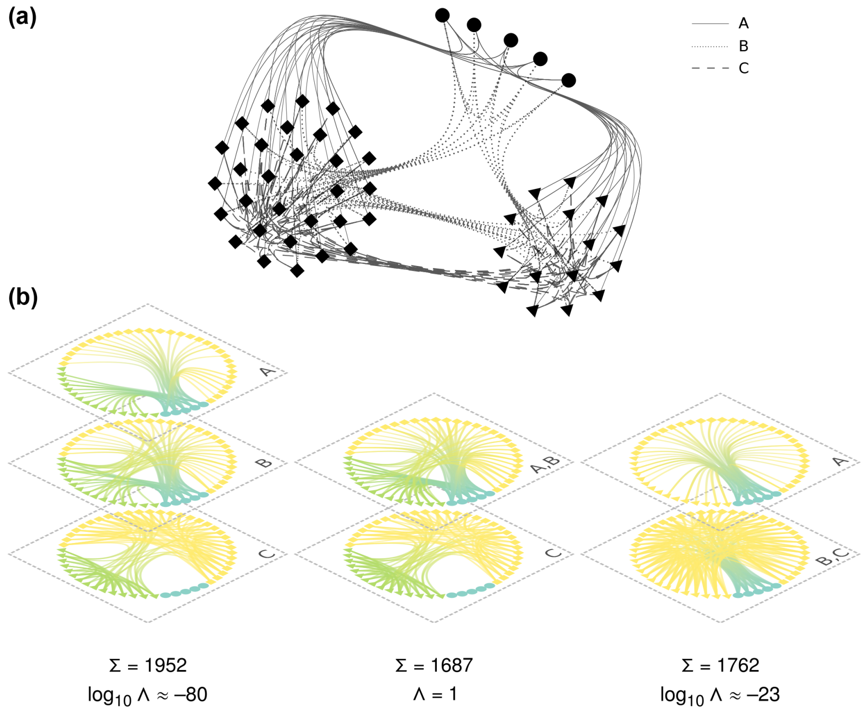



*Figure 2.3: Illustration of inferring the optimal granularity (OG) for a small network with the coarse-grained layered SBM. Images created with the graph-tool Python library [61].* ***(a)*** *An undirected and unweighted network with three types of edges (representing three generic interactions of type A, B, C) and three types of nodes (circle, square, and triangle). The interaction types define the network layers. Their network structure can be described as: perfect core-periphery (CP) (layer A), imperfect CP (B), and community structure (C).* ***(b)*** *The network is shown with three levels of granularity. From left to right, these are: complete differentation $\{A, B, C\}$; merging of layers A and B with a lonestanding C $\{\{A, B\}, C\}$; and the merging of layers B and C with a lonestanding A $\{A, \{B, C\}\}$. The nodes are coloured according to the group index $b_i$ inferred by the coarse-grained layered SBM. The description length $\Sigma$ [bits] and posterior odds ratio $\Lambda$ relative to the OG for each representation are also indicated. The OG for this network is $\{\{A, B\}, C\}$. This can be understood by realising that both layers A and B have CP structure so that the merged layer $\{A, B\}$ can be described more efficiently by just one CP model.*

group structure. Indeed, the wiring patterns in layers A and B can be summarised neatly by merging them, since the inferred groups in $\{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ and $\{\{\mathrm{A}, \mathrm{B}\}, \mathrm{C}\}$ are identical. In contrast, merging layers B and C induces a change in group structure where the distinction between the squares and triangles is lost and these two groups are aggregated into one with many internal interactions. It is worth mentioning that the $\{\mathrm{A}, \{\mathrm{B}, \mathrm{C}\}\}$ model is still a more appropriate description of the

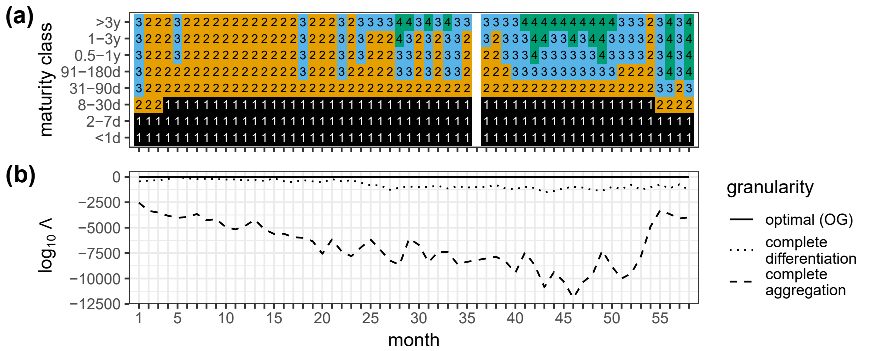



*Figure 2.4: The optimal granularity (OG) with respect to the loan maturity classes for the Russian interbank lending network. The OG corresponds to the maturity bin set $\{\ell\}$ parameter of the best-fit coarse-grained layered SBM. **(a)** Monthly time series of the OG inferred from the monthly interbank lending network. Each OG bin (OGB) holds one or more maturity classes and is labelled by an OGB index $1, 2, 3, 4$ and indicated by a colour. The OGB index runs from short to long maturities. The OGBs correspond to lending patterns between the bank groups that differ from each other in a statistically significant way. **(b)** Temporal evolution of the $\log_{10}$ of the posterior odds ratio $\Lambda$ (for its definition see Eq. 2.3) of the coarse-grained layered SBM for three different granularities: (i) the OG; (ii) complete loan maturity differentiation; (iii) complete loan maturity aggregation. Complete aggregation is decisively rejected as an optimal representation of the bank group structure of the monthly interbank lending network. The largest $\log_{10} \Lambda$-value different from zero is $\log_{10} \Lambda = -6.33$ and occurs for complete differentiation at $t = 5$.*

network than the $\{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ one.

## Fitting the coarse-grained layered SBM to the monthly ILNs

For each monthly ILN $\{G_l\}$ we fit the coarse-grained layered SBM by the MAP estimate of its parameters $(\{\theta\}, \{\ell\})$. From this we obtain a time series of the OGs and the bank groups $\{b_i\}$.

### The monthly OGs

Fig. 2.4 shows the monthly OGs together with the PORs relative to complete differentiation and complete aggregation. As noted before, each OG consists of a set of bins numbered from short to long maturity by the OG bin index (OGB index). The most important result is that complete aggregation is always rejected as the appropriate level of granularity for $\{G_\ell\}$. The maturity layers are thus informative of the monthly ILN's network structure in the sense that including them in a layered SBM yields an improved description compared to a layered SBM of the collapsed monthly ILN $\{G_c\}$, because the lending patterns between the bank groups depend significantly on the maturity classes of the loans. In other words, the lending patterns in the monthly ILNs $\{G_l\}$ correlate with the maturity classes in a way that cannot be captured completely by just considering the loans alone, i.e. without maturity information. This is also indicated by the fact that complete differentiation is seen to yield substantially better fits to $\{G_l\}$ than complete aggregation. The $\log_{10}\Lambda$ of Eq. 2.3 that measures the degree of rejection, tends to increase until roughly month 50 (February 2004), after which the degree of rejection becomes weaker. This aligns with the timing rumours surfaced about a large scale government investigation into money laundering by banks. This eventually caused several bank licenses to be withdrawn, see Appendix 2.F for a timeline.

The eight maturity classes reflect the CBR's reporting standards. One may therefore ask whether all eight classes also have an economic function in the actual lending and borrowing between banks. Interestingly, the OGs are always different from complete differentiation, which means that the maturity classes as defined by the CBR are partly redundant and that the lending patterns may be described more effectively by merging maturity layers into bins according to the OG. The OGBs are indicative of the fact that lending patterns between the bank groups can be combined in a more comprehensive form. For example, the short layers ($G_{<1\mathrm{d}}$, $G_{2\text{-}7\mathrm{d}}$, $G_{8\text{-}30\mathrm{d}}$) which are characterised by the ILN stylised features in Table 2.1 are almost always merged together. We emphasise that this does not necessarily indicate pointwise similarity between these layers; rather the merging of the layers in the OGB induce a new lending pattern between the bank groups that is significantly different from the other OGBs.

Figure 2.4(a) shows that the number of OGBs increases with time, which points to a developing interbank money market as more significantly differing lending patterns emerge between the bank groups. An additional argument is that the OGs differ more and more from the complete aggregation (see Fig. 2.4(b)). Indeed, the monthly number of bins in the OG correlate with the known two phases in the Russian money market's development: early development (before month 35) and emerging maturity (from month 35 onward) (see Methods). This indicates that the market's emerging maturity phase is characterised by a more complex layered SBM with up to four statistically significantly lending patterns between the bank groups. This is in contrast with the early development phase, where mostly only two lending patterns are discerned (maturities up to 30 days and maturities longer than 30 days). In other words, the interbank market can be characterised during the early development phase by the existence of only two distinct lending strategies between the bank groups.

### The bank groups

In Fig. 2.5(a) we display the number of groups $B$ $(1 \leq b_i \leq B)$ inferred for each monthly ILN. As $B > 1$ across time, the monthly ILNs do contain group structure.

Second, $B$ follows the same upward trend as in Figs. 2.2(a),(b); 2.4(b), indicating that the increase in the number of groups goes hand in hand with the development of the Russian ILN into a more mature phase as the months in our dataset pass.

We now gauge the time correlations between the inferred monthly group structure by looking at the similarity between the inferred bank groups $\{b_i\}$ in consecutive months. Fig. 2.5(b) shows the normalised mutual information (NMI) between the $\{b_i\}$ in a given month $t$ and the previous month $t-1$. In the Russian ILN, one discerns non-important banks. They display little activity over time and volume, and accordingly they are inclined to fluctuate between groups. To account for this, we condition the NMI on $q$, a measure to control which bank strengths [63] are included. The strength $s_i^l$ of a bank $i$ is defined by the total amount it borrows and lends in a layer $l$ during a certain month:

$$s_i^l = s_i^{l,\text{in}} + s_i^{l,\text{out}} = \sum_{j,k} x_{jik}^l + \sum_{j,k} x_{ijk}^l \; . \tag{2.4}$$

We also define the size of a maturity layer $S^l = \sum_{i<j,k} x_{ijk}^l$ as the total amount borrowed or lent during a certain month. The size of the monthly ILN is given by $S = \sum_l S^l$. To construct Fig. 2.5(b), we calculate for each month $t$ the relative strength of each active bank $s_i = \sum_l s_i^l / 2S$. Then we create a list of banks by adding one bank at a time, *starting out with the strongest bank and proceeding in order of decreasing bank strength*, until the cumulative relative strength reaches $q$, and we include only the banks on this list in the analysis. In this way, we exclude banks that are only responsible for an insignificant amount of lending and borrowing in the network. We intersect the included banks with those at time $t-1$ and calculate the NMI from the two intersected bank group memberships for $q = 0.95, 0.99, 1$. Figure 2.5(b) shows that with decreasing $q$ the NMI grows, pointing to increasing similarity between the inferred bank groups at consecutive months. It is hard to draw conclusions from the NMI without a reference. By excluding the "5% least important banks by strength" ($q = 1 \longrightarrow q = 0.95$) a considerable increase in correlations between the inferred bank groups in consecutive months is observed. This is indicative of the temporal stability of the inferred groups. This temporal stability emerges from the analysis without explicitly imposing intertemporal correlations in the algorithm. The temporal stability of the inferred groups is indicative for the robustness of the findings and corroborates the role of relationship lending in interbank markets. Explicitly modeling intertemporal correlations by means of a dynamic layered SBM model [19, 64] turned out to be computationally prohibitive when applied to our data. Another advantage of the adopted "static" layered SBM methodology is that it can be applied to both time series of ILN data and ILN data covering a specific time period. In addition, treating the monthly data independently provided an additional robustness check as the anticipated intertemporal correlations are recovered by the algorithm. Inversely,

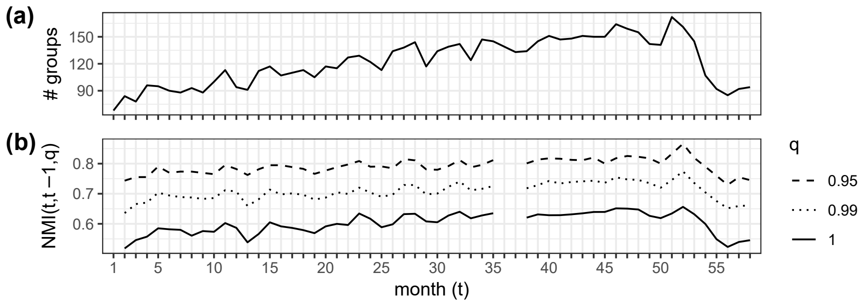



*Figure 2.5: Information about the bank groups $\{b_i\}$ for the monthly Russian interbank lending networks as inferred from the optimal granularity (OG). **(a)** Temporal evolution of the number of different bank groups. The number of groups roughly increases until $t = 50$. This indicates that the mesoscale structure of the monthly interbank lending networks becomes progressively more complex as more bank groups are needed to fit the lending patterns. **(b)** The temporal evolution of the normalised mutual information (NMI) between the OG group memberships of the set of banks active in two consecutive months $(t, t-1)$ for three strength fractions q. For example, for $q = 0.95$, banks responsible for 95% of the network's lending and borrowing are included. In many fields, including community detection, the NMI is a popular quantity to measure the similarity between two partitions of a set [65].*

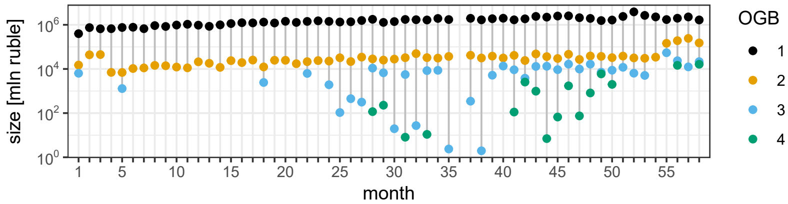



*Figure 2.6: Correlations between aggregate loan sizes and lending patterns: the size of the optimal granularity bins (OGBs) through time. The size of an OGB is defined as the total amount of issued money in the interbank lending network defined by the OGB. The OGB indices and colours correspond to those of Fig. 2.4(a).*

our results also indicate that intertemporal correlations can be exploited to reduce the computational complexity of the modelling of "current" ILN data by using the inferred structure of the ILN in preceding months as a reasonable initial estimate for the underlying structure.

#### 2.4.0.1 Characterisation of the OGBs

The OGBs can be interpreted as corresponding with statistically significant differing lending patterns between the bank groups. For consistency, we numbered the bins in the OG with an OG bin index (OGB index) from short to long maturity.

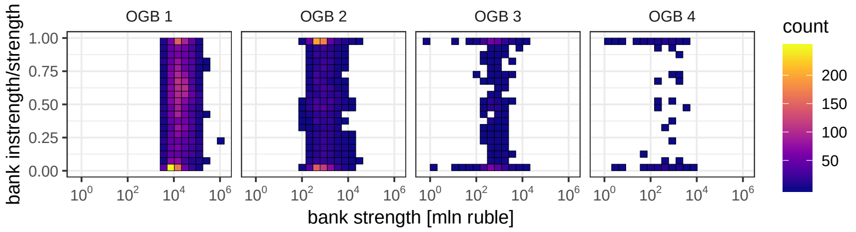



*Figure 2.7: The time-integrated distribution of the strength and the instrength/strength ratio of the monthly "important banks" for the four optimal granularity bins (OGBs). As in Fig. 2.5(b) we gauge a bank's importance by its strength: For a given month a bank is deemed important if its strength lies in the top 10%. With growing OGB index, the important banks increasingly tend to either lend or borrow. Banks in OGB 1 and OGB 2 tend to balance lending and borrowing. This suggests that the economic function of the important banks in the various OGBs changes from financial intermediation (short maturity bins) to financing (long maturity bins) on a monthly time scale.*

Even though their actual content can vary considerably through time, we find that this numbering scheme reveals surprisingly consistent patterns.

At the *network level*, Fig. 2.6 displays the OGB sizes $S^\ell = \sum_{l\in\ell} S^l$ throughout time. We see that the order of magnitude of the OGBs corresponding with the shorter maturities stays consistent across time, even though the composition of the OGBs (i.e. the maturity layers in the OGBs) changes considerably through time. This is especially so for OGB 2 (see Fig. 2.4(a)): it contains the long maturity layers during the early development phase but not during the emerging maturity phase. It is interesting to note that while the sizes of the "longer" OGBs are much smaller than the short ones, these are not merged in the emerging maturity phase of the Russian ILN. This points to lending patterns that are sufficiently different from those in the short OGBs. (This still holds, except in one month, for the non-contiguous binning in Appendix 2.B.)

Finally, we look at the *bank level*, i.e. the lending behaviour of individual banks. Again we use the bank strength to single out the "important" banks. In Fig. 2.7 we have shown the time-integrated distribution of $s_i^{\text{in},\ell}/s_i^\ell$ of the top 10% most important banks of each month, separately for each OGB $\ell$. (Note that the conclusions hold for other cutoffs – see Appendix 2.E.) The instrength of a bank $i$ in maturity bin $\ell$ during a certain month is $s_i^{\ell,\text{in}} = \sum_{l\in\ell}\sum_{j,k} x^l_{jik}$, and the strength $s_i^\ell$ is defined analogously as in Eq. 2.4. At monthly time scales the important banks in "short" bins tend to both lend and borrow equal amounts of cash, while the important banks in the "long" bins tend to either lend or borrow, i.e. act either as sources or sinks of liquidity. Together with the indications for CP structure in the short layers, this suggests that the economic function of the important banks changes from financial intermediation (short bins) to financing (long bins). The apparent patterns in the short bins are indeed reminiscent of the functions in CP structures, while those in the long bins have, to the best of our knowledge, not often been included in the literature. Thus our coarse-grained layered SBM of the ILN uncovers a correlation between statistically significant lending patterns between the bank groups, the economic functions of the important banks, and the maturity classes of the loans involved.

# Discussion

In this paper, we investigate the importance of loan maturity information in interbank lending networks towards understanding its mesoscale structure, i.e. the higher-level organisation of the banks into groups. We do this to better understand the possible diversity in lending and borrowing patterns between bank groups in ILNs and their accompanying economic functions. We find that the representation of ILNs common in the literature, where either one maturity layer is studied or an aggregated view on the maturity layers is used, is unable to fully uncover information on the diversity in structures and underlying functions of interbank loans. Even after introducing multi-directedness, degree distribution, and loan size distribution, the economic functions carried out by banks are obscured by aggregating maturity information in the ILN representation.

We employ the complete population of all Russian unsecured interbank lending contracts over a 57 months period (January 2000 - October 2004, except January 2003) containing uniquely granular loan maturity classes as defined by the CBR. Descriptive statistics on the associated maturity layers uncover non-homogeneity in network measures along loan maturity layers, indicating a diversity in lending and borrowing patterns.

To investigate the optimal maturity granularity, we apply a coarse-grained layered stochastic block model [19] to the data. The subsequent analysis confirms the suboptimality of complete maturity aggregation for our dataset, as it obscures the existence of four maturity layer bins that contain significantly distinct lending- and borrowing patterns with various underlying functional economic interpretations. We find, for example, a consistent shorter term maturity layer bin that behaves in line with the theory around tiered banking, with important banks intermediating liquidity. We also detect another layer bin that aligns with long-term financing of bank activities, with the important banks acting either as sources or sinks of liquidity.

These findings immediately imply that the common practice of complete maturity collapse or focus on a single maturity layer, by choice or for data limitations, obscures important information about the functions banks perform in an ILN. This leads potentially to unrealistic ILN models and misguided policy conclusions about systemic stability, especially so in times of liquidity crunches when the short term layer of the ILN issuance network collapses.These insights also align with and add to recent findings that banks use the different loan maturities to manage their duration gap, providing a direct link from interbank markets to financial stability and the allocative efficiency of the economy at large [2].

The lending patterns between the bank groups depend not only on the loan maturities but also on the phase of development of the interbank money market. The longest maturity loans, for example, only show structure related to interbank financing at later phases of market development. This makes economic sense since long-term loans entail greater counter-party risk and counter-party trust is only established over time through engagement in long term relationships.

Our analysis builds on the strand of interbank literature that leverages the network representation of interbank systems. Thereby, one often starts from the topology of exposure networks while largely disregarding loan maturity and other meta-data on the banks and their connections. Our work conclusively illustrates that after accounting for distinct loan issuance and degree- and loan size distributions, loan maturity is still informative of structure and economic functions. Alternate approaches have specifically focused on systemic risk and contagion and concluded that funding risk and CDS spreads play a central role in the mechanisms of risk propagation. It remains an open question how the structural information in maturity layers could complement these approaches. The trust crisis in Russia has indicated that the occurrence of a non-fundamentals crisis is clearly reflected in both the interest rates and the mesoscale structure of the maturity layers. This connection offers opportunities for future research.

All in all our results imply that theoretical and empirical research can neither adequately grasp the generative process of ILNs nor arrive at reliable policy conclusions from ILN modelling and simulation in the absence of appropriate granular maturity information. This underlines the importance for policy and regulatory bodies to collect maturity information on interbank loans if they desire to arrive at reliable insights into the health and systemic stability of the interbank lending market. It should also stimulate further theoretical and empirical research that incorporates loan maturity in modelling of the ILN generative process, its dynamics and its occasional transition to phases of instability or collapse.

# Methods

## Dataset

### Overview

The interbank data analysed in this work offers a rich and unique account of Russian commercial bank activities over a six-year timeline (August 1998 - October 2004). The data is provided by Schoors and Karas [17] and has been painstakingly assembled from public and private sources. It originates directly from the information reported to the CBR about interbank contracts. The dataset used in this paper can be retrieved on demand from the authors. To the best of our knowledge the dataset is unique [13] in quite a number of aspects. We mention the availability of information about the issuance months of loans (most often only the derived exposure is available) and rather detailed granular information about the maturity (see Appendix 2.G). The fact that the Russian interbank market went through many stages of development during 1998-2004 adds an additional layer of dynamics. Note that the CBR is not included as a node and that we do not have information about any of its transactions. A classification of the banks together with a recent description of the money market can be found in [46]. On average, about half of the Russian banks are active on the interbank market [24]. The dataset starts a few weeks before the so-called "1998 Russian default" [66], which caused a complete collapse of the interbank money market. In response the CBR imposed to little avail several extraordinary measures to stabilise the market. These exceptional circumstances greatly disrupted the workings of the interbank market and because of this we have restricted our analysis to the period of 57 months from January 2000 until October 2004. In our numbering scheme month 1 corresponds to January 2000, month 13 to January 2001, and month 58 to October 2004. As can be inferred from Fig. 2.2, during months 1-35 the market is developing. We refer to this period as the early development phase. Roughly starting from month 35 (December 2003) the interbank market enters an emergent maturity phase: the amount of active banks stabilises and the share of overnight loans declines in favour of longer-maturity loans. The emergent maturity phase includes the trust crises in the second half of 2003 and the summer of 2004 caused by a money laundering scandal (see Appendix 2.F for more details). As compared to the Russian default of 1998, the crises of 2003 and 2004 were less disruptive for the interbank lending market.

### Contents

The data consists of the issuance of domestic unsecured interbank loans, annotated by lender bank ID, borrower bank ID, month of issuance, loan size, interest rates and maturity class. As noted earlier, the loan interest rates are not incorporated into the coarse-grained layered SBM; we discuss them in Appendix 2.H. The loan issuances are reported to the legislator on the first day of each month throughout August 1998 up until November 2004 in one of the following eight maturity classes: overnight (less than one day, <1d), 2-7 days (2-7d), 8-30 days (8-30d), 31-90 days (31-90d), 91-180 days (91-180d), 0.5-1 year (0.5-1y), 1-3 years (1-3y), more than 3 years (>3y). Because of this reporting standard, the precise issuance month of the loans is known. For January 2003 (month 36) there is no data available. Whenever possible, we interpolate the missing month in the time series. For some variables (Figs. 2.4(a) and 2.5(b)) this is not feasible and this is at the origin of the missing data point in the time series.

#### Aggregation window

Loans are aggregated by issuance month into monthly ILNs (as in [24]) for two reasons. First, monthly aggregation is the most granular time scale available in the data. Second, monthly compliance with regulatory requirements for banks (e.g. for liquidity and capital) induces a monthly periodicity in the data.

### Implementation

#### Inference algorithm with agglomerative hierarchical clustering

We infer the OG for a layered network using an agglomerative hierarchical clustering heuristic as suggested in [19]. At the start of the procedure each layer is put in its own bin. In the next steps, bins are merged so as to reduce the overall DL. The overall DL consists of the DL of the layered SBM plus an "SBM extension term" that accounts for the model selection between the possible levels of granularity (Eq. 17 in [19]). With contiguous binning only contiguous bins are merged, while for non-contiguous binning any pair of bins may be merged at each step. In this way a series of layered networks is generated, starting with the original layered network and ending with the collapsed network. The layered network with the smallest overall DL defines the OG.

We use the efficient inference algorithm [67] implemented in the graph-tool library [61] to fit the layered SBMs to the monthly ILNs. The algorithm employs an agglomerative heuristic to fit the layered SBM and a multilevel Markov Chain Monte Carlo to sample the posterior distribution. We used High Performance Computing resources to perform four runs of numerical calculations: Two of those used contiguous binning and two of those used non-contiguous binning for the maturity layers. In each run we used agglomerative hierarchical clustering to fit the layered SBM and to sample the posterior for the levels of granularity generated for each of the monthly ILNs. The number of samples were set to 10,000 and 25,000 for both the contiguous and non-contiguous runs. All runs (including the test runs) yielded similar results for each monthly ILN which alludes to the stability of the posteriors. The layered SBMs with smallest DLs were gathered from both runs.

# Appendix

## 2.A Maturity layer analysis

Here we give a detailed description and network analysis of the Russian interbank lending network to compare to other interbank networks used in research. We divide the Russian interbank loan network $\{G_\alpha\}$ by maturity class $\alpha$ into separate layers consisting only of the loans with loan maturity inside each maturity class. (Note that in the main text we use the symbol $l$ instead of $\alpha$.) We evaluate the properties per maturity layer (often also referred to as layer for simplicity) to show the differences between them. We denote the eight maturity layers as: $G_{<1d}$, $G_{2-7d}$, $G_{8-30d}$, $G_{31-90d}$, $G_{91-180d}$, $G_{0.5-1y}$, $G_{1-3y}$, $G_{>3y}$.

### 2.A.1 Activity and volume

One expects that the number of banks participating in a given loan maturity layer decreases as the loan maturity lengthens. Table 2.A.1 shows that this is approximately correct. More importantly, one sees that the number of active banks becomes very small for $G_{1-3y}$ and $G_{>3y}$. We define an active bank as a bank that has lent or borrowed at least once in a given time window and maturity layer. We will always only consider active banks to deal with the death and birth of banks in the Russian interbank market over the course of time [68]. For the layers with long loan maturity, this results in an exceedingly high average lending activity per bank, comparing to layers with shorter loan maturity. It seems that $G_{>3y}$ is occupied by a small club of banks that trade relatively intense.

Table 2.A.2 lists the yearly, and average monthly lending activities by loan maturity class, where we define the *lending activity* simply as the number of loans recorded during a certain period. A crucial observation is that the lending activity sharply decreases for longer maturities, which is also reported for other interbank markets in [1]. We see that the overnight segment ($G_{<1d}$) is the most active, together with $G_{2-7d}$.

Table 2.A.3 lists the total loan volumes (i.e. loan sizes) by loan maturity layer and year. We observe that the relative importance of each maturity segment, as measured by the total volume of loans traded within it, follows the ranking of the loan maturity lengths. The loan volumes are log-normally distributed [24],

| | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|---|---|
| <1d | 610 | 810 | 878 | 954 | 1,022 | 1,027 | 983 |
| 2-7d | 589 | 907 | 953 | 1,026 | 1,074 | 1,098 | 1,058 |
| 8-30d | 375 | 849 | 920 | 980 | 1,018 | 1,054 | 986 |
| 31-90d | 166 | 535 | 648 | 698 | 706 | 775 | 686 |
| 91-180d | 63 | 203 | 265 | 310 | 321 | 429 | 378 |
| 0.5-1y | 49 | 120 | 132 | 169 | 204 | 258 | 259 |
| 1-3y | 24 | 64 | 80 | 60 | 58 | 113 | 107 |
| >3y | 11 | 45 | 29 | 38 | 33 | 23 | 20 |

| | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|---|---|
| <1d | 13 | 118 | 187 | 258 | 286 | 337 | 225 |
| 2-7d | 7 | 57 | 108 | 161 | 183 | 229 | 147 |
| 8-30d | 3 | 15 | 32 | 43 | 44 | 59 | 43 |
| 31-90d | 1 | 5 | 10 | 11 | 15 | 18 | 12 |
| 91-180d | 1 | 2 | 3 | 4 | 5 | 7 | 5 |
| 0.5-1y | 1 | 2 | 2 | 4 | 4 | 7 | 7 |
| 1-3y | 2 | 3 | 3 | 3 | 3 | 8 | 8 |
| >3y | 5 | 3 | 6 | 10 | 78 | 74 | 44 |

*Table 2.A.1: (top panel) The number of active banks involved in loans by maturity layer and year. A total of 1,040 banks are present in the data. (lower panel) The number of loans per active bank by maturity layer and year, obtained by dividing the top panel of Table 2.A.2 and the top panel of this Table element-wise. Care must be taken when interpreting; this is not indicative of an 'average' bank since the degree distributions are heavy-tailed (as will be discussed below). Also note that the columns '1998' and '2004' are biased because they are incomplete in the data (respectively 5 and 10 months missing months).*

especially for shorter loan maturities. Because of the considerable variance on a linear scale, these averages may be interpreted only as rough order of magnitude estimates. In general, log-normal distributions are reasonable descriptions for observers that only know the order of magnitude of the mean and the variance of a random variable [52].

| | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|---|---|
| <1d | 7,917 | 95,389 | 163,855 | 245,840 | 292,048 | 346,221 | 220,979 |
| 2-7d | 4,325 | 51,673 | 103,376 | 165,602 | 196,765 | 251,006 | 155,834 |
| 8-30d | 958 | 12,914 | 29,250 | 42,067 | 45,204 | 62,190 | 42,181 |
| 31-90d | 245 | 2,770 | 6,365 | 7,939 | 10,606 | 14,184 | 8,443 |
| 91-180d | 82 | 460 | 742 | 1,189 | 1,754 | 2,797 | 2,010 |
| 0.5-1y | 63 | 237 | 307 | 593 | 886 | 1,901 | 1,899 |
| 1-3y | 52 | 207 | 279 | 202 | 180 | 914 | 845 |
| >3y | 55 | 128 | 180 | 367 | 2,588 | 1,712 | 890 |

| | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|---|---|
| <1d | 1,583 | 7,949 | 13,655 | 20,487 | 26,550 | 28,852 | 22,098 |
| 2-7d | 865 | 4,306 | 8,615 | 13,800 | 17,888 | 20,917 | 15,583 |
| 8-30d | 192 | 1,076 | 2,438 | 3,506 | 4,109 | 5,182 | 4,218 |
| 31-90d | 49 | 231 | 530 | 662 | 964 | 1,182 | 844 |
| 91-180d | 16 | 38 | 62 | 99 | 159 | 233 | 201 |
| 0.5-1y | 13 | 20 | 26 | 49 | 81 | 158 | 190 |
| 1-3y | 10 | 17 | 23 | 17 | 16 | 76 | 84 |
| >3y | 11 | 12 | 15 | 31 | 235 | 143 | 89 |

*Table 2.A.2: (top panel) Lending activity, defined as number of of loans, by loan maturity class and year. Note that 1998 and 2004 are incomplete years in the data, counting 5 and 10 months respectively. (lower panel) Monthly average lending activity by loan maturity layer. Time series of the lending activity can be found in Fig. 1 in the main text.*

| | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|---|---|
| <1d | 226.3 | 2,674.4 | 5,984.3 | 8,855.9 | 10,306.4 | 13,844.6 | 13,040.1 |
| 2-7d | 83.1 | 1,030.4 | 2,910 | 5,114.7 | 6,104.6 | 9,058.1 | 7,880.2 |
| 8-30d | 15.9 | 220.7 | 620.8 | 1,103.1 | 1,102.9 | 1,871.8 | 1,833.3 |
| 31-90d | 4.4 | 31.9 | 105.2 | 205.3 | 318.3 | 415.4 | 253.6 |
| 91-180d | 1.7 | 5.3 | 9.6 | 23.4 | 46.9 | 73.6 | 73.7 |
| 0.5-1y | 3.2 | 2.4 | 8 | 20.2 | 24 | 50.8 | 69.9 |
| 1-3y | 2.2 | 5.5 | 6 | 6 | 3.3 | 10.1 | 11 |
| >3y | 2.6 | 0.8 | 1.2 | 1.4 | 1.3 | 0.8 | 2.5 |

*Table 2.A.3: Total loan volumes by loan maturity and per year in billions of rubles.*

### 2.A.2 Maturity layer activity of banks

A node $i$ is defined as active on layer $\alpha$ if it has at least one connection on this layer, i.e. its degree $k_i^\alpha > 0$. In symbols,

$$b_i^\alpha = \begin{cases} 1 & \text{if } k_i^\alpha > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

The *total activity* $B_i$ measures the number of layers the node participates in, i.e. $B_i = \sum_\alpha b_i^\alpha$ [9]. Figure 2.A.1 shows the distribution of $B_i$ on monthly time scales. We see that the average total activity grows steadily as the network develops, finally settling around at a value of about three. Almost no banks make use of more than six maturity layers on a monthly basis. The distribution of the total activity is quite broad and relatively unpeaked, especially during the stable phase of the network. This has been reported as a typical quality of real-world multiplex network [69].

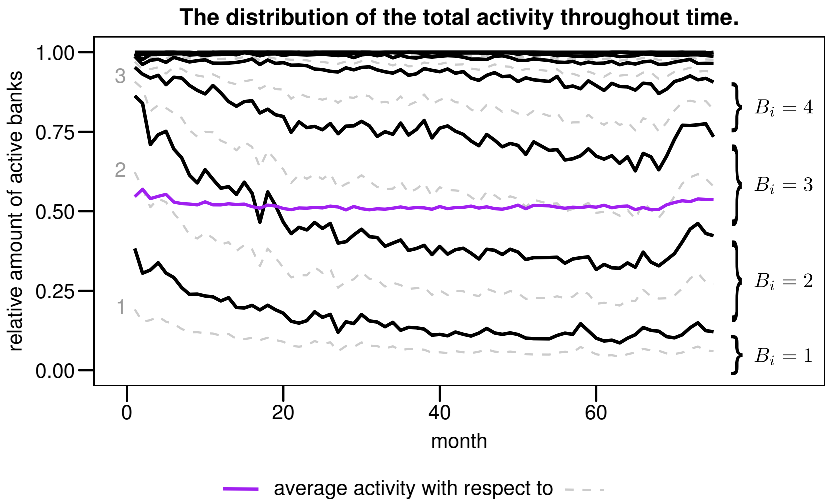



*Figure 2.A.1: For a given month, the area between the thick lines is proportional to the relative number of active banks with a certain $B_i$ value, which is indicated on the right. The dashed lines make up the moving coordinate system of $\langle B_i \rangle$, in relation to which the average total activity indicated by the thick purple line must be understood. For example, $\langle B_i \rangle$ is about 2 (3) for month 1 (75). The summer of 2004 crisis is mirrored by the drop in $\langle B_i \rangle$ during the last months.*

### 2.A.3 Density

Real-world interbank networks are typically *sparse*, meaning that in directed and undirected views of the network only a small fraction of all possible edges exist. One can then ask how the edges are distributed among the banks. To this end, we define the *degree* of a node. For undirected networks, this is the number of edges connected with a given node, i.e. the number of the bank's counterparties. For directed networks, the in-degree (out-degree) of a node is the number of incoming (outgoing) edges. The in-degree (out-degree) of a bank is then simply the number of loans borrowed (lent) by it.

The above seen differences in lending and bank activity per loan maturity layer is reflected in the directed density of the layers, shown in Fig. 2.A.2. All layers can be considered sparse, except for $G_{1-3y}$ and $G_{>3y}$, which have respectively moderate and an extremely high density [70]. We also observe that the density in almost all layers grows steadily as the network develops, pointing to an increasing interconnectedness with increasing development of the market [71].

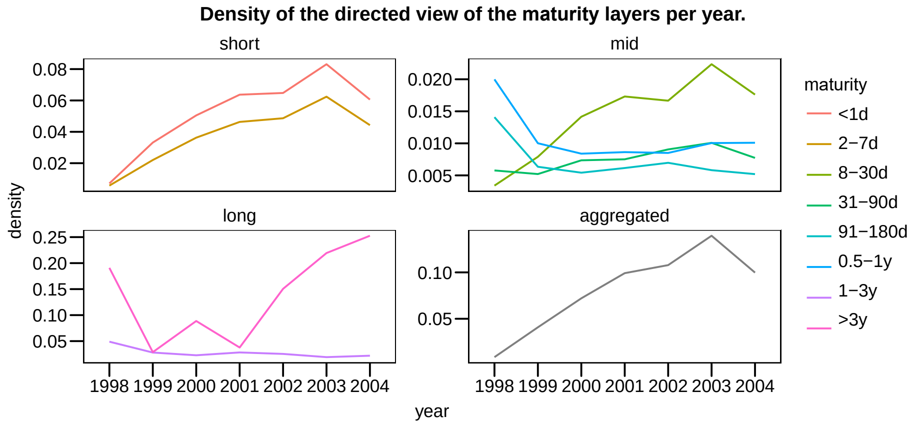



*Figure 2.A.2: The yearly density per loan maturity layer in the directed Russian interbank lending network. The directed density for a simple graph with $N$ nodes and $E$ edges is defined as $d = \frac{E}{N(N-1)}$.*

### 2.A.4 Degree distribution

Interbank networks, like many real-world networks [43], exhibit heavy-tailed degree distributions. In a nutshell, this means that *few nodes have many links and many nodes have few links*. In general, many typical distributions are heavy-tailed – in fact, they abound in descriptions of natural events like avalanches, earthquakes, turbulent flow and rainfall [72, 73]; another example in economics are the non-Gaussian return distributions in financial markets.

Early work in the previous decade tended to postulate power laws for the degree distributions. Networks whose degree distributions follow a power law (at least asymptotically) are called *scale-free* networks. However, while agreeing on the heavy-tail character, recent literature has cast serious doubt on the idea that power laws are the best candidate for the degree distributions, and thus on the scale-free character of interbank networks. In our case, power laws have been decisively rejected as best fit candidates for the heavy-tailed degree distributions of the ¡1d and 2-7d loan maturity by Vandermarliere et al. (2015) [24].

We plot the heavy-tailed degree distributions with complementary cumulative distribution functions (ccdfs) [73] with doubly logarithmic scales. If we denote the degree and its distribution by $k$ and $p(k)$ respectively, the ccdf is given by

$$\mathrm{ccdf}(k) \equiv 1 - \mathrm{cdf}(k) = P\left(k \leq k'\right) = \int_k^\infty p(k')\,\mathrm{d}k', \tag{2.6}$$

where $\mathrm{cdf}(k)$ is the usual cumulative distribution function.

The degree distributions are shown per loan maturity layer in Figures 2.A.3 and 2.A.4. Each loan maturity layer exhibits fat-tailed in- and out-degree distributions, with the maximum degrees separated some 5 ($G_{>3y}$, in-degree) to 18 ($G_{8-30d}$, in-degree) standard deviations from the mean. Vandermarliere et al. (2015) [24] established that stretched exponentials of the form

$$f(d) = Cd^{\beta-1}\exp\left(-\lambda d\right)^\beta \tag{2.7}$$

provide the best overall fit for the bulk+tail multi-directed in- and out-degree for monthly and yearly time windows. In Eq. 2.7, $d$ is the in-degree or out-degree, $C$ a normalization constant and $\lambda$, $\beta$ distribution parameters. The stretched exponential can be understood as a Weibull distribution, with $\beta$ being the shape parameter, and $1/\lambda$ the scale parameter. If $0 < \beta < 1$, the distribution has a fat tail, with a smaller $\beta$ putting more weight towards smaller $d$. $1/\lambda$ widens the distribution, the mean of $f$ being proportional to it.

We fitted $f$ to the degree distributions to see if the result by Vandermarliere et al. (2015) [24] can be extended to longer maturities. The conclusions are identical for the in- and out-degrees. First, the stretched exponential fit does seem to describe the time-aggregated degree distributions of $G_{<1d}$, $G_{2-7d}$ and $G_{8-30d}$. Then the leap in residual sum-of-squares suggests that $G_{31-90d}$ and longer maturities can not be satisfactorily described by the stretched exponential. This is confirmed by visual inspection. While (parts of) the bulk distributions of the longer loan maturities are reasonably well captured by $f$, the fits systematically underestimate the ccdf in the tail; put differently, they underestimate for a given large degree $d$ how many nodes exists with an even larger degree $d'$.

To investigate whether frequent lenders are also frequent borrowers, we first rank the banks with regard to in- and out-degree, then we calculate Kendall's coefficient of concordance $W$, which measures the rate of agreement $0 < W < 1$

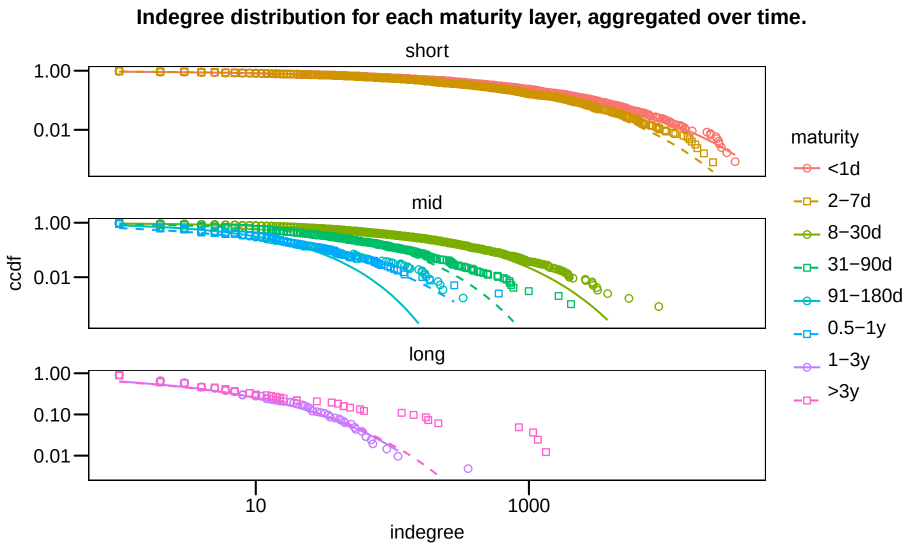



*Figure 2.A.3: In-degree distribution with attempts to fit the bulk+tail with stretched exponentials, drawn as piecewise linear functions connecting each predicted value. The correspondence with this distribution seems to break down for layers with loan maturity longer than 8-30d. Inferred* $(\lambda, \beta)$ *parameters for* $G_{<1d}$, $G_{2-7d}$, $G_{8-30d}$ *are* $(2.6e-3, 4.3e-1)$, $(3.4e-3, 4.8e-1)$, $(1.0e-2, 5.8e-1)$ *respectively, with all standard errors below 1%.*

between two rankings, per loan maturity layer. The significant results together with the joint degree distributions are shown in Fig. 2.A.5. We conclude that lending and borrowing are highly correlated for $G_{<1d}$, $G_{2-7d}$, and $G_{8-30d}$; in other words, the dominant lenders are likely to be dominant borrowers, and vice versa. This correspondence breaks down markedly for the layers with longer loan maturity, with orderly decreasing $W$ but insignificant test results. This can also be seen in the shapes of the contour plots in Fig. 2.A.5: high in-degree - out-degree correlation is marked by squeezed contours around the plot diagonal, which does not hold for maturity longer than 31-90d. $G_{1-3y}$ and $G_{>3y}$ even display modest anticorrelation for the degree tails, meaning that the most frequent lenders (borrowers) area unlikely to be among the most frequent borrowers (lenders).

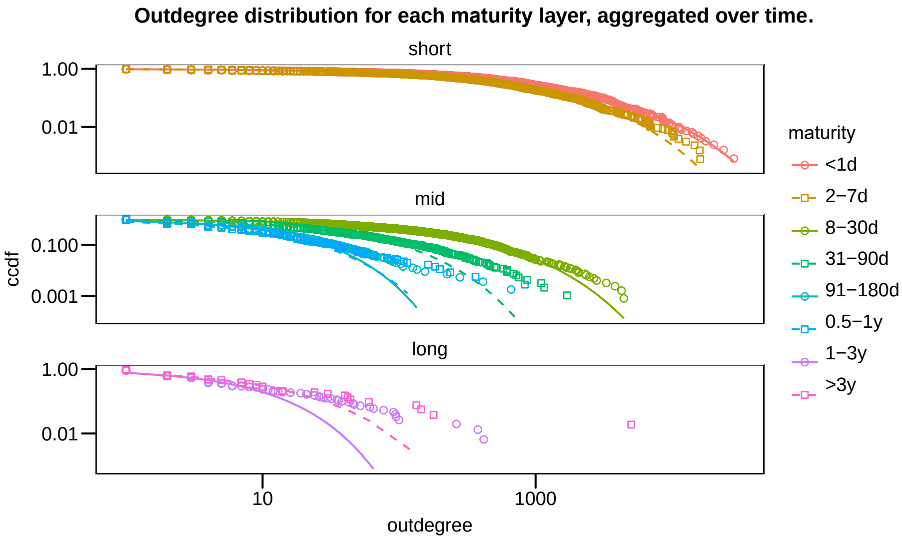



*Figure 2.A.4: Out-degree distribution with with attempts to fit the bulk+tail with stretched exponentials, drawn as piecewise linear functions connecting each predicted value. It is clear the correspondence with this distribution breaks down for layers with loan maturity longer than 8-30d. Inferred* $(\lambda, \beta)$ *parameters for* $G_{<1d}$, $G_{2-7d}$, $G_{8-30d}$ *are* $(1.5e-3, 5.3e-1)$, $(2.3e-3, 5.7e-1)$, $(8.3e-3, 6.1e-1)$ *respectively, with all standard errors below 1%.*

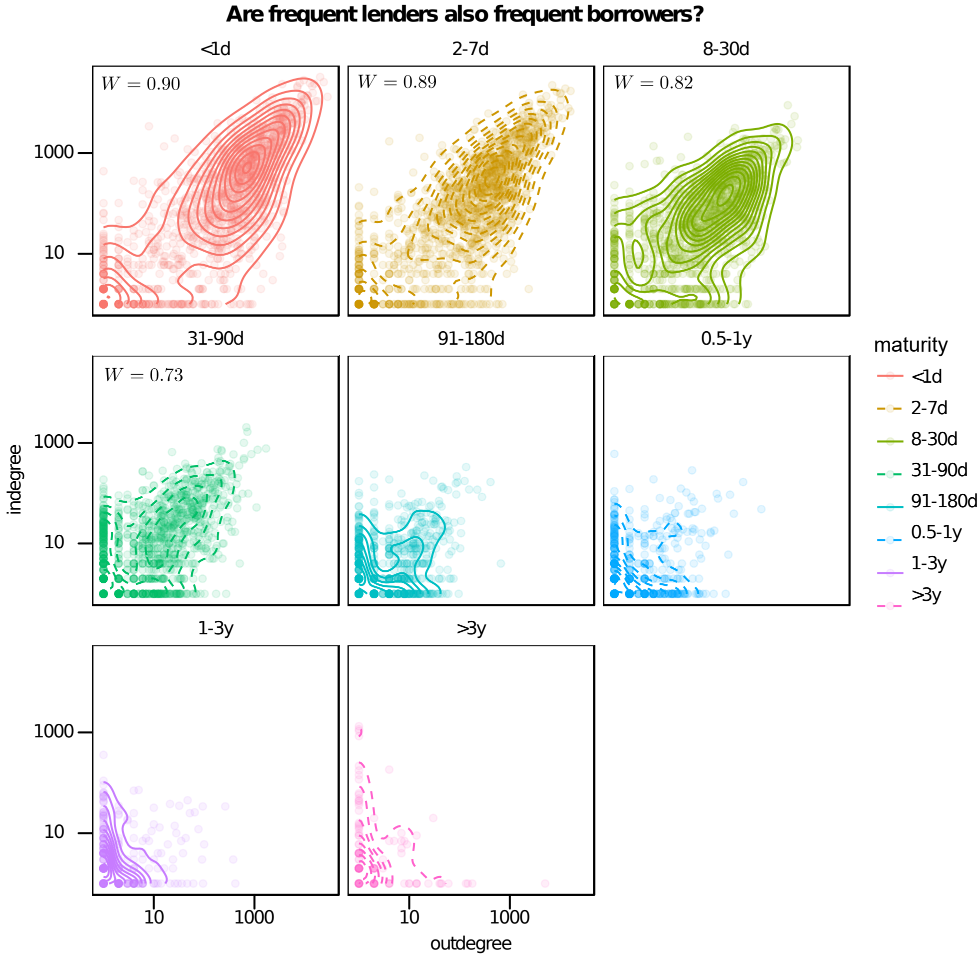



*Figure 2.A.5: The time-aggregated joint degree distribution for each layer, drawn as density contour plots with almost transparent data points. For the first four maturity bins, a highly significant Kendall $W$ was obtained and put in the top left corner of the panel. Note that all scales are equal for easy comparison.*

### 2.A.5 Average shortest path length

The average shortest path length (denoted from here on as just 'average path length') $D$ is defined for the largest connected component of the undirected view of interbank networks. It indicates the typical distance between two randomly chosen nodes, i.e. the smallest number of edges needed to reach one node from the other. *Small-world networks* are characterized by high clustering and small average path length. Both should be compared to their mean values in an ensemble of random networks with the same number of nodes and edge [74].

Most studies find that $D$ is small for interbank networks, which indicates compact network structure, but not all conclude that they are small-worlds. Hubs tend to lower the average path length; scale-free networks are ultra-small-worlds. $D$ is a measure of the typical length of intermediation chains that are taking place among the market participants (at least in the largest connected component). Longer intermediation chains arise when $D$ is large, which effectively contribute to slowing down the market transactions between participants and consequently harming the liquidity allocation between financial institutions. In contrast, when $D$ is small, the information between the market participants flows quickly in the network, giving rise to a well-functioning liquidity allocation in the market [27].

In the case of the Russian interbank network, the average path length peaks at times when the network is in crisis, and decreases gradually when the network is maturing. The same patterns have been reported in other empirical literature, such as in [26].

Figure 2.A.6 shows that exclusively for the layers with loan maturity ¡90d the average path length $D$ *decreases* roughly linearly when the size of the largest connected component (LCC) increases *logarithmically*. Herein the LCC covers the vast majority of the active banks within yearly time windows, see Fig. 2.A.7. Thus we see that core-periphery structure, which we expect in these layers from our analysis in the Topology section, is extremely effective in channeling and intermediating liquidity; of course this comes at the cost of systemic risk in the core hubs.

The remaining layers display the familiar pattern: the size of the largest connected component shrinks for longer maturities, while simultaneously the network becomes increasingly fragmented as the number of unconnected components increase, relative to the number of active banks (Fig. 2.A.8) – both tend to suppress $D$.

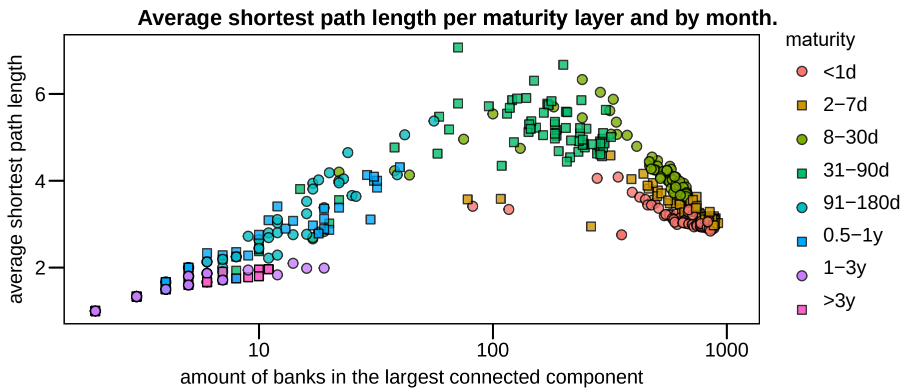



*Figure 2.A.6: The average (shortest) path in function of the size of the largest (weakly) connected component. For the four shortest loan maturity layers, the points lying astray from the dense cloud all occur during the first five months. Note that the number of banks in the largest connected component for these layers is notably larger than those of the remaining layers.*

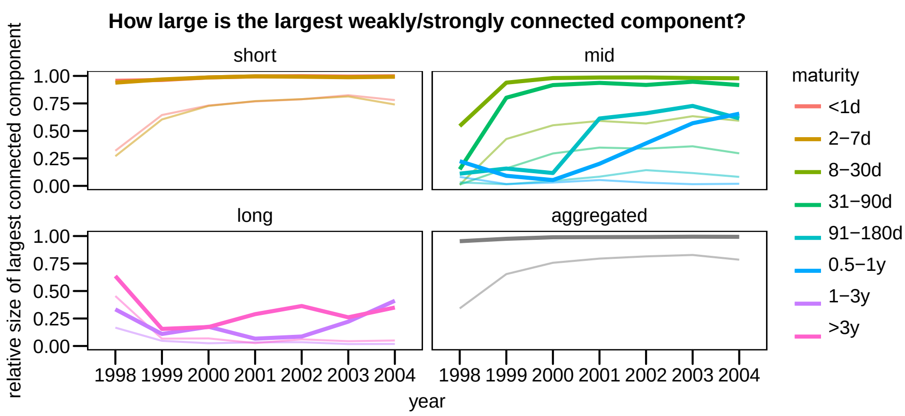



*Figure 2.A.7: The relative size of the largest connected component, given the weak or strong requirement, is calculated by dividing the number of banks in the largest connected component by the total number of active banks in a specific layer and year. The fat (thin) lines indicate the weak (strong) version.*

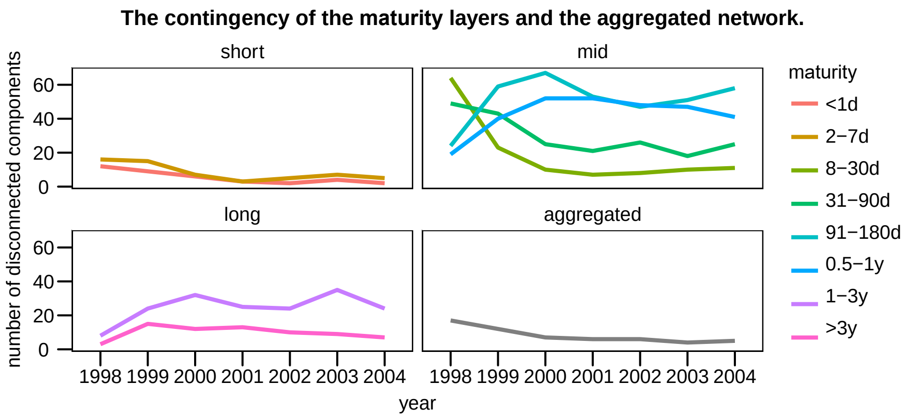



*Figure 2.A.8: To compare the number of (weakly) disconnected components with the number of active banks per year, consult Table 2.A.1 on page 48. The relative size of the largest weakly connected can be found in Fig. 2.A.7.*

### 2.A.6 Clustering coefficient

Clustering coefficients measure to which extent banks form triangles. Put differently, they measure the tendency of connected nodes to have common neighbors in undirected views of interbank networks.

According to Bargigli et al. [1], an inverse relationship between the degree and the clustering of a node is observed quite commonly. In the core-periphery picture, low clustering values of core nodes indicate that they essentially behave as star centers. A star graph exhibits zero clustering, as the periphery nodes are unconnected amongst themselves. Consequently, deviations from the star graph, which has idealized core-periphery structure, can be probed by measuring the local clustering coefficients of the periphery nodes. One can identify these heuristically by simply considering the nodes with low degree, so the aforementioned inverse relationship hints that the core-periphery structure may exhibit considerable complexity. A second and more robust implication is that the clustering coefficient of the complete interbank network is dominated by the clustering of low-degree nodes if this inverse relationship is observed together with heavy-tailed degree distributions.

High clustering has obvious implications for systemic risk [24] and therefore clustering coefficients are of interest to the interbank network literature. Table 1 in the main text shows, that at least a few studies claim opposite 'typical' values for clustering coefficients. This may be due to the fact that the clustering tends to increase with longer time windows [1], or that the coefficients may not have been compared to the mean clustering coefficient obtained for random networks of the same size and number of edges.

The *local* clustering coefficient [43] of a node $i$ is defined in function of triangle motifs in undirected networks:

$$C_i = \frac{\text{number of of triangles connected to } i}{\text{number of triples on } i}, \tag{2.8}$$

where a triple on $i$ means an unordered pair of nodes connected via $i$, and possibly connected directly. If that is the case, the triple is counted as a triangle, ending up in the nominator of Eq. 2.8 as well. Thus $C_i$ expresses the degree of connectedness among the neighbors of $i$ [75]. The (*global*) clustering coefficient of an undirected $(N, E)$ network is then simply the average of the local coefficients:

$$C = \frac{1}{N}\sum_i C_i. \tag{2.9}$$

To assess whether an observed network possesses a non-random clustering structure and thus a significant clustering coefficient $C_0$, we generate $n$ random networks of the same size, i.e. the same number of nodes and edges, having clustering coefficients $C_{\text{rand}} = \{C_1, C_2, \ldots, C_n\}$. Then we can calculate the $z$-score of the observed $C_0$ as

$$z = \frac{C_0 - \langle C_{\text{rand}} \rangle}{\text{sd}(C_{\text{rand}})}. \tag{2.10}$$

Large $z$ indicates a significant value of $C_0$, the sign indicating more $(+)$ or less $(-)$ clustering than a random network of the same size.

We have calculated clustering coefficients and $z$-scores for the undirected view of the loan maturity layers using yearly time windows which are displayed in Fig. 2.A.9. A clear pattern can be seen: the clustering varies from significantly high to moderate (in order $G_{<1d}$, $G_{2-7d}$, and $G_{8-30d}$), via moderately low to insignificant (in order $G_{31-90d}$, $G_{91-180d}$, and $G_{0.5-1y}$), to significantly low ($G_{1-3y}$ and $G_{>3y}$). Since all layers exhibit heavy-tailed degree distributions, the sum in Eq. 2.9 is dominated by the (local) clustering coefficients of banks with small degree. We can use $C$ to proxy the clustering of the periphery banks

Within that approximation, one sees that the core-periphery structure expected in the short maturity layers deviates considerably from the idealized star network, at least when looked at with yearly resolution. Notwithstanding the strong presence of intermediation found earlier, the lower-tier banks still trade extensively with each other, indicating that contagion risk is not located entirely in the high-tier banks.

In contrast, the two longest maturity layers, where source-sink structures are expected, exhibit considerable star-like structure, low degree nodes and periphery banks being equivalent in most cases. In addition to the low clustering, we recall the modest anticorrelation between in- and outdegree; these observations lead us to believe that $G_{1-3y}$ and $G_{>3y}$ behave as a sort of *source-sink star layers*. These layers have star-like hub structure, but hardly any intermediation occurs in the hubs. They act simply as sources and sinks, generating and dissipating excess liquidity in the interbank network. Of course, one could argue that for loans with maturities of at least one year, no intermediation is possible within the scope of one year. Looking at Fig. 2.A.5, which completely aggregates time, however, we see that the banks with the largest degrees tend to be *either* sinks or sources, especially in $G_{>3y}$. Thus the tendency of the hubs to be either sink or source, but not an intermediary, holds for time scales longer than one year. The next question is whether the star centers in the source-sink star layers are connected, or rather at the center of disconnected components. Disconnected components are plentiful compared to the number of active banks per year (see Fig. 2.A.8 and Table 2.A.1), and less than half of the latter participate in the largest (weakly) connected component, as Fig. 2.A.7 shows. Furthermore, we report that in the two source-sink star layers the average shortest path length is between 2 and 3 (see Fig. 2.A.6, which shows the average path length for $G_{1-3y}$ and $G_{>3y}$ on a *monthly* basis as being typically 2), which, together with the proven existence of hubs, indicates a compact star-like structure. In a nutshell, the source-sink star layers $G_{1-3y}$ and $G_{>3y}$ are composed in general of many disconnected 'islands', of which the largest exhibit an almost perfect star-like structure around a sink or source hub.

All results are highly significant, except for $G_{91-180d}$ and $G_{0.5-1y}$: these layers do not possess any structural clustering structure on their own. As always, the clustering in the aggregated network resembles mainly the first two layers. We further note that clustering is most volatile in times of crisis, and increases during the growing phase of the network [24].

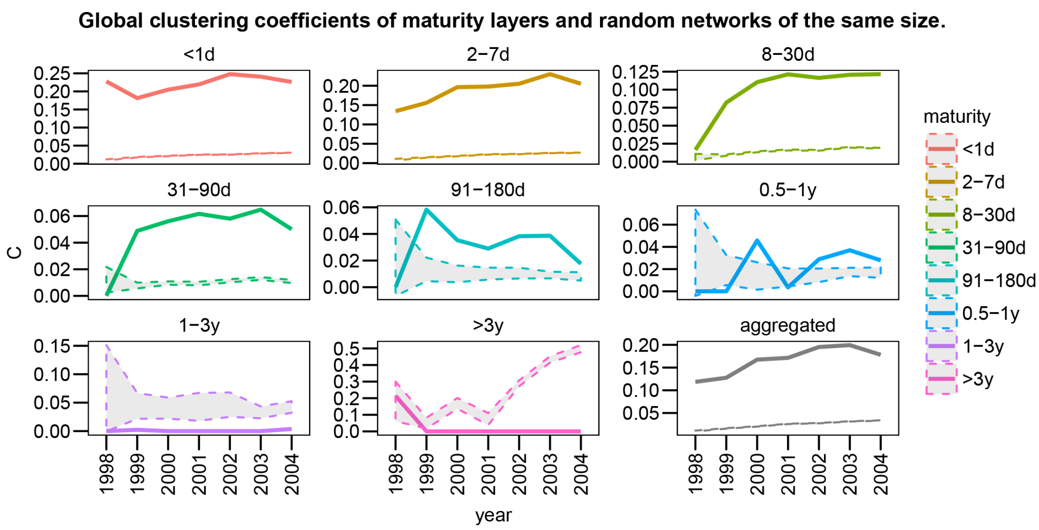



*Figure 2.A.9: Time series of clustering coefficients for the loan maturity layers and the aggregated network, together with a ribbon centered around $\langle C_{rand}\rangle$ of total width $2 \times sd(C_{rand})$. Following the methodology in [1], a total of 100 Erdős-Rényi networks were generated to test the significance of the observed $C$'s.*

### 2.A.7 Bank size mixing as degree size mixing

Many studies point to *disassortative mixing* with respect to the bank size, meaning that small banks trade mainly with large banks and vice versa. Given the core-periphery picture in these studies, one expects that the core banks are the large banks. Indeed, it is found that total bank assets are significant in explaining core membership [28]. As bank size correlates with total degree, we would also expect disassortative mixing of the banks with respect to the (total) degrees, i.e. high (low) degree nodes tend to be connected to low (high) degree nodes [43]. In fact several studies report this as an additional stylized fact [1, 25].

Figure 2.A.10 shows the assortativity with respect to the total node degrees on yearly basis, which is also called the total degree correlation. The layers with short maturity show a clear preference for the low-degree nodes to attach to the high-degree ones. This preference weakens when we look at longer maturity layers. The assortativity or the long maturity layers is pushed up by the large number of disconnected clusters, in most cases simply isolated pairs of trading banks. The disassortative degree mixing in the larger clusters is present, caused by their star-like structure. As the number of unconnected components grows drastically, we would expect that the same mechanism is behind the high assortativity for $G_{91-180d}$ and especially for $G_{0.5-1y}$. Interestingly, this expectation turns out to be false: evaluating the degree assortativity only in the largest connected component results in lower values for all layers (notably in those where star-like structures were expected based on the clustering coefficient), but it remains strongly positive for $G_{91-180d}$ and $G_{0.5-1y}$ during the stable phase of the network (Fig. 2.A.11).

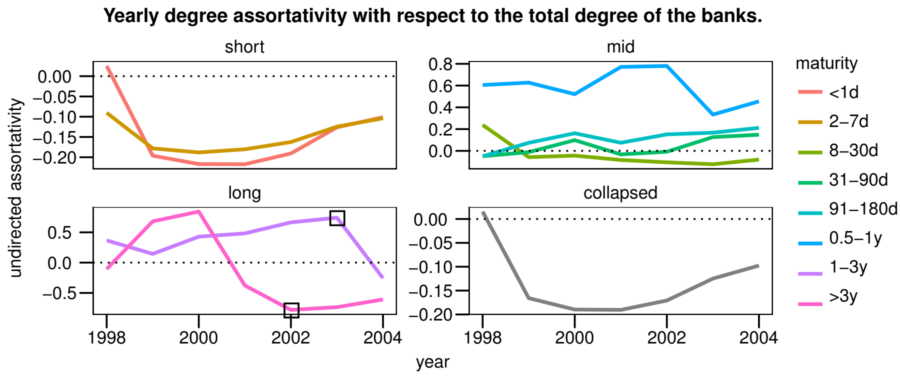



*Figure 2.A.10: The degree assortativity measures to which extent nodes with a given degree associate preferentially with other nodes of similar degree – see Equation (2) in [76] for the formal definition. In this case the type of degree considered is the total degree, i.e. the sum of the in- and outdegree. A yearly resolution was chosen because monthly aggregation suffered from bad statistics starting from* $G_{91-180d}$.

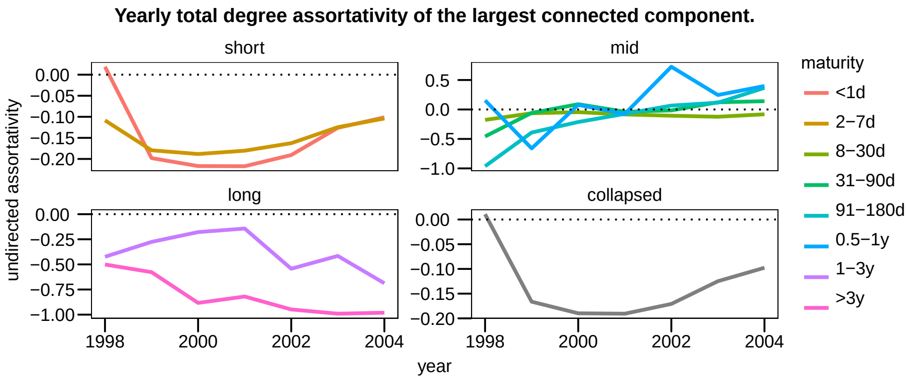



*Figure 2.A.11: The degree assortativity for the total degrees throughout the years. The degree assortativity is also called degree correlation [43].*

### 2.A.8 Summary

A summary of the results of this maturity layer analysis can be found in Table 1 in the main text. While certain features here are indicative of known structures in the interbank lending network literature, we refrain from making any such claims. Rather we report the findings here as is and only make claims about function and structure in the main text where steps have been taken to gather statistical evidence for both the construction of the layers as the occurrence of structure in such layers.

## 2.B OG inference with non-contiguous binning

Here we present the results of the OG inference where non-contiguous binning was allowed in Fig. 2.B.1 (as opposed to the contiguous binning in the main text). This allows for partitionings where short maturity loans are merged with very long maturity loans. One should keep in mind that the algorithm only makes a new bin in the OG when there is enough statistical evidence that it contains a different structure than the other bins. It is thus possible that in those cases where very long maturity loans are merged with short maturity loans, there is just hardly any actual structure (and thus not enough statistical evidence) in the long maturity layer. It then makes sense that such layer is merged with the densest layer available since this merger will cause the least amount of new connections to be introduced. This is corroborated by the fact that this kind of merging only happens in the early stages of market development and at the moment the trust crisis hit the interbank market in the summer of 2004 at the end of the data period.

Comparing Fig.3 in the main text to Fig. 2.B.1 shows the results to be qualitatively the same.

As mentioned in the main text, for comparability the partitionings are again always numbered according to the shortest maturity class in the bin.

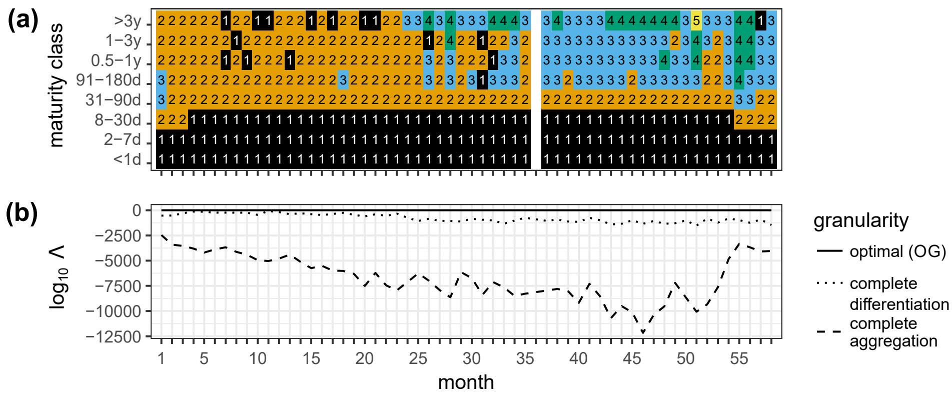



*Figure 2.B.1: Results of the layered SBM with non-contiguous binning. The optimal granularity (OG) with respect to the loan maturity classes for the Russian interbank lending network. The OG corresponds to the granularity of the best-fit layered SBM. **(a)** Monthly time series of the OG inferred from the monthly interbank lending network. Each OG bin (OGB) holds one or more maturity classes and is labelled by an OGB index* $1, 2, 3, 4$ *and indicated by a colour. The OGB index runs according to the shortest maturity class in the bin. The bin with the shortest maturity class in it is numbered OGB 1 and so forth. The OGBs correspond to lending patterns between the bank groups that differ from each other in a statistically significant way. **(b)** Temporal evolution of the* $\log_{10}$ *of the posterior odds ratio* $\Lambda$ *(for its definition see Eq. 4 in the main text) of the layered SBM for three different granularities: (i) the OG; (ii) complete loan maturity differentiation; (iii) complete loan maturity aggregation. Note that the OG inferred from the algorithm is consistently preferred. Complete aggregation is decisively rejected as an optimal representation of the mesoscale of the monthly interbank lending network*

# 2.C Number of active banks per group

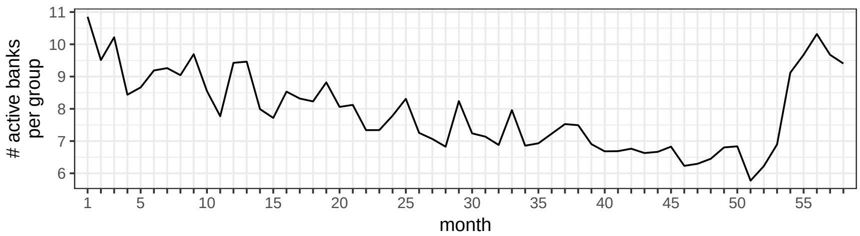



*Figure 2.C.1: The number of active banks per group is obtained by dividing the number of active banks by $B$, i.e. the number of groups in the inferred bank groups $\{b_i\}$.*

## 2.D Activity characteristics per OGB

We investigate the OGBs and their typical characteristics. Fig. 2.D.1 shows the distribution of the log group strengths across time and per OGB. We discern three types of group strength: outstrength, instrength and internal strength. The group outstrength in a certain month of bank group $b$ is the total amount lent by all banks belonging to $b$ to banks belonging to another group $b' \neq b$. In symbols,

$$s_b^{\text{out}} = \sum_l \sum_{i,j,k} x_{ijk}^l \delta_{b_i,b}(1 - \delta_{b_j,b}) \tag{2.11}$$

The two other strength types and the constraint that the maturity layer $l$ must be in a given OGB are obtained likewise from this definition.

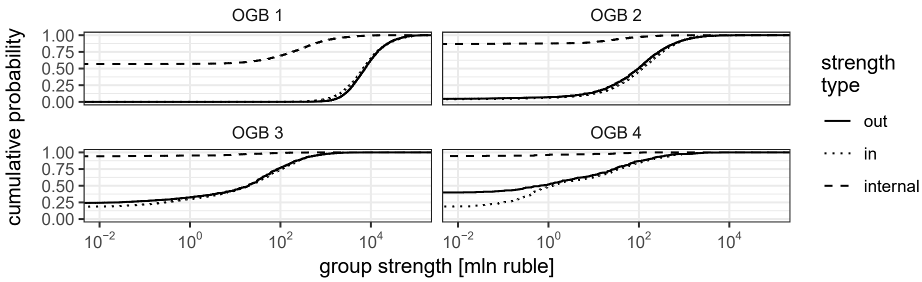



*Figure 2.D.1: The time-integrated cumulative density functions (CDFs) of the group instrengths, outstrengths and internal strengths for each OGB. Zero group strengths are represented by non-vanishing CDFs at the origin.*

We see that the groups show anti-community characteristics, they hardly lend or borrow internally. Looking at the group strengths, each OGB also seems to function on its own magnitude scale.

## 2.E Other percentiles for instrength/strength of important banks

We look at the lending behaviour of individual banks. As in the main text we again use the bank strength to single out the “important” banks. We show that the conclusions from the main text hold for different bank strength cutoffs. Each row of subplots in Fig. 2.E.1 uses a different cutoff, respectively with the bank strength (s) not limited, limited to the top $10$ percentiles, top $5$ percentiles, and the top $1$ percentile of bank strengths. At monthly time scales the important banks in “short” bins tend to both lend and borrow equal amounts of money, while the important banks in the “long” bins tend to either lend or borrow.

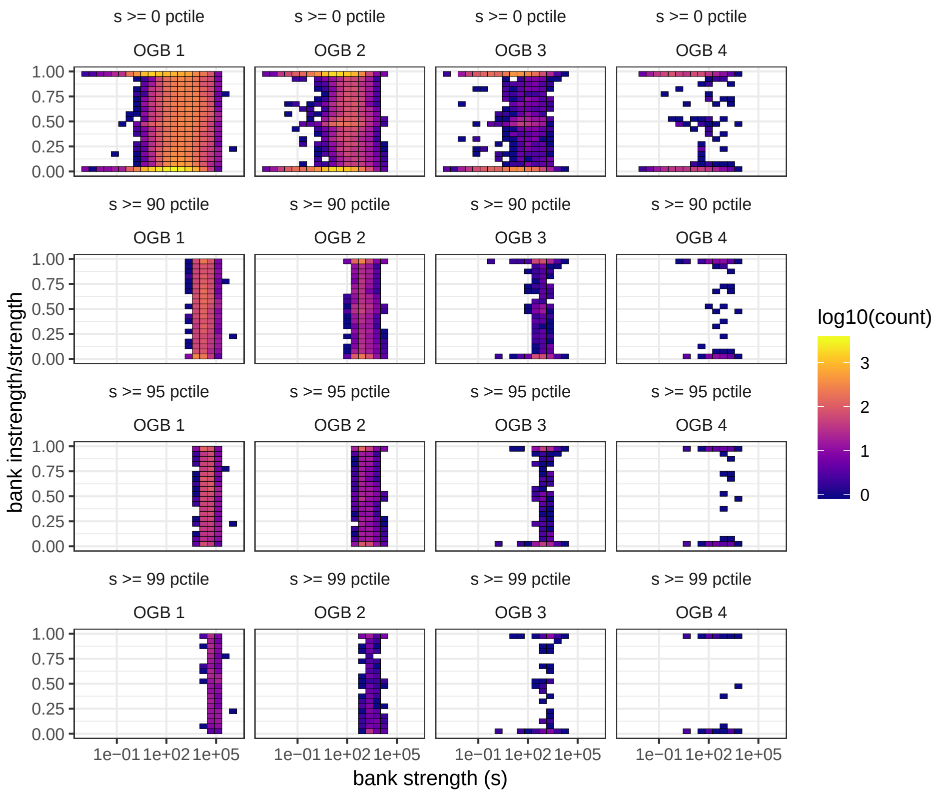



*Figure 2.E.1: The time-integrated distribution of the strength and the instrength/strength ratio of the monthly "important banks" for the four OGBs. As in Fig. 4(b) in the main text we gauge a bank's importance by its strength: For a given month a bank is deemed important if its strength lies in the top X percentiles, with X here differing and given per row of subplots. With growing OGB index, the important banks increasingly tend to either lend or borrow. Banks in OGB 1 and OGB 2 tend to balance lending and borrowing. This suggests that the economic function of the important banks in the various OGBs changes from financial intermediation (short maturity bins) to financing (long maturity bins) on a monthly time scale.*

## 2.F Timeline of the announcement crisis

The trust crises consist of the announcement crisis in the second half of 2003 and the summer of 2004 crisis. This Appendix has been compiled from [66, 71, 77, 78].

Investigations of money laundering led the CBR to deprive Sodbusinessbank of its license in May 2004. The following mutual suspicion led to a drying up of liquidity on the interbank market in the summer of 2004. Roughly one year earlier, this investigation was announced, which caused a smaller trust crisis (i.e. the announcement crisis).

| | |
|---|---|
| end of July 2003 | Start of announcement crisis. Decline in reciprocal bank interactions. |
| September 2003 | The network starts to fall apart because of the distrust among banks. |
| December 2003 - January 2004 | Back to normal liquidity. |
| March | Peak of interbank lending market recovery from early trust crisis. The bank lending reciprocity, however, does not regain pre-crisis levels. |
| following weeks | Three stages of the summer of 2004 crisis. |
| First stage: April | Volatility on the interbank market with even higher lending rates than later stages but without significant outflow of individual's deposits. Demand for liquidity was caused by policy changes and statements of the CBR; no perception of crisis by the banks themselves, yet this financial instability undoubtedly impacted the crisis to come. |
| Second stage: May | CBR deprives Sodbusinessbank of its license. |
| May 19-21 | Several conflicting statements from authorities about the deposit insurance for the clients of Sodbusinessbank make the depositors increasingly uneasy. |
| Third stage: June 3 | Crisis now definitively developed beyond Sodbusinessbank alone. Banks start to introduce additional control measures; several suspend lending activities on the interbank market. |
| June 11 | CBR changes policy rates to accommodate the banks, officially for the low liquidity on the interbank short-term market. |
| June 21-22 | Peak of the crisis. |
| July 13 | Interbank market starts to stabilize. |
| July 16 | The 'crisis of confidence' is declared to be at an end. |

Investigations of money laundering led the CBR to deprive Sodbusinessbank of its license in May 2004. The following mutual suspicion led to a drying up of liquidity on the interbank market in the summer of 2004. Roughly one year earlier, this investigation was announced, which caused a smaller trust crisis (i.e. the announcement crisis). In May 2004, a crisis on the Russian interbank market was triggered. In particular, on May 13 the Central Bank of Russia recalled the license from Sodbiznesbank on accusations of money laundering and sponsorship of terrorism. It was the first bank to be closed on these grounds in Russia. This unexpected closure caused panic on the interbank market since banks suspected other banks would follow suit, but they had no reliable information on who these banks might be.

On May 16 the head of Federal Financial Monitoring Service, Viktor Zubkov, announced publicly that his agency suspected another ten banks of similar violations. Because this announcement was not accompanied by the actual list of suspected banks, it just fuelled the hysteria on the interbank market. Rumours about the identities of these ten banks started to spread rapidly. Soon several inconsistent blacklists were circulating in the banking community as bankers tried to guess which banks were officially suspected of money laundering. Anecdotal evidence suggests that banks were actively helping to spread the rumour by removing themselves from the list and adding competitors in an attempt to escape the carnage. By consequence, The union of all the blacklists circulating in the banking community expanded in a few days to include dozens of banks, including several market leaders.

In the presence of total uncertainty about the quality of their counterparties, banks began to reduce limits on each other, which reverberated into an acute liquidity drought on the interbank market [79]. The turnover volume on the interbank market dropped spectacularly. Later the original source of the panic, Viktor Zubkov, shockingly announced that “the Federal Financial Monitoring Service has no blacklist”. The deputy Minister of Internal Affairs similarly announced that “the Interior Ministry has no such list. We have no plans to persecute any banks.”For further descriptions (in Russian) of this episode in Russian history see [80, 81]. Up till now it remains unclear whether an official list actually existed. Important for us is that the 2004 meltdown on the Russian interbank market was based on rumours and unrelated to shocks to the fundamentals of Russian banks or the Russian economy at large.

This non-fundamental and exogenous mutual trust crisis provides a great example of what structural changes occur during a shock to the non-fundamentals of the ILN. The advantage of the trust crisis compared to the 2007-2009 financial crisis is that, while the financial crisis was caused elsewhere and only indirectly found its way into Russia (via oil-prices, etc.), the trust crisis had its epicenter inside of Russia.

## 2.G Maturity classes in the ILN literature

An overview of the ILN literature using multilayer approaches and how their granularity compares to the current dataset.

| | <1d | 2-7d | 8-30d | 31-90d | 91-180d | 0.5-1y | 1-3y | >3y |
|---|---|---|---|---|---|---|---|---|
| Central Bank of Russia [82] | instant | short term | | long term | | | | |
| Bargigli et al. (2015) [1] | overnight | short term | | | | | long term | |
| Aldasoro and Alves (2016) [25] | short term | | | | | | long term | |
| Montagna and Kok (2016) [16] | short term | | | | long term | | | |

*Table 2.G.1: Classifications of loan maturity in terms of our maturity classes used by the Central Bank of Russia and several studies that have included multilayered ILNs.*

## 2.H Term structure of interest rates and loan volumes in the interbank loan market

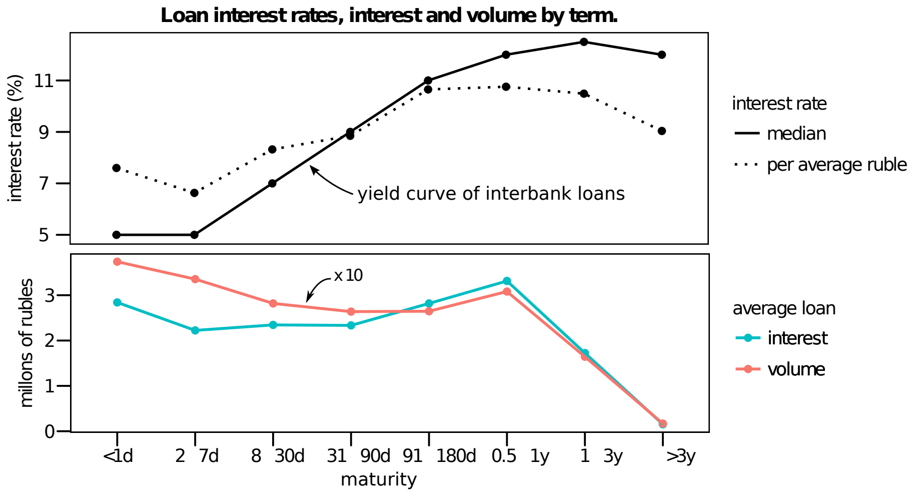



*Figure 2.H.1: (top panel) The global categorical yield curve is the median interest rate over time per term. An alternative characterization is the interest rate per average ruble lent/borrowed, defined as the average interest weighed by the loan volumes. This statistic is robust to whether one excludes or includes the interest rate outliers. (lower panel) The average loan interest received/paid per term in millions of rubles and average loan volume (size) lent/borrowed in tens of millions of rubles.*

### The global categorical yield curve

The yield curve is obtained after "averaging" interest rates $r$ over time and by maturity class. The averaging technique chosen is the median due to extreme outliers which disproportionately affect the mean: the standard deviation of the complete $r$ population (including outliers) is $\sigma \approx 40\%$, yet $99\%$ of the population is contained in $[0, \sigma]$. Within that interval the standard deviation is about 7%, which is a more sensible measure of the interest rate dispersion. By using the median, we avoid choosing any cutoff.

Yield curves are usually considered with *continuous* maturities ranging from one day to several years, but the available data only records maturity *classes*. Although methods exist to estimate the continuous yield curve from discrete data [83], we will only consider interest rates (and interests etc.) per category, i.e. per maturity class.

The solid line in the top panel of Figure 2.H.1 exhibits a typical stylized fact of the yield curve. It is upward sloping and has a convex shape, except for the $>$3y maturity, i.e. maturities three years or longer. The unusual steepness of the curve in Figure 2.H.1 is an artifact caused by the term categories; as they progress, they bucket a growing amount of maturities, so a continuous yield curve would be horizontally stretched with respect to the categorical yield curve. The upward slope is usually explained by classical *expectations theory* (ET). According to ET, interbank lending rates dynamics are determined by the structure of liquidity supply and demand [84]; the long-term interest rate is an average of expected future short-term rates, plus a term premium that increases with longer terms to compensate risk-averse lenders for the interest risk, which arises for lenders from fluctuating interest rates with respect to the base deposit policy rate. One can also extend ET by including separate premiums for liquidity risk (selling loans on the secondary market tends to be harder as their maturities lengthen) and default risk, also called counterparty risk, which is in theory governed by the credit rating of the borrower. Normally the short-end (long-end) of the yield curve is dominated by the liquidity (default) risk; both are considered components of interbank lending risk [55, 85]. Each risk contributes to the upward slope of the yield curve given normal market conditions.

The interest per average ruble is also displayed in the top panel of Figure 2.H.1. Loans with longer maturities are the most profitable for lenders. The drop for the >3y maturity could point to the fact that these contracts may be made on more amicable (flexible) terms; this would imply lower perceived interbank risk which could explain the negative slope between the 1-3y and >3y categories.

To illustrate the dynamics that underlie the global yield curve, Figure 2.H.2(a) plots the categorical yield curve for each of the last 16 months in the data. In this period the by now mature interbank network deals with the trust crises which are explained in more detail in Appendix 2.F. Figure 2.H.2(b) contains the *(yield) spread*, which we define as the difference in average yield between "long" maturities and "short" maturities. We define these maturities as two classes: the "long maturity" class consists of maturity classes (1-3y, >3y) and the "short maturity" class consists of maturity classes (<1d, 2-7d, 8-30d). The general shape of the curve is not sensitive to the definition of the "long maturity" class; the "short maturity" class is motivated in the paper, while the "long maturity class" is based on Basel recommendations and is also used in [1, 25].

In normal times, interbank markets are among the most liquid in the financial sector: banks prefer to lend out excess cash since the central bank's interest rate on excess reserves is smaller than rates available in interbank markets [86]. During trust crises, the perceived default risk grows, which inflates interest rates according to ET. Riskier banks, i.e. banks at risk of being in financial distress, exert an externality on safer banks who subsidize their liquidity [86]. If the crisis gets worse, this externality on safer banks is so costly that they leave the unsecured market, and liquidity rich banks may prefer to hoard liquidity instead of lending it

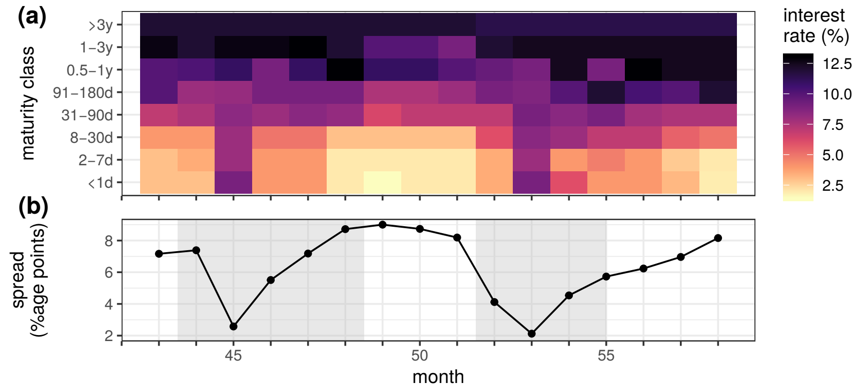



*Figure 2.H.2: Monthly categorical yield curves (i.e. aggregated by maturity class; top panel) together with the spread (bottom panel) from July 2003 (month up until November 2004. This period is embedded in the emergent maturity phase of the network and includes the announcement and summer of 2004 crises (these are the trust crises – see Section 2.F), both shaded in grey on the bottom panel. The first shaded region spans the period from August 1st 2003 until January 1st 2004. The second shaded region spans the April 1st 2004 until July 16th 2004. The CBR lowered the overnight rate from 16 to 14% on January 15, 2004 and again to 13% on June 15, 2004. The yield curve is the median interest rate per maturity class, as in Figures 2.H.1; the spread is defined in the text.*

out to an adverse selection of borrowers; the interbank lending market dries up.

We see that this mechanism is indeed captured by the spread curve in the bottom panel of Figure 2.H.2: low spread seems to indicate abnormal market conditions with low liquidity because the short term interest rates increase quite relatively fast. Note in the upper panel that the long term interest rates stay almost constant during the crises and drop slightly during the recovery in between (roughly from January until April 2004). This suggests that we cannot make an analogy with the typical *inverted yield curves* of e.g. treasury securities that are associated with (predicting) recessions. Indeed, according to ET these are yield curves with negative slopes because investors have poor expectations of future interest rates. In contrast, we see for our data that the short term interest rates rise quickly during the trust crisis while the long term rates hardly change.

## Characteristics of loan volumes and interests

The lower panel of Figure 2.H.1 shows the average interest and volume per term. The loan volumes are log-normally distributed [24] and this holds remarkably well for the interest too, especially for shorter terms. Because of the considerable variance on a linear scale, these averages may be interpreted only as rough order of magnitude estimates. With this in mind, Figure 2.H.1 shows that interest rate and volume are roughly negatively correlated: except for the bump at 0.5-1y and the case >3y, the volumes decrease as the interest rates increase. This can also be seen in the slowly varying average interest for the first five term categories.

Table 2.A.3 lists the total loan volumes by term and per year. As stated there (p. 49), we observe that the relative importance of each term segment, as measured by the total volume of loans traded within it, follows the ranking of the loan terms. This, together with the typical volumes and interest rates in Figure 2.H.1 and the lending activity in Table 2.A.2 (p. 49), supports the conclusion that the Russian interbank network exhibits a distinct hierarchy with respect to the loan maturities, which we could summarize by saying that *banks lend greater volumes at lower interest rates more often for shorter loan terms.*

# References

[1] L. Bargigli, G. di Iasio, L. Infante, F. Lillo, and F. Pierobon. *The multiplex structure of interbank networks*. Quantitative Finance, 15(4):673–691, April 2015.

[2] Marcel Bluhm, Co-Pierre Georg, and Jan-Pieter Krahnen. *Interbank intermediation*. Bundesbank Discussion Paper 16, Deutsche Bundesbank, 2016.

[3] Jrgen Wiemers and Ulrike Neyer. *Why do we have an interbank money market?* IWH Discussion Papers, 182, 2003.

[4] Prasanna Gai, Andrew Haldane, and Sujit Kapadia. *Complexity, concentration and contagion*. Journal of Monetary Economics, 58(5):453 – 470, 2011.

[5] Fabio Caccioli, Paolo Barucca, and Teruyoshi Kobayashi. *Network models of financial systemic risk: a review*. Journal of Computational Social Science, 1(1):81–114, Jan 2018.

[6] Anne-Caroline Hser. *Too interconnected to fail: A survey of the interbank networks literature*. SAFE Working Paper Series 91, Research Center SAFE - Sustainable Architecture for Finance in Europe, Goethe University Frankfurt, 2015.

[7] Jorge A Chan-Lau, Marco Espinosa, Kay Giesecke, and Juan A Solé. *Assessing the systemic implications of financial linkages*. IMF Global Financial Stability Report, 2:73–110, April 2009.

[8] ECB. *Recent advances in modelling systemic risk using network analysis*. Technical report, ECB, 1 2010.

[9] Stefano Battiston, Guido Caldarelli, Robert M. May, Tarik Roukny, and Joseph E. Stiglitz. *The price of complexity in financial networks*. Proceedings of the National Academy of Sciences, 113(36):10031–10036, 2016.

[10] Tarik Roukny, Stefano Battiston, and Joseph E. Stiglitz. *Interconnectedness as a source of uncertainty in systemic risk*. Journal of Financial Stability, 35:93–106, April 2018.

[11] Jean-Charles Rochet and Jean Tirole. *Interbank Lending and Systemic Risk*. Journal of Money, Credit and Banking, 28(4):733–762, 1996.

[12] Craig H. Furfine. *Banks as Monitors of Other Banks: Evidence from the Overnight Federal Funds Market*. The Journal of Business, 74(1):33–57, 2001.

[13] Kartik Anand, Iman van Lelyveld, dm Banai, Soeren Friedrich, Rodney Garratt, Grzegorz Haaj, Jose Fique, Ib Hansen, Serafn Martnez Jaramillo, Hwayun Lee, Jos Luis Molina-Borboa, Stefano Nobili, Sriram Rajan, Dilyara Salakhova, Thiago Christiano Silva, Laura Silvestri, and Sergio Rubens Stancato de Souza. *The missing links: A global study on uncovering financial network structures from partial data*. Journal of Financial Stability, 35:107–119, April 2018.

[14] Grzegorz Halaj and Christoffer Kok. *Modelling the emergence of the interbank networks*. Quantitative Finance, 15(4):653–671, 2015.

[15] Tao Xu, Jianmin He, and Shouwei Li. *A dynamic network model for interbank market*. Physica A: Statistical Mechanics and its Applications, 463:131 – 138, 2016.

[16] Mattia Montagna and Christoffer Kok. *Multi-layered interbank model for assessing systemic risk*. Kiel Working Papers 1873, Kiel Institute for the World Economy (IfW), 2013.

[17] Alexei Karas and Koen J. L. Schoors. *A Guide to Russian Banks Data*. SSRN Electronic Journal, 2010.

[18] Tarik Roukny, Hugues Bersini, Hugues Pirotte, Guido Caldarelli, and Stefano Battiston. *Default Cascades in Complex Networks: Topology and Systemic Risk*. Scientific Reports, 3:2759, September 2013.

[19] Tiago P. Peixoto. *Inferring the mesoscale structure of layered, edge-valued, and time-varying networks*. Phys. Rev. E, 92:042807, Oct 2015.

[20] Tiago P. Peixoto. *Nonparametric weighted stochastic block models*. Physical Review E, 97(1):1–20, 2018.

[21] José-Luis Molina-Borboa, Serafin Martinez-Jaramillo, Fabrizio López-Gallo, and Marco van der Leij. *A multiplex network analysis of the Mexican banking system: link persistence, overlap and waiting times*. The Journal of Network Theory in Finance, 1(1):99–138, mar 2015.

[22] Sam Langfield, Zijun Liu, and Tomohiro Ota. *Mapping the UK interbank system*. Journal of Banking & Finance, 45:288–303, 2014.

[23] Karl Finger, Daniel Fricke, and Thomas Lux. *Network analysis of the e-MID overnight money market: the informational value of different aggregation levels for intrinsic dynamic processes*. Computational Management Science, 10(2):187–211, June 2013.

[24] Benjamin Vandermarliere, Alexei Karas, Jan Ryckebusch, and Koen Schoors. *Beyond the power law: Uncovering stylized facts in interbank networks*. Physica A: Statistical Mechanics and its Applications, 428:443–457, June 2015.

[25] Iaki Aldasoro and Ivn Alves. *Multiplex interbank networks and systemic importance: An application to European data*. Journal of Financial Stability, 35(C):17–37, 2018.

[26] Francisco Blasques, Falk Bruning, and Iman van Lelyveld. *A dynamic network model of the unsecured interbank lending market*. Journal of Economic Dynamics and Control, 90:310–342, May 2018.

[27] Thiago Christiano Silva, Marcos Soares da Silva, and Benjamin Miranda Tabak. *Liquidity Performance Evaluation of the Brazilian Interbank Market using a Network-Based Approach*. Working Papers Series 401, Central Bank of Brazil, Research Department, September 2015.

[28] Ben Craig and Goetz von Peter. *Interbank tiering and money center banks*. Journal of Financial Intermediation, 23(3):322–347, July 2014.

[29] G. De Masi, G. Iori, and G. Caldarelli. *Fitness model for the Italian interbank money market*. Phys. Rev. E, 74:066112, Dec 2006.

[30] Kimmo Soramäki, Morten L. Bech, Jeffrey Arnold, Robert J. Glass, and Walter E. Beyeler. *The topology of interbank payment flows*. Physica A: Statistical Mechanics and its Applications, 379(1):317–333, June 2007.

[31] Stephen P Borgatti and Martin G Everett. *Models of core/periphery structures*. Social Networks, 21(4):375 – 395, 2000.

[32] Fabio Della Rossa, Fabio Dercole, and Carlo Piccardi. *Profiling core-periphery network structure by random walkers*. Scientific Reports, 3:1467, March 2013.

[33] Sean ZW Lip. *A fast algorithm for the discrete core/periphery bipartitioning Problem*. arXiv:1102.5511, 2011.

[34] Sadamori Kojaku and Naoki Masuda. *Core-periphery structure requires something else in the network*. New Journal of Physics, 20(4):043012, 2018.

[35] Xiao Zhang, Travis Martin, and Mark EJ Newman. *Identification of core-periphery structure in networks*. Physical Review E, 91(3):032803, 2015.

[36] Paolo Barucca and Fabrizio Lillo. *Disentangling bipartite and core-periphery structure in financial networks*. Chaos, Solitons & Fractals, 88:244–253, 2016.

[37] Paolo Barucca and Fabrizio Lillo. *The organization of the interbank network and how ECB unconventional measures affected the e-MID overnight market*. Computational Management Science, 15(1):33–53, Jan 2018.

[38] P.R. Krugman. *The Self Organizing Economy*. Wiley-Blackwell Publishers, 1996.

[39] Giulia Iori, Rosario N. Mantegna, Luca Marotta, Salvatore Miccich, James Porter, and Michele Tumminello. *Networked relationships in the e-MID interbank market: A trading model with memory*. Journal of Economic Dynamics and Control, 50:98 – 116, 2015. Crises and Complexity.

[40] Daan Laurens Veld. *Complex systems in financial economics: Applications to interbank and stock markets*. phdthesis, Amsterdam School of Economics Research Institute, UvA, 2014.

[41] Ben R. Craig, Falko Fecht, and Gnseli Tmer-Alkan. *The role of interbank relationships and liquidity needs*. Journal of Banking & Finance, 53:99–111, April 2015.

[42] Falk Bruning and Falko Fecht. *Relationship Lending in the Interbank Market and the Price of Liquidity***. Review of Finance, 21(1):33–75, March 2017.

[43] M. E. J. Newman. *The Structure and Function of Complex Networks*. SIAM Review, 45:167–256, January 2003.

[44] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. *Stochastic blockmodels: First steps*. Social Networks, 5(2):109–137, June 1983.

[45] Koen Schoors and Ksenia Yudaeva. *Russian Banking as an Active Volcano*. In Michael Alexeev and Shlomo Weber, editors, The Oxford Handbook of the Russian economy, chapter 4, pages 544–573. Oxford University Press, Oxford, 2013.

[46] Andrei Vernikov. *A Guide to Russian Bank Data: Breaking Down the Sample of Banks*. SSRN, 2018.

[47] Tiago P. Peixoto. *Bayesian stochastic blockmodeling*. In Patrick Doreian, Vladimir Batagelj, and Anuska Ferligoj, editors, Advances in Network Clustering and Blockmodeling. Wiley, New York, 2018.

[48] Emmanuel Abbe. *Community Detection and Stochastic Block Models: Recent Developments*. Journal of Machine Learning Research, 18(177):1–86, 2018.

[49] Jean-Gabriel Young, Guillaume St-Onge, Patrick Desrosiers, and Louis J. Dubé. *Universality of the stochastic block model*. Phys. Rev. E, 98:032309, Sep 2018.

[50] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2005.

[51] Tiago P Peixoto. *Hierarchical block structures and high-resolution model selection in large networks*. Physical Review X, 4(1):011047, 2014.

[52] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, April 2003.

[53] Tiago P Peixoto. *Nonparametric Bayesian inference of the microcanonical stochastic block model*. Physical Review E, 95(1):012317, 2017.

[54] Brian Karrer and M. E. J. Newman. *Stochastic blockmodels and community structure in networks*. Phys. Rev. E, 83:016107, Jan 2011.

[55] Steven H Russell. *Understanding the term structure of interest rates: The expectations theory*. Review, 74, 1992.

[56] Sheri Markose, Simone Giansante, Mateusz Gatkowski, and Ali Rais Shaghaghi. *Too Interconnected To Fail: Financial Contagion and Systemic Risk In Network Model of CDS and Other Credit Enhancement Obligations of US Banks*. Working Papers 033, COMISEF, 2010.

[57] Alan D. Morrison, Michalis Vasios, Mungo Ivor Wilson, and Filip Zikes. *Identifying Contagion in a Banking Network*. SSRN Electronic Journal, 2016.

[58] Gideon Schwarz et al. *Estimating the dimension of a model*. The annals of statistics, 6(2):461–464, 1978.

[59] Hirotugu Akaike. *A new look at the statistical model identification*. In Selected Papers of Hirotugu Akaike, pages 215–222. Springer, 1974.

[60] Xiaoran Yan, Cosma Shalizi, Jacob E Jensen, Florent Krzakala, Cristopher Moore, Lenka Zdeborová, Pan Zhang, and Yaojia Zhu. *Model selection for degree-corrected block models*. Journal of Statistical Mechanics: Theory and Experiment, 2014(5):P05007, 2014.

[61] Tiago P. Peixoto. *The graph-tool python library*. figshare, 2014.

[62] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C (2nd Ed.): The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 1992.

[63] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. *The architecture of complex weighted networks*. Proceedings of the national academy of sciences, 101(11):3747–3752, 2004.

[64] Tiago P Peixoto and Martin Rosvall. *Modelling sequences and temporal networks with dynamic community structures*. Nature communications, 8(1):582, 2017.

[65] Santo Fortunato. *Community detection in graphs*. Physics Reports, 486(3):75 – 174, 2010.

[66] Abbigail J. Chiodo and Michael T. Owyang. *A Case Study of a Currency Crisis: The Russian Default of 1998*. In The Federal Reserve Bank of St. Louis Review, November/December, Nov/Dec 2002.

[67] Tiago P Peixoto. *Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models*. Physical Review E, 89(1):012804, 2014.

[68] A. Karas and K. Schoors. *Heracles or Sisyphus? Finding, cleaning and reconstructing a database of Russian banks*. Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium 05/327, Ghent University, Faculty of Economics and Business Administration, August 2005.

[69] F. Battiston, V. Nicosia, and V. Latora. *The new challenges of multiplex networks: measures and models*. ArXiv e-prints, June 2016.

[70] V. Zlatic, G. Bianconi, A. Díaz-Guilera, D. Garlaschelli, F. Rao, and G. Caldarelli. *On the rich-club effect in dense and weighted networks*. The European Physical Journal B, 67(3):271–275, Feb 2009.

[71] Milan van den Heuvel. *Journey through the controllability of interbank credit networks*. mathesis, Ghent University, 2015.

[72] Kim Christensen and Nicholas R Moloney. *Complexity and criticality*, volume 1. Imperial College Press, 2005.

[73] M. E. J. Newman. *Power laws, Pareto distributions and Zipf's law*. Contemporary Physics, 46(5):323–351, September 2005.

[74] D. J. Watts and S. H. Strogatz. *Collective dynamics of small-world networks*. Nature, 393(6684):440–442, June 1998.

[75] Benjamin Vandermarliere. *Network analysis of the Russian interbank system*. mathesis, Ghent University, 2012.

[76] M. E. J. Newman. *Mixing patterns in networks*. Physical Review E, 67(2):026126+, February 2003.

[77] Alexei Karas, Koen Schoors, et al. *Bank networks, interbank liquidity runs and the identification of banks that are too interconnected to fail*. Second CInST Banking Workshop, Moscow, 2012.

[78] Boris Porfiriev. *Crises in Russia: Contemporary Management Policy and Practice From A Historical Perspective*. Routledge, New York, NY, USA, 2012.

[79] Dimitry Paranyushkin. *Russian Money Market & Bank System: Problems*, 2009.

[80] Nikola Sergeevich Simonov. *Banks and Money, 1988-2008*. RAGS, 2009.

[81] Alexey Boyarskiy. *Modern Russian Banking System. Brief Chronology. Period 1988-2004.*, 2018.

[82] Valentina Y. Guleva, Maria V. Skvorcova, and Alexander V. Boukhanovsky. *Using Multiplex Networks for Banking Systems Dynamics Modelling*. Procedia Computer Science, 66:257–266, January 2015.

[83] Stanislav Anatolyev and Sergey Korepanov. *The term structure of Russian interest rates*. Applied Economics Letters, 10(13):867–870, 2003.

[84] Alexey Egorov and Olga Kovalenko. *Structural features and interest-rate dynamics of Russia's interbank lending market*. BOFIT Discussion Papers 23/2013, Bank of Finland, Institute for Economies in Transition, August 2013.

[85] Benjamin Tabak and Sandro Andrade. *Testing the Expectations Hypothesis in the Brazilian Term Structure of Interest Rates*. Brazilian Review of Finance, 1(1):19–43, 2003.

[86] Florian Heider, Marie Hoerova, and Cornelia Holthausen. *Liquidity hoarding and interbank market rates: The role of counterparty risk*. Journal of Financial Economics, 118(2):336 – 354, 2015.

# Part II

# Belgian client data of a large European bank

# 3

# The wealth origins of income mobility

Co-authored with: Tarik Roukny, Benjamin Vandermarliere, Koen Schoors, and Jan Ryckebusch.

## 3.1 Schematic overview

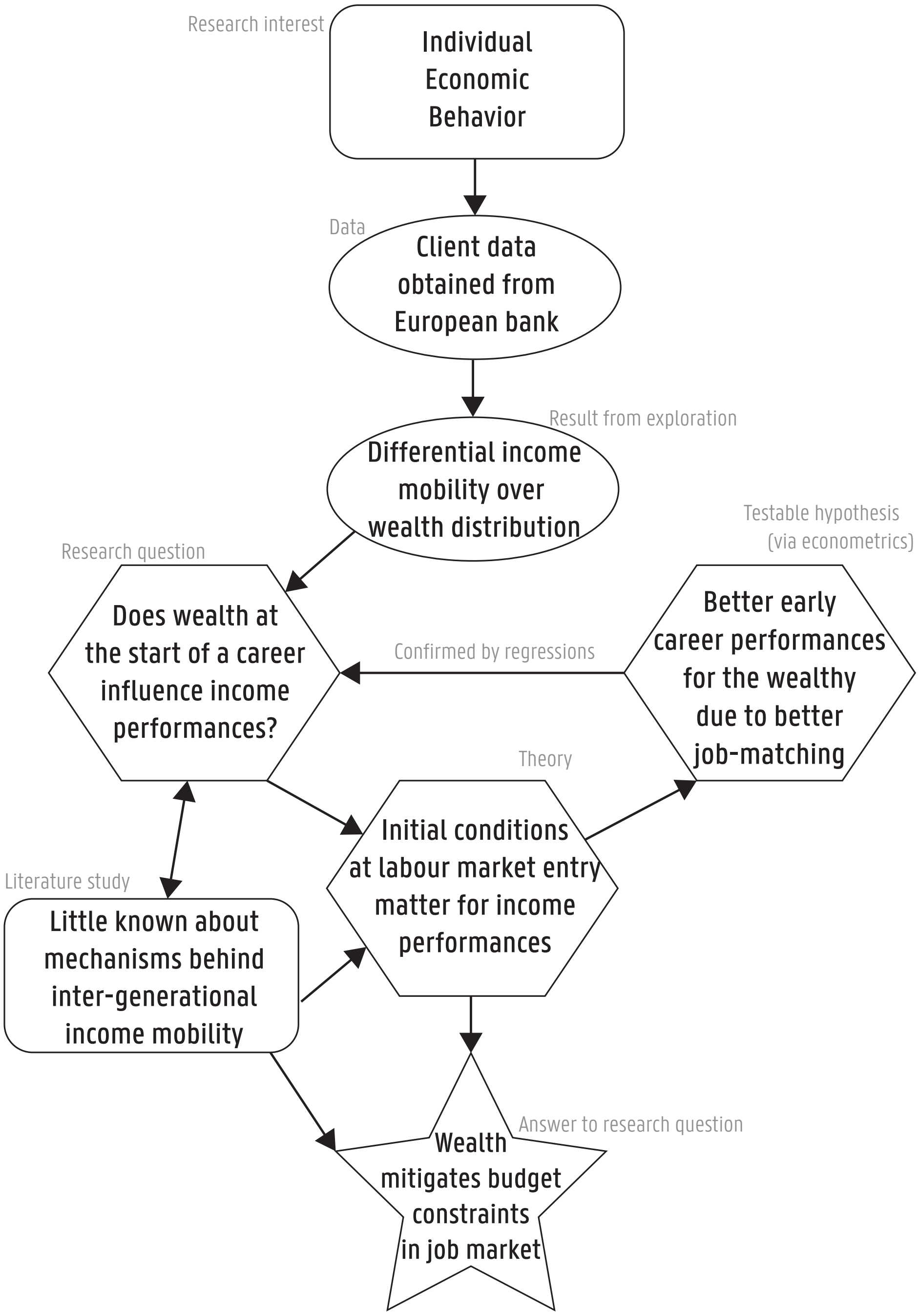



*Figure 3.1: Schematic overview of the workflow of the project on the wealth origins of income mobility.*

## 3.2 Abstract

The social costs of wealth inequality and lack of income mobility have become a central focus of both public and academic debate. In this paper, we study the interaction between financial wealth and income growth using client-level data on career starters in Belgium, obtained from a large European bank. We find higher income performances for individuals with higher financial wealth at the start of their careers. This effect holds as early as three years into a career. While the roles of social capital and innate abilities appear limited, evidence is found for a transmission channel via job search conditions. Our results suggest that individuals with higher disposable wealth are more likely to find a job matching their human capital, in turn, boosting their chances of higher performance. Policies addressing an individual's capacity to accommodate frictions in the market for first jobs could therefore substantially promote economic mobility.

## 3.3 Introduction

The stagnation of income mobility combined with the increase of wealth inequality in recent times has become a central focus in both public and academic debate [1–3]. Concerns over the related social costs, ranging from reduced health [4, 5] to lower economic growth [6–8], have triggered calls for dedicated policy action [1, 9]. However, designing adequate policies requires a careful understanding of the micro-mechanisms at play.

In modern economies, the start of a professional career stands as a major life event and offers unique opportunities to study how and when wealth inequality interacts with income mobility. Would an individual with more financial support achieve a greater income performance during the first years of her professional career? Would more financial support instead disengage an individual from income performance at early stages? Through which transmission channel and how early does the relationship materialize? We address these questions by studying the impact of financial wealth accumulated before the start of an individual's professional career on her income performances over a window of up to seven years after job market entry.

In general, countries with more inequality are also associated with less earning mobility across generations, as exhibited empirically by the so-called "Great Gatsby Curve" [10, 11]. Over the last decade, the increase of granular data has boosted research on the micro-mechanisms at play in *inter*-generational transfers and income mobility [12, 13]. Large drivers of *inter*-generational income mobility include parental income [14–17], parental health [18], geographical location in childhood [16, 19], access to and quality of education [20, 21] – which in turn also relates to parental income [16], and wealth transfers [22–26].

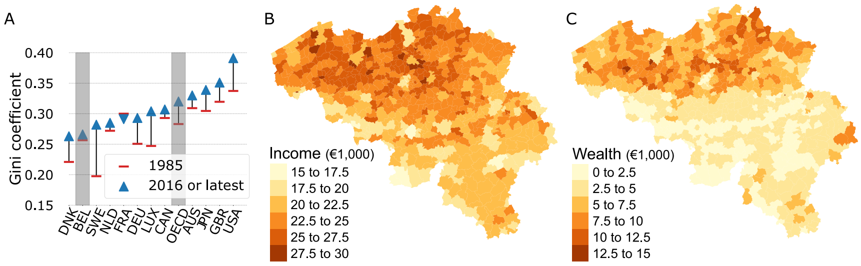



*Figure 3.1: Economic equality in Belgium.* ***A:*** *The Gini coefficient of disposable income (post taxes and transfers) for* 12 *OECD countries for the mid-*1980*s and for* 2016 *or the latest available year. Belgian and the OECD average incomes are highlighted. Note that a Gini of 0(1) stands for complete (in)equality.* ***B:*** *The observed median income per municipality in the data for all clients (aged* 18–60*) in* 2006*.* ***C:*** *The observed median financial wealth per municipality in the data for all clients (aged* 18–60*) in* 2006*.*

At the *intra*-generational level, upward income mobility is usually promoted by professional performances and career opportunities [27]. In essence, an individual's level of income is driven by a combination of human capital, financial wealth, and learning capabilities [28]. The contribution of each of these dimensions results from both innate abilities and socioeconomic variables (e.g., family, neighborhood, school, etc.) [12]. In particular, initial conditions at the moment of entry in the labor market can substantially impact lifetime earnings, wealth, and utility [28]. Such conditions include budget constraints (e.g., rate of depletion of financial wealth) and optimal allocation of human capital in the form of search and matching dynamics in the job market [29, 30]. However, little is known about the short-run materialization of the transmission channel between wealth inequality and income mobility. This paper seeks to fill this gap.

We study the effects of three transmission channels between financial wealth inequality and early career income performance: social capital, innate ability and job search conditions. We combine data from all financial transactions, financial wealth and demographic characteristics of several million Belgian clients of a large European bank between 2006 and 2016. As a result, we identify career starters as individuals exhibiting a transition to a wage income during the time window of the dataset. We then create an annual account of the financial progress of each career starter from a year before their start to seven years into their careers. In our setting, social capital is assessed via the demographic and geographic characteristics of an individual; innate abilities are reflected by an individual's own capacity to generate financial wealth before the start of her career; and job search conditions are captured by the economic pressures to secure a job.

First, we find consistently and significantly higher early career income performances for individuals with higher financial wealth at the start of their careers. For example, the €$5,500$ difference in initial financial wealth between the $25$th and the $75$th percentiles produces a $4\%$ difference in income growth three years into a career. This effect holds after controlling for demographic and geographical variables such as age, gender, and postal and neighborhood codes. Next, we address innate abilities by separating financial wealth into two sources: self-generated financial wealth (e.g., accumulated through student jobs) and transferred financial wealth (e.g., parental wealth transfers). We construct instruments to extract transferred financial wealth using the transactions data We construct instruments to isolate transferred financial wealth using the transactions data. Therefore, by focusing on this source of financial wealth, we exclude sources of wealth related to innate abilities. The results show that transferred financial wealth contributes to early career income performance with economic significance comparable to the original effects of financial wealth. Henceforth, the contribution of financial wealth originating from external transfers to early career income performance applies beyond the effects of social and innate advantages.

Finally, we address the interaction between job search conditions and human capital allocation in the market for first jobs. While timing and initial conditions are critical in job markets [28, 31], first-time jobseekers are particularly constrained by the length of the period they can sustain without income. We hypothesize that an individual entering the job market with more financial wealth possesses a higher capacity to sustain an extended absence of income, irrespective of the origin of her financial wealth. In turn, such an individual would be more likely to find a job matching her human capital, thus increasing her chances of stronger early career performances through higher productivity [32]. In order to test the job search condition hypothesis, we construct a measure of "thriftiness," that is, an individual's share of financial wealth that is not consumed before the start of a career. Intuitively, thriftiness generates higher savings and should therefore compensate for lower financial wealth transfers. We find that higher thriftiness is indeed consistently associated with higher early career income performances. While thrift and financial wealth are found to have comparable effects on income growth, they exhibit substituting effects on each other: benefits from financial wealth transfers before the start of a career are not boosted by higher individual thrift or vice versa.

In the absence of any compensating behavior such as thriftiness, less financially wealthy individuals face larger budget constraints when entering the job market for the first time. In turn, they exhibit lower income performances early in their career. Stronger pressure when entering the job market for the first-time may push jobseekers to accept job offers promptly at the cost of a low match thus reducing their expected income performances. In particular, this result offers evidence of adverse selection in the inequality distribution for early contracting [31].

Overall, our findings suggest that differential frictions in the hlsearch for first jobs due to financial pressure may contribute to the perpetuation of economic inequality at least on par with other more classic factors such as social capital and innate abilities. Policies addressing this precise issue may be key to promoting income mobility.

While the data used in this paper covers a relatively small country with 11 million inhabitants, Belgium has consistently ranked as a top-tier country in terms of economic equality and mobility [8, 33]. Figure 3.1.A shows that Belgium's Gini index of income inequality is listed as one of the lowest among the OECD countries. Additionally, Belgians do not rely on debt to finance their graduate studies, in contrast with other countries like the US and the UK [34]. Most likely, student debt adds pressure on the budget constraints of a first-time jobseeker. All together, these considerations indicate that the results from this paper provide a lower bound for the effects of wealth transfers and savings behavior on income mobility and early career income performances. For illustrative purposes, Figure 3.1.B and Figure 3.1.C also show median income and financial wealth per municipality, respectively, for all work-eligible individuals in the dataset.

# Data

We use anonymized data for the Belgian customers of a large European bank between 2006 and 2016. The full dataset contains more than 3 million clients (only retail individuals) and covers detailed information about personal characteristics, financial portfolio, and every monetary transaction (incoming and outgoing) from their accounts. All monetary variables are deflated to 2006 money values.

Each transaction contains a flag indicating whether it refers a point of sale in Belgium, a cash withdrawal, a payment via credit card, etc. From those codes (see Supplementary Information (S.I.) for an overview), we construct relevant variables in order to adopt flexible definitions of an individual's imputed income and consumption. Table 3.1 reports all the variables used in this study.

In addition to transactional information, we include data on personal characteristics covering date of birth, civil state, gender, postal code, neighborhood code and nationality. The neighborhood code is a statistical sector variable constructed by the Belgian statistical office (Statbel) and constitutes the most granular geographical subdivision for government data collection [35].

Finally, using complementary data on every client's financial assets held at the bank, we construct a measure of financial wealth aggregating saving deposits, checking deposits, trading account deposits, pension savings, and financial insurance.

Our analysis is exclusively concerned with active clients starting their career between 2006 and 2016 and remaining active through the full time window of the study data. In the following we provide a brief overview of the identification

| Variable | Definition |
|---|---|
| Consumption (Yearly) | Debit card payments at P.O.S., cash withdrawals, written cheques, credit card payments, outgoing electronic transfers below a €$10,000$ single transaction value. |
| Income (Yearly) | Incoming electronic transfers below a given single transaction threshold during a year. If not stated otherwise, the cut-off value is €$5,000$. |
| Financial wealth (Yearly) | Saving deposits, checking deposits, term account deposits, trading account deposits, pension savings, and financial insurances. |
| Married | Dummy variable equal to $1$ if an individual is married. |
| Female | Dummy variable equal to $1$ if the individual is female. |
| Belgian | Dummy variable equal to $1$ if the individual is Belgian. |
| Start age | Age when individual started working (imputed as a categorical variable). |
| Start year | Year when individual started working (imputed as a categorical variable). |
| Region | Region where an individuals lives. There are $3$ distinct regions in Belgium. |
| Postal code | Postal code where an individuals lives. There are almost $3,000$ distinct postal codes in Belgium. |
| Neighborhood code | Neighborhood or statistical sector where an individuals lives. There are around $20,000$ distinct statistical sectors in Belgium with on average $540$ citizens per sector in 2008. |

*Table 3.1: Overview of all variables extracted from the data and their definition.*

procedure (see S.I. for a detailed description).

Conditioning on the presence of regular client activity (i.e., no inactivity lasting longer than 3 months), we are left with $852,331$ individuals who are eligible to work for the full sample period. The next step is to identify career starters and homogenize their starting point. Our identification consists of a series of thresholds based on consumption surveys used for living standards statistics in Belgium [36]. Throughout the procedure, we adopt a conservative approach: cleaning decisions are driven by the need to avoid false positives at the potential cost of losing false negatives.

The following conditions distinguish between ‘active and not working’ and 'active and working' for an individual in year *y*. An individual *i* is 'active and not working' in year *y* if their $Consumption_{i,y}$ >€300 and their $Income_{i,y}$ <€$6,000$. If their $Consumption_{i,y}$ >€$2,400$ and $Income_{i,y} \geq$ €$10,000$, they are categorised as 'active and working'.

The upper income bound for 'non-working years' (€$6,000$) is meant to account for the student worker earning limits in Belgium. The lower income bound for 'working years' (€$10,000$) is based on minimum wage statistics in Belgium. Further analyses of our main results with varying values of income bounds are provided in the S.I. and show no qualitative effects. Individuals with a yearly income between €$6,000$ and €$10,000$ are excluded from the sample. This condition excludes individuals who experienced discontinuous job experiences in their first year or started working late in the year. Using public statistics on first jobseeking individuals in Belgium, we assess that our sampling strategy covers the majority of the population (see S.I.). To control for outliers, we also trim those in the top $1\%$ of starting wealth and the top $1\%$ of income growth (other choices presented in S.I.).

Our framework identifies a career starter as a person for which there is exactly one persistent switch from 'active and not working' to 'active and working'. Their starting year is then the first calendar year in which they are categorized as 'active and not working'. To further reduce possible false positives, we focus on starting years between 2009 and 2014. Doing so ensures that at least three consecutive working and non-working years are observable. Since individuals have different starting years, we synchronize them on a working-years timeline. Finally, we restrict the starting age to a window between 18 and 30 years old. This window covers all the standard starting ages for individuals completing basic or higher education in Belgium.

Our final dataset consists of $17,576$ career starters with an observed income performance coverage between three and seven years into their careers. While the reduction in size from the original dataset is significant, the final population sample contains a unique level of granularity and reliability to study the effects of financial wealth on income mobility in the first years of a career.

In order to capture those exposed to a common environment, we filter out isolated individuals by dropping locations for which we have too few datapoints: the threshold is set to more than or equal to 10 and 3 career starters in the same year for postal codes and neighborhoods to be included, respectively.

## 3.4 Results

We first assess whether the data show evidence of differential income mobility for individuals with higher financial wealth.

To this end, let $T_{\Delta t}\left(i \rightarrow j|q\right)$ be the share of the population starting in the $q^{th}$

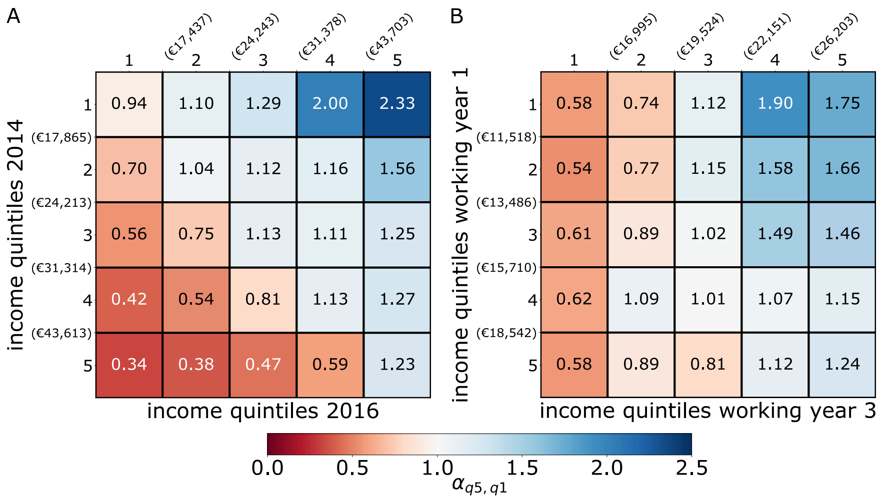



*Figure 3.2: Observed income mobility over three years for individuals starting in the highest quintile of financial wealth relative to those starting in the lowest financial wealth quintile, read from row to column.* ***A:*** *calculation based on the work-eligible sample.* ***B:*** *calculation based on the career-starter sample.*

quintile of the financial wealth distribution at $t_0$, moving from the $i^{th}$ income quintile at $t_1$, to the $j^{th}$ income quintile over $\Delta t$ years. Further let $\alpha_{q5,q1}$ be the ratio in this conditional transition probability between the highest ($5^{th}$) and lowest ($1^{st}$) financial wealth quintile, such that $\alpha_{q5,q1} = T_{\Delta t}\left(i \rightarrow j|q5\right)/T_{\Delta t}\left(i \rightarrow j|q1\right)$. Intuitively, this ratio indicates the comparative performance in income mobility between individuals with the highest and the lowest financial wealth. Note that financial wealth here is computed with a one-year lag. We motivate this choice below. Figure 3.2 reports values of $\alpha_{q5,q1}$ for all combinations of income quintile mobility for $\Delta t = 3$. Figure 3.2.A includes all work eligible individuals, whereas Figure 3.2.B is restricted to the population of career starters (synchronised on starting year $t_1$). The results indicate that over a relatively small timespan of three years, upward income mobility is dominated by individuals from the top financial wealth quintile (i.e., almost all right triangle cells are larger than one). Furthermore, the best possible income transition, $T_{\Delta t}\left(1 \rightarrow 5\right)$, is achieved by up to 2.3 times (1.75 times) more individuals from the top financial wealth quintile than individuals from the bottom financial wealth quintile for all work-eligible individuals (career starter-individuals). While the results for the career starters (Fig. 3.2.B) are less pronounced, they provide evidence at first glance of a positive relationship between financial wealth and early career income performance.

We report summary statistics on the sample of career starters in the S.I. Most career starters are Belgian, unwed, and around 22-23 years old in their first working calendar year. Gender is almost equally divided between males and females.

In the following, we unpack the relationship between financial wealth owned in the last non-working year (henceforth *start wealth*) and early career income performance. The reason we lag financial wealth by one year is driven by a specific Belgian rule, according to which first-time jobseekers (e.g., new graduates) have a legal waiting period of one year before being eligible for unemployment benefits. As such, a one-year lag ensures that the level of financial wealth we use is captured at the beginning of the waiting period or earlier. Our *start wealth* measure is therefore not affected by potential desaving or financial depletion during the waiting period.

Our first analysis quantifies the association of growth in income and *start wealth*. The baseline model is defined as follows:

$$\begin{aligned}\log(income_{i,t_{0+x}} + 1) = {} & constant + \alpha_1 \log(income_{i,t_1} + 1) + \alpha_2 \log(start\ wealth_i + 1) \\ & + \alpha_3 married_{i,t_{0+x}} + \alpha_4 female_i + \alpha_5 start\ age_i + \alpha_6 start\ year_i \\ & + \alpha_7 location_{i,t_0} + \alpha_8 Belgian_{i,t_{0+x}} + \alpha_9 married_{i,t_{0+x}} : female \\ & + \alpha_{10} \log(income_{i,t_1} + 1) : start\ age_i + \epsilon_i.\end{aligned} \tag{3.1}$$

The parameter $x$ is the year horizon, $t_1$ is the first working year, and $t_0$ the last non-working year. Note that *start age* and *start year* are not lagged and thus taken at time $t_1$. The $location$ variable controls for environmental factors, including location-based social capital and available amenities. It covers three different levels of geographical granularity: region, postal code, and neighborhood code. This ensures that the individuals were exposed to the same environmental influences. We note that while education and job occupation is not explicitly known, we control for *start age* and its interaction with starting income ($income_{i,t_1}$). This should largely capture the heterogeneity in education and the linked occupations within the sample. The model parameters are estimated using OLS methods with robust standard errors. The estimation for Eq. 3.1 is reported in Table 3.1. An extended version of the table, which explicitly shows the estimates for the control variables is presented in the S.I.

The results confirm that financial wealth in the last non-working year has a significantly positive effect on income performance during the first working years. More importantly, the granularity of the location controls has a limited effect on the main result. This indicates that social capital variables, as proxied by fixed effects for increasingly granular bins of living location, hold a restrained role in the transmission channel between financial wealth and early career income performance. In the case of Belgium, most social interactions take place at the neighbouring municipalities level [37]. Using location as a measure of social capital therefore allows to include both strong and weak social ties [38, 39]. All levels of

*Table 3.1: The effect of starting wealth on income performance three years (columns 1–3), five years (columns 4–6), and seven years (columns 7–9) into a career using Ordinary Least Squares (OLS) regressions. The columns within each horizon have different levels of location controls. The estimates for the used control variables (Married, Female, Start age, Start year, Region, and Belgian) are given in the S.I.*

| | *Dependent variable:* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | log(income after 3y + 1) | | | log(income after 5y + 1) | | | log(income after 7y + 1) | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| log(start income + 1) | 0.225*** | 0.230*** | 0.061 | 0.097* | 0.067 | −0.107 | 0.124* | 0.316** | 0.069 |
| | (0.041) | (0.088) | (0.208) | (0.054) | (0.098) | (0.246) | (0.068) | (0.151) | (0.275) |
| log(start wealth + 1) | 0.012*** | 0.013*** | 0.016*** | 0.013*** | 0.014*** | 0.018*** | 0.011*** | 0.015*** | −0.004 |
| | (0.001) | (0.002) | (0.004) | (0.001) | (0.003) | (0.005) | (0.002) | (0.003) | (0.009) |
| Constant | 7.575*** | 7.526*** | 9.269*** | 8.896*** | 9.085*** | 10.884*** | 8.794*** | 6.955*** | 9.242*** |
| | (0.395) | (0.836) | (1.984) | (0.515) | (0.932) | (2.351) | (0.651) | (1.436) | (2.618) |
| Personal controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | | ✓ | | | ✓ | | |
| Postal code | | ✓ | | | ✓ | | | ✓ | |
| Neighborhood code | | | ✓ | | | ✓ | | | ✓ |
| F Statistic | 76.73*** | 6*** | 5.37*** | 57.34*** | 5.13*** | 4.26*** | 36.55*** | 4.32*** | 7.91*** |
| Observations | 17,576 | 4,856 | 1,416 | 12,573 | 3,761 | 1,090 | 6,682 | 2,121 | 609 |
| Adjusted $R^2$ | 0.159 | 0.158 | 0.195 | 0.150 | 0.148 | 0.102 | 0.161 | 0.146 | 0.086 |

*Note:* $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ and robust standard errors in parentheses.
✓indicates that the controls are included.

location precision except the seven-year horizon with neighborhood controls are consistent. The drop in this case comes from the fact that sample size per neighborhoods in the seven-year horizon becomes too small. For this reason we do not include neighborhood controls in the remainder of the paper.

Absent social capital as the main driver of the positive impact of financial wealth on income mobility, we study the effect of innate ability as our second hypothesis. In fact, a high level of *start wealth* might originate from hard work and other innate abilities which could, in turn, translate into faster-rising income in the early stages of a career. In order to address this concern of endogeneity, we construct instrumental variables using categorized transactional data in the non-working years. Using said variables allows to distinguish between transferred financial wealth and self-originated financial wealth. The former is an exogenous transfer from another account (e.g., from parents) whereas the latter captures the financial wealth earned by the individual herself (e.g., from student jobs).

The first dummy instrument indicates whether an individual is the recipient of a large amount on her trading account (*financial boost*). The second and third instruments indicate whether an individual is the recipient of a transfer on a savings and/or current account (*cash boost*). These instruments mainly differ in the magnitude of the transfer and do not require the money to remain present on the respective account until $t_0$. Here, we set limits for the two *cash boosts* at respectively €$5,000$–$10,000$, and above €$10,000$; for the *financial boost*, the limit is set to €$5,000$ or more. In the S.I., we report several other settings and show that the related results produce qualitatively similar outcomes. Finally, we also construct a merged dummy variable (*recipient*) which is equal to $1$ if an individual has been subject to at least one of the above mentioned transfers in their non-working years.

We classify the population of career starters according to the *recipient* dummy variable and report the complete summary statistics of each resulting sample (recipients and non-recipients) in the S.I. Personal characteristics of the two samples do not vary by more than $3\%$ in the median. Income after five years however, is larger by almost $6\%$ for the *recipient* sample compared to the *non-recipient* sample.

The *IV* columns in Table 3.2 report the estimates of the model, where *start wealth* is instrumented with the three dummies, at different horizons and with different location controls. Instrumenting with these variables effectively excludes other sources of starting wealth including the ones related to innate abilities. The *Dummy* columns report the results where the *start wealth* parameter is replaced in the model by the merged *recipient* dummy variable. Directly using this dummy variable captures the effect of transferred financial wealth, independent of the amount of *start wealth* that was saved. Apart from column (8), all results from Table 3.2 maintain the sign and significance of the effect of *start wealth* from Table 3.1. The estimates for the control variables are presented in the S.I.

*Table 3.2: The effect of transferred financial wealth on income performance three years (columns 1–4), five years (columns 5–8), and seven years (columns 9–12) into a career using OLS regressions. The effect is estimated in two ways: first, in the* ***IV columns*** *by instrumenting start wealth with three dummies indicating the presence of different kinds of financial wealth transfers (+€5k financial boost, €5-10k cash boost, and +€10k cash boost); secondly, in the* ***Dummy columns*** *by combining the aforementioned dummies into one Recipient dummy indicating at least one kind of received (+€5k) financial wealth transfer and replacing the start wealth parameter by it. The columns within each estimation model have different levels of location controls. The estimates for the used control variables (Married, Female, Start age, Start year, Region, and Belgian) are given in the S.I.*

| | *Dependent variable:* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | log(income after 3y + 1) | | | | log(income after 5y + 1) | | | | log(income after 7y + 1) | | | |
| | *IV* | *IV* | *Dummy* | *Dummy* | *IV* | *IV* | *Dummy* | *Dummy* | *IV* | *IV* | *Dummy* | *Dummy* |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| log(start income + 1) | 0.220*** | 0.228** | 0.228*** | 0.230*** | 0.094* | 0.063 | 0.105* | 0.080 | 0.124* | 0.301** | 0.133* | 0.336** |
| | (0.041) | (0.088) | (0.041) | (0.087) | (0.054) | (0.098) | (0.054) | (0.099) | (0.068) | (0.149) | (0.069) | (0.156) |
| log(start wealth + $1)_{IV}$ | 0.020*** | 0.024*** | | | 0.017*** | 0.017** | | | 0.011* | 0.026** | | |
| | (0.003) | (0.006) | | | (0.004) | (0.008) | | | (0.006) | (0.011) | | |
| (+5k) Recipient Dummy | | | 0.037*** | 0.037*** | | | 0.032*** | 0.022 | | | 0.025** | 0.059*** |
| | | | (0.006) | (0.012) | | | (0.008) | (0.016) | | | (0.011) | (0.022) |
| Constant | 7.567*** | 7.477*** | 7.617*** | 7.607*** | 8.902*** | 9.105*** | 8.909*** | 9.045*** | 8.793*** | 7.026*** | 8.772*** | 6.875*** |
| | (0.394) | (0.843) | (0.396) | (0.827) | (0.514) | (0.931) | (0.520) | (0.938) | (0.651) | (1.407) | (0.663) | (1.485) |
| Personal controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | |
| Postal code | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ |
| Observations | 17, 576 | 4, 856 | 17, 576 | 4, 856 | 12, 573 | 3, 761 | 12, 573 | 3, 761 | 6, 682 | 2, 121 | 6, 682 | 2, 121 |
| Adjusted $R^2$ | 0.156 | 0.153 | 0.155 | 0.152 | 0.150 | 0.147 | 0.144 | 0.141 | 0.161 | 0.142 | 0.158 | 0.142 |
| Weak instruments | 0 | 0 | | | 0 | 0 | | | 0 | 0 | | |
| Wu-Hausman | 0 | 0.06 | | | 0.3 | 0.65 | | | 0.94 | 0.31 | | |
| Sargan | 0.37 | 0.97 | | | 0.36 | 0.49 | | | 0.13 | 0.1 | | |
| F Statistic | | | 73.56*** | 5.66*** | | | 54.43*** | 4.81*** | | | 36*** | 4.27*** |

*Note:* *p<0.1; **p<0.05; ***p<0.01 and robust standard errors in parentheses.
✓indicates that the controls are included.

Furthermore, Table 3.2 shows a consistently larger effect of transferred financial wealth, thus suggesting the limited explanatory power of innate abilities for the observed relationship between *start wealth* and income mobility. This result holds over all time horizons considered. The typical instrumental variable regression test statistics finds no complications with the validity of these results. Moreover, the Wu-Hausman test, which tests for endogeneity and thus for the consistency of ordinary least squares, does not find sufficient evidence of endogeneity beyond the three-year horizon. Interestingly, the results suggest that recipients experience higher income performance irrespective of whether they save or spend their transferred financial wealth.

Table 3.1 and 3.2 highlight that social capital and innate abilities do not fully cover the mechanisms driving the positive relationship between financial wealth and early career income performances. We now turn to a third candidate mechanism: the effect of *start wealth* on improving the search conditions of first-time jobseekers. Recall that Table 3.2 has shown that financial wealth transfers positively correlate with income mobility irrespective of the consumption behavior. Such finding supports the claim that recipients of financial wealth transfers may be less economically pressured to rush into securing their first job. As a result, these individuals might find opportunities which would match better their human capital and aspirations, therefore increasing their chances of higher productivity and higher income growth [32].

In order to test the validity of this job search condition hypothesis further, we explicitly estimate the effects of savings behavior. We proceed by constructing a *thrift* variable for the population and adding it to the model. An individual's level of *thrift* is computed as the share of the total incoming and available financial wealth that is consumed in a given time frame. Let this measure be $\textit{Thrift} \in [0, 1]$, with $0$ representing no saving and $1$ complete saving of income.

In principle, saving more extends the capacity of an individual to sustain themselves during periods without an income. Relaxing the budget constraints should provide a first-time jobseeker with less economic pressure. This alternative scenario should then also increase job matching quality in turn generating higher expected income performances for individuals exhibiting higher thrift [29, 30, 32]. By the same token, thrift should only have an effect on individuals whose external sources of financial wealth are limited. In the following, we test this hypothesis.

It is important to note that lower budget constraints do not need to translate into more time on the job market [40, 41]. In general, matchings in the optimal job market stopping problem is relative to both the sequence of job offers and the increased risk of running out of liquidity. Individuals with more financial wealth seeking a better match benefit from the lower costs of rejecting a job offer.

Formally, we have the following definition:

$$thrift_{i,y} = \frac{financial\ wealth_{t_0}}{financial\ wealth_{t_0-y} + \sum_{x=1}^{y}(incoming\ financial\ wealth_{t_0-x})} \tag{3.2}$$

where

$$incoming\ financial\ wealth_t = financial\ wealth_{t+1} - financial\ wealth_t \tag{3.3}$$

and $y$ is either equal to two or three years as we only consider the years before the start of the waiting period. Limiting the analysis to this period avoids reverse causality with income. It also treats *thrift* as a personal characteristic resulting from personal and family endowment, in line with the literature [12, 42]. Complete summary statistics on *thrift* are reported in the S.I. In the median, recipients save almost half of their incoming financial wealth, while only the top quartile of non-recipients match the same level of *thrift*. The combination of *thrift* and *recipient* can be used to re-assess the *start wealth* parameter in Table 3.1 according to two dimensions: one capturing the origin of *start wealth* (i.e., how much of the initial wealth results from transfers) and the other one capturing the savings behavior (i.e., how much of the wealth is saved or spent).

Table 3.3 shows the estimates for the model where *wealth* is replaced by the *two-* and *three-year thrift* variable and the *recipient* dummy variable. Table 3.3 also includes several horizons with different *location* controls. Note that only the *2y (two-year) thrift* can be shown for the seven-year horizon due to the extra year required (see Eq. 3.2). Once again, the estimates for the used control variables are presented in an extended version of Table 3.3 in the S.I. The results confirm a significant positive effect of being the recipient of transferred financial wealth on early career income performance. Furthermore, the results show a comparable positive effect of *thrift* on early career income performance. This suggests that financial wealth frugality before the start of a career is associated with greater income performance. Importantly, the negative interaction term between *thrift* and *recipient* shows that the marginal effect of *thrift* diminishes when an individual is simultaneously recipient of transferred financial wealth. The benefits of thriftiness and financial wealth transfer thus tend to substitute each other: a lack of transferred financial wealth can be compensated by more thriftiness but the effects do not add up. This substitutability also confirms the importance of budget constraints: being subject to large financial wealth transfers may indicate access to financial resources (e.g., parents) to support any absence of income thus alleviating pressure on the job market independently of the way such transfers are consumed. Overall, the results confirm the positive relationship between financial wealth, job search conditions, and early career income mobility.

*Table 3.3: The effect of transferred financial wealth and thrift on income performance three (columns 1–4), five years (columns 5–8), and seven years (columns 9–10) into a career using OLS regressions. Thrift is measured over the last two (2y thrift) or over the last three non-working years (3y thrift). The columns within each estimation model have different levels of location controls. The estimates for the used control variables (Married, Female, Start age, Start year, Region, and Belgian) are given in the S.I.*

| | *Dependent variable:* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | log(income after 3y + 1) | | | | log(income after 5y + 1) | | | | log(income after 7y + 1) | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| log(start income+1) | $0.219^{***}$ | $0.286^{***}$ | $0.217^{***}$ | $0.361^{***}$ | $0.160^{**}$ | $0.165$ | $0.189^{**}$ | $0.211^{*}$ | $0.101$ | $0.381$ |
| | (0.049) | (0.105) | (0.058) | (0.125) | (0.063) | (0.121) | (0.078) | (0.123) | (0.114) | (0.233) |
| (+5k) Recipient Dummy | $0.045^{***}$ | $0.044^{**}$ | $0.041^{***}$ | $0.041^{**}$ | $0.051^{***}$ | $0.045^{**}$ | $0.051^{***}$ | $0.028$ | $0.048^{**}$ | $0.097^{***}$ |
| | (0.010) | (0.019) | (0.010) | (0.018) | (0.012) | (0.020) | (0.014) | (0.024) | (0.019) | (0.037) |
| 2y thrift | $0.051^{***}$ | $0.073^{***}$ | | | $0.066^{***}$ | $0.082^{***}$ | | | $0.063^{***}$ | $0.100^{***}$ |
| | (0.008) | (0.016) | | | (0.011) | (0.019) | | | (0.016) | (0.032) |
| 3y thrift | | | $0.051^{***}$ | $0.079^{***}$ | | | $0.067^{***}$ | $0.084^{***}$ | | |
| | | | (0.010) | (0.020) | | | (0.014) | (0.028) | | |
| (+5k) Recipient Dummy :2y thrift | $-0.042^{**}$ | $-0.054$ | | | $-0.074^{***}$ | $-0.096^{***}$ | | | $-0.074^{**}$ | $-0.153^{**}$ |
| | (0.018) | (0.033) | | | (0.019) | (0.028) | | | (0.034) | (0.069) |
| (+5k) Recipient Dummy :3y thrift | | | $-0.038^{**}$ | $-0.047^{*}$ | | | $-0.064^{***}$ | $-0.043$ | | |
| | | | (0.018) | (0.028) | | | (0.023) | (0.035) | | |
| Constant | $7.703^{***}$ | $7.081^{***}$ | $7.737^{***}$ | $6.323^{***}$ | $8.389^{***}$ | $8.253^{***}$ | $8.139^{***}$ | $7.778^{***}$ | $9.107^{***}$ | $6.485^{***}$ |
| | (0.468) | (0.997) | (0.556) | (1.189) | (0.600) | (1.154) | (0.750) | (1.182) | (1.079) | (2.205) |
| Personal controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | |
| Postal code | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ |
| F Statistic | $68.14^{***}$ | $5.8^{***}$ | $53.38^{***}$ | $4.73^{***}$ | $50.28^{***}$ | $4.74^{***}$ | $35.45^{***}$ | $3.83^{***}$ | $32.55^{***}$ | $8.3^{***}$ |
| Observations | 17,109 | 4,735 | 13,382 | 3,547 | 12,103 | 3,640 | 8,383 | 2,458 | 6,216 | 2,001 |
| Adjusted $R^2$ | 0.155 | 0.155 | 0.153 | 0.144 | 0.145 | 0.141 | 0.140 | 0.132 | 0.162 | 0.147 |

*Note:* $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ and robust standard errors in parentheses.
✓indicates that the controls are included.

## 3.5 Discussion

The major contribution of this paper is to show a substantial effect of wealth distribution on income mobility as early as three years into a career. We find a limited contribution of social capital and innate abilities to fully explain this effect. Finally, we find evidence that a channel of transmission exists when wealth inequality translates into differential economic conditions among first-time jobseekers.

Financial wealth, by extending an individual's capacity to remain without income, relaxes search frictions in the job market. More precisely, it alleviates the optimal stopping problem in labor markets resulting from the tradeoff between job opportunities and the uncertainty in the job search sequence [29]. Less economic pressure increases the expected quality of the achieved match with a job seeker's human capital. A better match, in turn, increases the likelihood of higher productivity and income performance in the early stages of a career [30, 32]. In contrast, less financially wealthy individuals face higher economic pressure and will tend to accept job offers at the cost of a low match, thus reducing their expected income mobility. This interplay provides a channel of transmission between intergenerational financial wealth transfers and income mobility at the earliest stage. Additionally, the interaction between financial wealth and job search conditions also offers evidence of adverse selection in the wealth inequality distribution for early contracting and market unraveling [31, 43]. Policies addressing these issues may be key to promoting income mobility.

Our analysis was limited to a seven-year time frame and had limited knowledge of parental affiliation, educational background and job occupation for each career starter. Future work featuring a longer time horizon, with explicit intergenerational mapping and a full knowledge of the career path is necessary to assess the scope of our findings and assess potential heterogeneous effects in education background and job occupation cross sections. Such results would help guide policy designs to support relevant segments of the population.

While testing the robustness of our results for other economies is paramount, the case of Belgium already provides interesting insights for three main reasons. First, Belgium ranks high in both equality and mobility worldwide (see Figure 3.1). Second, in contrast with several other developed economies, Belgium has a very low rate of consumer debt [34]. In particular, the Belgian educational system is mainly public and affordable (e.g., Belgium does not have a student debt system). Finally, according to Belgian law, an individual with no income record must wait for one year before being entitled to unemployment benefits, thus providing a suitable test-bed with a one-year lag between initial financial wealth and early career income performances. Overall, these remarks suggest that our results should provide lower bar estimates for the interaction between wealth inequality and income mobility. Countries such as the United States featuring higher wealth inequality and larger sources of budget pressure, including large tuition fees and student debts will likely exhibit more-pronounced effects.

# Appendix

## 3.A Detailed cleaning procedure

Here we describe the procedure used to construct the variables and extract the subset in detail from the raw financial data.

For clarity the data is subdivided in two parts. One part contains every single financial transaction of all the clients, henceforth 'the transaction dataset'. The other part contains a yearly snapshot of the personal characteristics, henceforth 'the characteristics dataset'. The time frame of both datasets is from 2006 to 2016.

The characteristics dataset contains an individual start-of-year snapshot of personal as well as financial characteristics. The personal characteristics are: age, civil state, and location of domicile on the level of region, postal code, and neighbourhood code (government defined statistical sector). The financial characteristics are the balances of their savings, checking, term, and trading accounts, as well as their pension savings and any financial insurances they might have at the bank. These financial characteristics are aggregated into the variable *Financial wealth*.

In the transaction dataset, every transaction contains a flag to describe its type. Table 3.A.1 gives an overview of the utilized flags from the transaction dataset and how they were categorized. Because the characteristics dataset only provides a yearly start-of-year snapshot, we aggregate the transaction over one calendar year.

The total amount of clients included in these datasets is over three million (exact number redacted due to confidentiality). However, among these are also inactive clients, clients with only a savings account, etc. To exclude these individuals, we condition on the presence of a current account and on account activity in a three month rolling window over the complete time frame of the data. Since we are interested in career performances, we also limit the age between 15-50 in 2006 such that the subsample is work-eligible throughout the whole time frame. This active, work eligible subsample contains $852,331$ individuals and is used in Fig. 3.1.B, 3.1.C, and 3.2.A in the main text.

In light of the literature however, as explained in the main text, we further focus on career starters. Two conditions are used to categorize each person-year as 'active and working' or 'active and not working':

| Transaction flag | Explanation |
|---|---|
| Consumption | The flags below sum to the consumption of an individual. |
| outgoing cash | Cash withdrawals. |
| outgoing cheque | Consumption cheques. Cheques commonly used for large purchases (e.g. houses) are excluded. |
| credit card purchase | |
| P.O.S. purchase | Purchases paid for at point of sale. |
| Trans. out -1k | Outgoing electronic transfer below €$1,000$ value. |
| Trans. out 1-2k | Outgoing electronic transfer between €$1,000$ and €$2,000$ value. |
| Trans. out 2-5k | Outgoing electronic transfer between €$2,000$ and €$5,000$ value. |
| Trans. out 5-10k | Outgoing electronic transfer between €$5,000$ and €$10,000$ value. |
| Trans. out +10k | Outgoing electronic transfer above €$10,000$ value. |
| Income | Depending on the definition used, (some of) the flags below make up the income of an individual. The 'Narrowplus' definition includes all flags below, the 'Narrow' definition excludes cheque and cash deposits, and the 'Supernarrow' definition further excludes Trans. in +10k. Note that none of the income definitions includes rents, interests, or dividends coming from financial wealth (neither in liquid form nor stocks or bonds). |
| Cash deposits | |
| Cheque deposit | |
| Trans. in -1k | Incoming electronic transfer below €$1,000$ value. |
| Trans. in 1-2k | Incoming electronic transfer between €$1,000$ and €$2,000$ value. |
| Trans. in 2-5k | Incoming electronic transfer between €$2,000$ and €$5,000$ value. |
| Trans. in 5-10k | Incoming electronic transfer between €$5,000$ and €$10,000$ value. |
| Trans. in +10k | Incoming electronic transfer above €$10,000$ value. |

*Table 3.A.1: Overview of all utilized flags extracted from the transaction dataset (indented), the category under which they fall (not indented), and extra explanation where the meaning of the flag is not self-explanatory.*

**Condition 1: Individual $i$ active and Not Working in year *y***

$$Consumption_{i,y} > €300 \quad \text{and} \quad Income_{i,y} < €6,000$$

**Condition 2: Individual $i$ active and Working in year *y***

$$Consumption_{i,y} > €2,400 \quad \text{and} \quad Income_{i,y} \geq €10,000$$

The upper income bound for ‘not working years’ ($€6,000$) is meant to account for the student worker earning limits in Belgium. The lower income bound for ‘working years’ ($€10,000$) is loosely based on minimum wage in Belgium A career starter is then identified by having one persistent switch from ‘not working’ to ‘working’ within the time frame. The year in which this switch occurs is called the *Start year*. To have at least three consecutive ‘working’ and ‘not working’ years available, we only keep individuals with *Start year* between 2009-2014. Lastly, to limit false positives further, we restrict the age of individuals in their *Start year* between 18-30. This choice also excludes the underage career starters (-18). This is done because minors still have part-time compulsory attendance at school and need to comply with special requirements before they are allowed to start working. We are left with $17,576$ career starters with an observed income performance coverage between three to seven years into their careers. This subset is used for the regressions and results in the main text.

## 3.B Summary statistics

Here we present summary statistics for our starters sample. Table 3.B.1 presents the summary statistics for the general career starter population. Table 3.B.2 shows the same summary statistics but now for the (+5k) recipients and non-recipient career starters separately. Lastly, Table 3.B.3 gives insights into the distribution of the *Thrift* parameter, as defined in Eq.2 in the main text.

| | count | mean (€) | std (€) | min (€) | 25p (€) | median (€) | 75p (€) | max (€) |
|---|---|---|---|---|---|---|---|---|
| variable | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Start income | 12,573 | 15,053.31 | 4,035.20 | 10,000.22 | 11,899.90 | 14,347.53 | 17,268.66 | 63,769.25 |
| Income after 5y | 12,573 | 23,407.27 | 7,335.20 | 10,000.01 | 18,746.72 | 22,070.37 | 26,588.98 | 80,557.07 |
| Start wealth | 12,573 | 4,584.54 | 7,135.48 | 0.00 | 302.26 | 1,670.50 | 5,801.79 | 52,637.47 |
| Married | 12,573 | 0.08 | 0.26 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Female | 12,573 | 0.48 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Belgian | 12,573 | 0.97 | 0.17 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Start age | 12,573 | 22.15 | 2.35 | 18.00 | 21.00 | 22.00 | 23.00 | 30.00 |
| Start year | 12,573 | 2010.42 | 1.10 | 2009 | 2009 | 2010 | 2011 | 2012 |
| Consumption | 12,573 | 12,952.35 | 5,849.63 | 2,402.16 | 9,455.14 | 12,143.73 | 15,440.74 | 106,833.38 |
| Start inhabitants (neighborhood code) | 12,573 | 1.42 | 0.74 | 1.00 | 1.00 | 1.00 | 2.00 | 6.00 |
| Start inhabitants (postal code) | 12,573 | 7.93 | 6.36 | 1.00 | 3.00 | 6.00 | 11.00 | 36.00 |

*Table 3.B.1: Summary statistics for all career starters that have worked for five years within the dataset's time frame. Timing of the variables is as specified in the model (Eq.1 main text).*

| | count | mean (€) | std (€) | min (€) | 25p (€) | median (€) | 75p (€) | max (€) |
|---|---|---|---|---|---|---|---|---|
| variable | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| **(+5k) recipients** | | | | | | | | |
| Start income | 1,571 | 15,589.04 | 4,829.58 | 10,001.60 | 12,001.14 | 14,781.93 | 17,896.22 | 61,040.13 |
| Income after 5y | 1,571 | 24,550.76 | 8,093.06 | 10,094.53 | 19,338.51 | 23,166.11 | 28,259.18 | 67,366.46 |
| Start wealth | 1,571 | 13,180.25 | 12,059.68 | 0.00 | 3,173.45 | 10,090.58 | 19,511.21 | 52,537.42 |
| Married | 1,571 | 0.09 | 0.29 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Female | 1,571 | 0.49 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Belgian | 1,571 | 0.98 | 0.14 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Start age | 12,573 | 22.54 | 2.36 | 18.00 | 21.00 | 22.00 | 24.00 | 30.00 |
| Start year | 1,571 | 2010.50 | 1.10 | 2009 | 2010 | 2010 | 2011 | 2012 |
| Consumption | 1,571 | 13,732.65 | 8,451.55 | 2,482.81 | 8,937.25 | 12,271.92 | 16,460.64 | 106,833.38 |
| Financial boost (+€5k) | 1,571 | 0.35 | 0.48 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Cash boost (€5-10k) | 1,571 | 0.48 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Cash boost (+€10k) | 1,571 | 0.36 | 0.48 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| **(+5k) non-recipients** | | | | | | | | |
| Start income | 11,002 | 14,976.81 | 3,902.88 | 10,000.22 | 11,884.39 | 14,299.50 | 17,178.03 | 63,769.25 |
| Income after 5y | 11,002 | 23,243.99 | 7,206.10 | 10,000.01 | 18,690.14 | 21,941.95 | 26,364.64 | 80,557.07 |
| Start wealth | 11,002 | 3,357.14 | 5,037.09 | 0.00 | 252.62 | 1,316.26 | 4,431.52 | 52,637.47 |
| Married | 11,002 | 0.07 | 0.26 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Female | 11,002 | 0.48 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Belgian | 11,002 | 0.97 | 0.17 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Start age | 11,002 | 22.16 | 2.34 | 18.00 | 21.00 | 22.00 | 23.00 | 30.00 |
| Start year | 11,002 | 2010.41 | 1.10 | 2009 | 2009 | 2010 | 2011 | 2012 |
| Consumption | 11,002 | 12,840.93 | 5,367.62 | 2,402.16 | 9,547.65 | 12,134.02 | 15,333.21 | 97,545.70 |

*Table 3.B.2: Summary statistics for two groups of career starters that have worked for* $5$ *years within the dataset's time frame.* ***Top:*** *Summary statistics for career starters who received over* €$5,000$ *in transferred financial wealth before the start of their careers ((+5k) recipients). The correlation between financial boost (+*€$5,000$*) and the combined cash boost is* $-0.72$ *(*$-0.22$ *for +*€$10,000$ *and* $-0.41$ *for* €$5,000$-€$10,000$*).* ***Bottom:*** *Summary statistics for career starters who received less than* €$5,000$ *in transferred wealth before the start of their careers ((+5k) non-recipients)*

| variable | count (1) | mean (2) | std (3) | min (4) | 25p (5) | median (6) | 75p (7) | max (8) |
|---|---|---|---|---|---|---|---|---|
| **Full sample** | | | | | | | | |
| 2y thrift | 12,103 | 0.27 | 0.30 | 0.00 | 0.04 | 0.17 | 0.46 | 1 |
| 3y thrift | 8,383 | 0.24 | 0.28 | 0.00 | 0.03 | 0.14 | 0.39 | 1 |
| **(+5k) recipients** | | | | | | | | |
| 2y thrift | 1,524 | 0.48 | 0.44 | 0.00 | 0.14 | 0.50 | 0.73 | 1 |
| 3y thrift | 1,116 | 0.45 | 0.44 | 0.00 | 0.12 | 0.44 | 0.67 | 1 |
| **(+5k) non-recipients** | | | | | | | | |
| 2y thrift | 10,579 | 0.24 | 0.26 | 0.00 | 0.03 | 0.15 | 0.40 | 1 |
| 3y thrift | 7,267 | 0.21 | 0.23 | 0.00 | 0.02 | 0.11 | 0.34 | 1 |

*Table 3.B.3: Summary statistics for the thrift of career starters in the dataset that worked for five years. The top part of the table shows thrift values for the full sample while the bottom two parts show thrift values for those that got transferred at least* €5,000 *wealth at least once, and those that did not, respectively.*

## 3.C Effects of income thresholds

As mentioned in the main text, the chosen thresholds in income (condition 1 and 2) cause some false negatives among the detected career starters. More specifically, if the timing of the start of an individual's career causes them to fall between €$6,000$ and €$10,000$ income in their first working year, they are dropped from the sample. Here we use publicly available data on recently graduated job seekers, collected by the public employment service of Flanders (VDAB) and made available via their online tool Arvastat [44]. We utilize this data to estimate the impact of the income thresholds on capturing career starters in the data.

Among the unemployed, we select only the BIT category. Individuals that have completed their education, before having the right to unemployment benefits, have to register for this BIT status. During periods of unemployment within the first 12 months after registration, the individual will remain in this category. After this 12 month period, the individual will have the right to unemployment benefits and will no longer appear in this category when unemployed. This 12 month period is the waiting period before being eligible for unemployment benefits which is referred to in the main text. When a person has worked the majority of a month, they automatically get unsubscribed from the BIT category in that month. A decline in BIT individuals thus means a net outflow of newly graduated unemployed. Monthly data on the amount of individuals with BIT status for the $< 25$ age category and per education level was extracted. Patterns did not differ significantly when including the $25 - 39$ age category.

We calculate, for several levels of income, how many of the yearly starters are captured. Figures 3.C.1, 3.C.2, and 3.C.3 show such plots for a starting net income of €$1,300$, €$1,600$, and €$2,100$ respectively. The three line per plot show the monthly amount of people per education level present in the BIT category. The shaded regions show the months in which starters are detected by our cut-offs and included in our sample. It can be seen that the majority of periods in which large declines of BIT membership occurs are captured. The plots show the following features:

- The periodicity is largely the same for all three levels of education.
- For the wages between €$1,250$-$1,500$ we capture the full wave of hiring.
- For wages between €$1,500$-$2,000$ we drop the people that started their career in September but you can see that there is a larger net outflow in the months from October to the end of May, which we do include.
- For the extremes of the distribution, namely those making less that €$1,250$ or more than €$2,000$, we do somewhat worse.

Since the VDAB reports that the median wages are: €$1,550$ for low-skilled (did not complete upper secondary education), €$1,635$ for medium-skilled (completed upper secondary education or did a trade year), and €$1,700$ for high-skilled (college or university) our selection criteria works well for a large majority of the people. Since these statistics only capture the net in-/out-flow of the BIT-category, it is hard to quantify the exact number of career starters we capture, however for a wage between €$1,500$ - €$2,000$ we should capture $67\%$ of the net outflow of job seekers when only taking into account the timing of outflow. Looking at national statistics, a rough estimation of yearly starters lies around $80,000$ individuals per year (based on statistics of -30 year old individuals that leave school and find a job within a year). This means that, given our sample size, we capture at least three to five percent of the total yearly population of job starters in Belgium, depending on the horizon looked at.

*Figure 3.C.1: Impact of false negatives from income thresholds in selection procedure of career starters. The amount of low-skilled (blue), medium-skilled (orange), and high-skilled (red) individuals present in the BIT category. The shaded area marks the months in which career starters, with a starting wage of €1,300, are included in the sample. The individuals starting in the darker shaded months are included among the career starters in the next year, those in the lighter shaded months in the same year.*

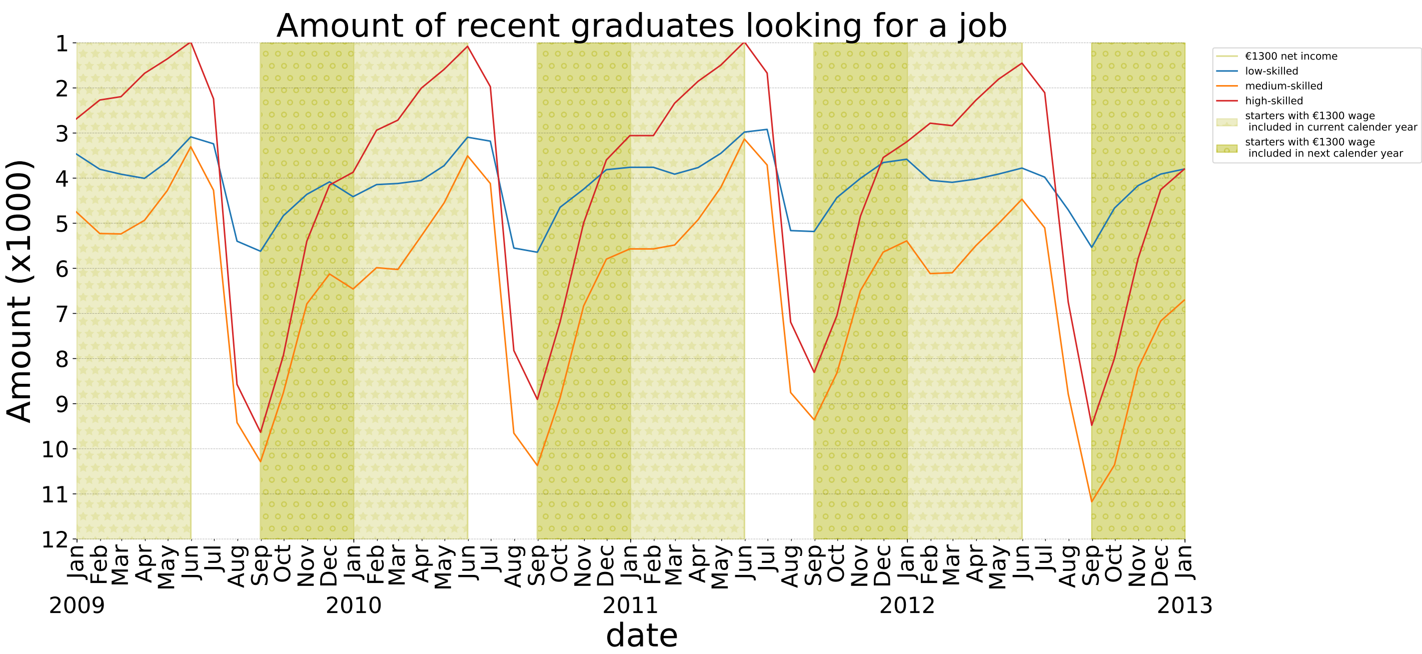


*Figure 3.C.2: Impact of false negatives from income thresholds in selection procedure of career starters. The amount of low-skilled (blue), medium-skilled (orange), and high-skilled (red) individuals present in the BIT category. The shaded area marks the months in which career starters, with a starting wage of €1,600, are included in the sample. The individuals starting in the darker shaded months are included among the career starters in the next year, those in the lighter shaded months in the same year.*

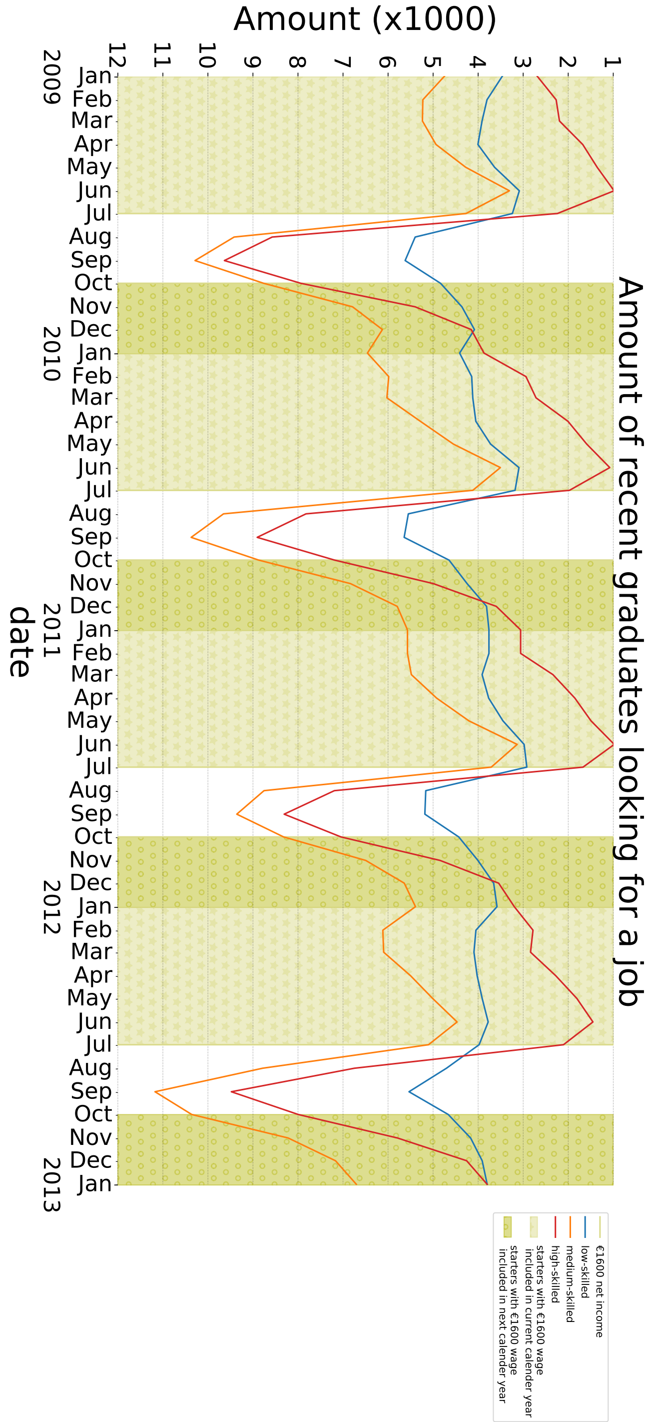
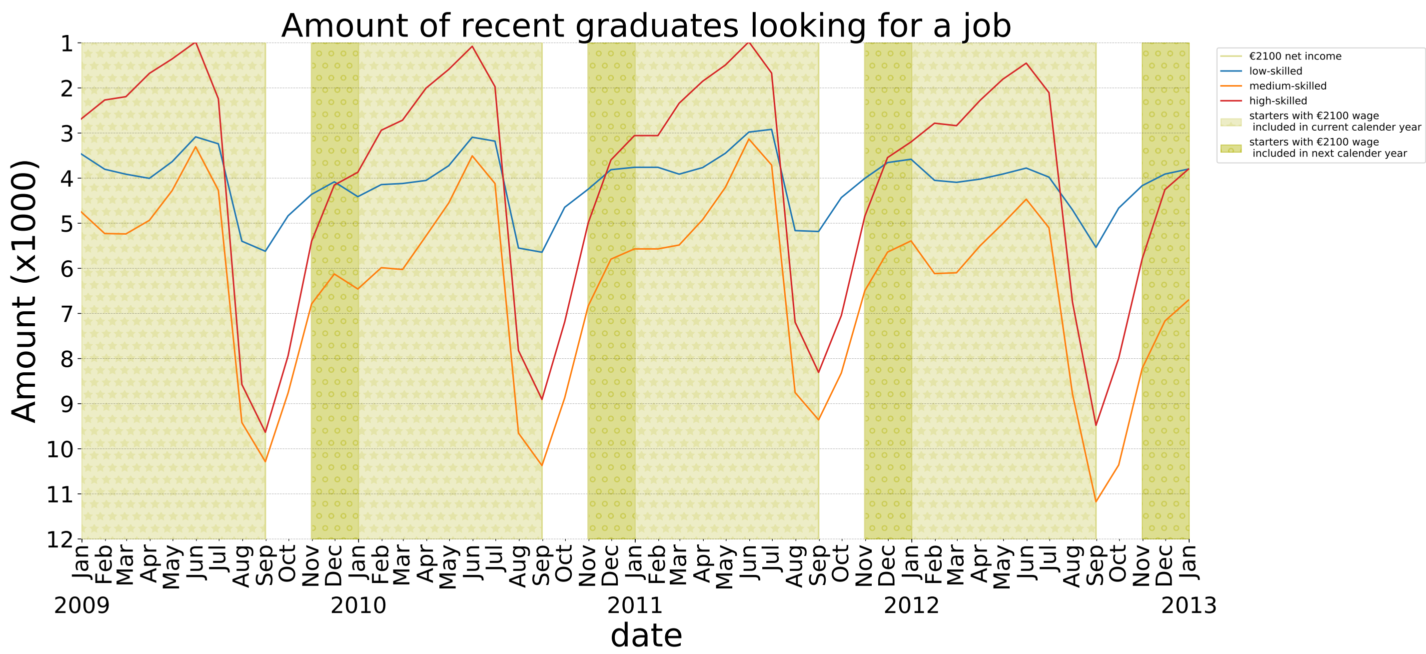


*Figure 3.C.3: Impact of false negatives from income thresholds in selection procedure of career starters. The amount of low-skilled (blue), medium-skilled (orange), and high-skilled (red) individuals present in the BIT category. The shaded area marks the months in which career starters, with a starting wage of €2,100, are included in the sample. The individuals starting in the darker shaded months are included among the career starters in the next year, those in the lighter shaded months in the same year.*



## 3.D Extended regression tables

Here we present the extended versions of the regression tables in the main text, which include explicit estimations of the most important control variables. To contain the length of the tables within page-limits, we present estimations for the *start year* and *start age* control variables in the next section( 3.E).

*Table 3.D.1: The effect of starting wealth on income performance three years (columns 1–3), five years (columns 4–6), and seven years (columns 7–9) into a career using Ordinary Least Squares (OLS) regressions. The columns within each horizon have different levels of location controls.*

| | *Dependent variable:* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | log(income after 3y + 1) | | | log(income after 5y + 1) | | | log(income after 7y + 1) | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| log(start income + 1) | $0.225^{***}$ | $0.230^{***}$ | 0.061 | $0.097^{*}$ | 0.067 | −0.107 | $0.124^{*}$ | $0.316^{**}$ | 0.069 |
| | (0.041) | (0.088) | (0.208) | (0.054) | (0.098) | (0.246) | (0.068) | (0.151) | (0.275) |
| log(start wealth + 1) | $0.012^{***}$ | $0.013^{***}$ | $0.016^{***}$ | $0.013^{***}$ | $0.014^{***}$ | $0.018^{***}$ | $0.011^{***}$ | $0.015^{***}$ | −0.004 |
| | (0.001) | (0.002) | (0.004) | (0.001) | (0.003) | (0.005) | (0.002) | (0.003) | (0.009) |
| Married | $0.193^{***}$ | $0.185^{***}$ | $0.413^{***}$ | $0.220^{***}$ | $0.239^{***}$ | $0.200^{***}$ | $0.247^{***}$ | $0.245^{***}$ | $0.244^{***}$ |
| | (0.021) | (0.037) | (0.074) | (0.017) | (0.031) | (0.051) | (0.019) | (0.032) | (0.065) |
| Female | $-0.077^{***}$ | $-0.064^{***}$ | $-0.043^{***}$ | $-0.108^{***}$ | $-0.086^{***}$ | $-0.065^{***}$ | $-0.136^{***}$ | $-0.109^{***}$ | $-0.075^{**}$ |
| | (0.004) | (0.008) | (0.016) | (0.005) | (0.010) | (0.021) | (0.008) | (0.014) | (0.035) |
| Belgian | $0.044^{***}$ | 0.026 | −0.024 | $0.043^{***}$ | 0.038 | −0.008 | 0.023 | 0.026 | 0.018 |
| | (0.013) | (0.020) | (0.051) | (0.016) | (0.025) | (0.049) | (0.023) | (0.037) | (0.087) |
| Married:Female | $-0.140^{***}$ | $-0.156^{***}$ | $-0.382^{***}$ | $-0.139^{***}$ | $-0.181^{***}$ | $-0.148^{**}$ | $-0.202^{***}$ | $-0.187^{***}$ | $-0.308^{***}$ |
| | (0.026) | (0.043) | (0.088) | (0.021) | (0.039) | (0.075) | (0.025) | (0.044) | (0.094) |
| Constant | $7.575^{***}$ | $7.526^{***}$ | $9.269^{***}$ | $8.896^{***}$ | $9.085^{***}$ | $10.884^{***}$ | $8.794^{***}$ | $6.955^{***}$ | $9.242^{***}$ |
| | (0.395) | (0.836) | (1.984) | (0.515) | (0.932) | (2.351) | (0.651) | (1.436) | (2.618) |
| Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start year | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start income:Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | | ✓ | | | ✓ | | |
| Postal code | | ✓ | | | ✓ | | | ✓ | |
| Neighborhood code | | | ✓ | | | ✓ | | | ✓ |
| F Statistic | $76.73^{***}$ | $6^{***}$ | $5.37^{***}$ | $57.34^{***}$ | $5.13^{***}$ | $4.26^{***}$ | $36.55^{***}$ | $4.32^{***}$ | $7.91^{***}$ |
| Breusch-Pagan | 0 | 0.003 | 0 | 0 | 0.033 | 0 | 0 | 0.079 | 0 |
| Observations | 17,576 | 4,856 | 1,416 | 12,573 | 3,761 | 1,090 | 6,682 | 2,121 | 609 |
| $R^2$ | 0.161 | 0.187 | 0.435 | 0.153 | 0.183 | 0.377 | 0.166 | 0.199 | 0.400 |
| Adjusted $R^2$ | 0.159 | 0.158 | 0.195 | 0.150 | 0.148 | 0.102 | 0.161 | 0.146 | 0.086 |
| Residual Std. Error | 0.245 | 0.251 | 0.239 | 0.270 | 0.276 | 0.269 | 0.295 | 0.295 | 0.304 |
| | ($df$ = 17538) | ($df$ = 4688) | ($df$ = 994) | ($df$ = 12537) | ($df$ = 3605) | ($df$ = 755) | ($df$ = 6648) | ($df$ = 1988) | ($df$ = 399) |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 and robust standard errors in parentheses. ✓ indicates that controls are included.

*Table 3.D.2: The effect of transferred financial wealth on income performance three years (columns 1–4), five years (columns 5–8), and seven years (columns 9–12) into a career using OLS regressions. The effect is estimated in two ways: first, in the **IV columns** by instrumenting start wealth with three dummies indicating the presence of different kinds of financial wealth transfers (+€5k financial boost, €5-10k cash boost, and +€10k cash boost); secondly, in the **Dummy columns** by combining the aforementioned dummies into one Recipient dummy indicating at least one kind of received (+€5k) financial wealth transfer and replacing the start wealth parameter by it. The columns within each estimation model have different levels of location controls.*

| | *Dependent variable:* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | log(income after 3y + 1) | | | | log(income after 5y + 1) | | | | log(income after 7y + 1) | | | |
| | *IV* | *IV* | *Dummy* | *Dummy* | *IV* | *IV* | *Dummy* | *Dummy* | *IV* | *IV* | *Dummy* | *Dummy* |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| log(start income + 1) | 0.220*** | 0.228** | 0.228*** | 0.230*** | 0.094* | 0.063 | 0.105* | 0.080 | 0.124* | 0.301** | 0.133* | 0.336** |
| | (0.041) | (0.088) | (0.041) | (0.087) | (0.054) | (0.098) | (0.054) | (0.099) | (0.068) | (0.149) | (0.069) | (0.156) |
| log(start wealth + 1)$_{IV}$ | 0.020*** | 0.024*** | | | 0.017*** | 0.017** | | | 0.011* | 0.026** | | |
| | (0.003) | (0.006) | | | (0.004) | (0.008) | | | (0.006) | (0.011) | | |
| (+5k) Recipient Dummy | | | 0.037*** | 0.037*** | | | 0.032*** | 0.022 | | | 0.025** | 0.059*** |
| | | | (0.006) | (0.012) | | | (0.008) | (0.016) | | | (0.011) | (0.022) |
| Married | 0.192*** | 0.180*** | 0.195*** | 0.188*** | 0.219*** | 0.238*** | 0.221*** | 0.243*** | 0.247*** | 0.239*** | 0.251*** | 0.252*** |
| | (0.021) | (0.037) | (0.021) | (0.037) | (0.017) | (0.031) | (0.017) | (0.030) | (0.019) | (0.033) | (0.019) | (0.032) |
| Female | −0.079*** | −0.066*** | −0.075*** | −0.061*** | −0.109*** | −0.087*** | −0.105*** | −0.083*** | −0.136*** | −0.111*** | −0.134*** | −0.105*** |
| | (0.004) | (0.008) | (0.004) | (0.008) | (0.005) | (0.010) | (0.005) | (0.010) | (0.008) | (0.014) | (0.008) | (0.014) |
| Belgian | 0.042*** | 0.024 | 0.045*** | 0.026 | 0.042** | 0.038 | 0.045*** | 0.039 | 0.023 | 0.027 | 0.024 | 0.023 |
| | (0.013) | (0.021) | (0.013) | (0.020) | (0.016) | (0.025) | (0.016) | (0.026) | (0.023) | (0.037) | (0.023) | (0.037) |
| Married:Female | −0.137*** | −0.149*** | −0.145*** | −0.164*** | −0.138*** | −0.180*** | −0.143*** | −0.186*** | −0.202*** | −0.180*** | −0.207*** | −0.195*** |
| | (0.026) | (0.043) | (0.026) | (0.043) | (0.021) | (0.039) | (0.021) | (0.039) | (0.025) | (0.044) | (0.025) | (0.044) |
| Constant | 7.567*** | 7.477*** | 7.617*** | 7.607*** | 8.902*** | 9.105*** | 8.909*** | 9.045*** | 8.793*** | 7.026*** | 8.772*** | 6.875*** |
| | (0.394) | (0.843) | (0.396) | (0.827) | (0.514) | (0.931) | (0.520) | (0.938) | (0.651) | (1.407) | (0.663) | (1.485) |
| Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start year | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start income:Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | |
| Postal code | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ |
| Observations | 17,576 | 4,856 | 17,576 | 4,856 | 12,573 | 3,761 | 12,573 | 3,761 | 6,682 | 2,121 | 6,682 | 2,121 |
| Breusch-Pagan | 0 | 0.003 | 0 | 0.001 | 0 | 0.033 | 0 | 0.021 | 0 | 0.079 | 0 | 0.06 |
| $R^2$ | 0.158 | 0.182 | 0.156 | 0.181 | 0.152 | 0.182 | 0.147 | 0.176 | 0.166 | 0.195 | 0.162 | 0.195 |
| Adjusted $R^2$ | 0.156 | 0.153 | 0.155 | 0.152 | 0.150 | 0.147 | 0.144 | 0.141 | 0.161 | 0.142 | 0.158 | 0.142 |
| Weak instruments | 0 | 0 | | | 0 | 0 | | | 0 | 0 | | |
| Wu-Hausman | 0 | 0.06 | | | 0.3 | 0.65 | | | 0.94 | 0.31 | | |
| Sargan | 0.37 | 0.97 | | | 0.36 | 0.49 | | | 0.13 | 0.1 | | |
| Residual Std. Error | 0.246 | 0.252 | 0.246 | 0.252 | 0.270 | 0.276 | 0.271 | 0.278 | 0.295 | 0.296 | 0.295 | 0.296 |
| | ($df$ = 17538) | ($df$ = 4688) | ($df$ = 17538) | ($df$ = 4688) | ($df$ = 12537) | ($df$ = 3605) | ($df$ = 12537) | ($df$ = 3605) | ($df$ = 6648) | ($df$ = 1988) | ($df$ = 6648) | ($df$ = 1988) |
| F Statistic | | | 73.56*** | 5.66*** | | | 54.43*** | 4.81*** | | | 36*** | 4.27*** |

*Note:* *p<0.1; **p<0.05; ***p<0.01 and robust standard errors in parentheses. ✓ indicates that controls are included.

*Table 3.D.3: The effect of transferred financial wealth and thrift on income performance three (columns 1–4), five years (columns 5–8), and seven years (columns 9–10) into a career using OLS regressions. Thrift is measured over the last two ($2y\ thrift$) or over the last three non-working years ($3y\ thrift$). The columns within each estimation model have different levels of location controls.*

| | *Dependent variable:* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | log(income after 3y + 1) | | | | log(income after 5y + 1) | | | | log(income after 7y + 1) | |
| | $2y\ thrift$ | $2y\ thrift$ | $3y\ thrift$ | $3y\ thrift$ | $2y\ thrift$ | $2y\ thrift$ | $3y\ thrift$ | $3y\ thrift$ | $2y\ thrift$ | $2y\ thrift$ |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| log(start income+1) | $0.219^{***}$ | $0.286^{***}$ | $0.217^{***}$ | $0.361^{***}$ | $0.160^{**}$ | $0.165$ | $0.189^{**}$ | $0.211^{*}$ | $0.101$ | $0.381$ |
| | (0.049) | (0.105) | (0.058) | (0.125) | (0.063) | (0.121) | (0.078) | (0.123) | (0.114) | (0.233) |
| (+5k) Recipient Dummy | $0.045^{***}$ | $0.044^{**}$ | $0.041^{***}$ | $0.041^{**}$ | $0.051^{***}$ | $0.045^{**}$ | $0.051^{***}$ | $0.028$ | $0.048^{**}$ | $0.097^{***}$ |
| | (0.010) | (0.019) | (0.010) | (0.018) | (0.012) | (0.020) | (0.014) | (0.024) | (0.019) | (0.037) |
| 2y thrift | $0.051^{***}$ | $0.073^{***}$ | | | $0.066^{***}$ | $0.082^{***}$ | | | $0.063^{***}$ | $0.100^{***}$ |
| | (0.008) | (0.016) | | | (0.011) | (0.019) | | | (0.016) | (0.032) |
| 3y thrift | | | $0.051^{***}$ | $0.079^{***}$ | | | $0.067^{***}$ | $0.084^{***}$ | | |
| | | | (0.010) | (0.020) | | | (0.014) | (0.028) | | |
| Married | $0.199^{***}$ | $0.186^{***}$ | $0.181^{***}$ | $0.116^{**}$ | $0.221^{***}$ | $0.241^{***}$ | $0.198^{***}$ | $0.212^{***}$ | $0.246^{***}$ | $0.245^{***}$ |
| | (0.021) | (0.037) | (0.025) | (0.047) | (0.017) | (0.031) | (0.020) | (0.036) | (0.019) | (0.033) |
| Female | $-0.074^{***}$ | $-0.061^{***}$ | $-0.075^{***}$ | $-0.063^{***}$ | $-0.105^{***}$ | $-0.083^{***}$ | $-0.112^{***}$ | $-0.090^{***}$ | $-0.134^{***}$ | $-0.107^{***}$ |
| | (0.004) | (0.008) | (0.004) | (0.009) | (0.005) | (0.010) | (0.006) | (0.012) | (0.008) | (0.015) |
| Belgian | $0.043^{***}$ | $0.022$ | $0.030^{*}$ | $0.017$ | $0.042^{**}$ | $0.037$ | $0.047^{**}$ | $0.042$ | $0.022$ | $0.022$ |
| | (0.014) | (0.021) | (0.016) | (0.024) | (0.017) | (0.026) | (0.019) | (0.030) | (0.023) | (0.038) |
| (+5k) Recipient Dummy :2y thrift | $-0.042^{**}$ | $-0.054$ | | | $-0.074^{***}$ | $-0.096^{***}$ | | | $-0.074^{**}$ | $-0.153^{**}$ |
| | (0.018) | (0.033) | | | (0.019) | (0.028) | | | (0.034) | (0.069) |
| (+5k) Recipient Dummy :3y thrift | | | $-0.038^{**}$ | $-0.047^{*}$ | | | $-0.064^{***}$ | $-0.043$ | | |
| | | | (0.018) | (0.028) | | | (0.023) | (0.035) | | |
| Married:Female | $-0.147^{***}$ | $-0.162^{***}$ | $-0.124^{***}$ | $-0.096^{*}$ | $-0.144^{***}$ | $-0.189^{***}$ | $-0.125^{***}$ | $-0.174^{***}$ | $-0.205^{***}$ | $-0.200^{***}$ |
| | (0.026) | (0.043) | (0.030) | (0.053) | (0.021) | (0.039) | (0.026) | (0.048) | (0.025) | (0.045) |
| Constant | $7.703^{***}$ | $7.081^{***}$ | $7.737^{***}$ | $6.323^{***}$ | $8.389^{***}$ | $8.253^{***}$ | $8.139^{***}$ | $7.778^{***}$ | $9.107^{***}$ | $6.485^{***}$ |
| | (0.468) | (0.997) | (0.556) | (1.189) | (0.600) | (1.154) | (0.750) | (1.182) | (1.079) | (2.205) |
| Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start year | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start income:Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | |
| Postal code | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ |
| F Statistic | $68.14^{***}$ | $5.8^{***}$ | $53.38^{***}$ | $4.73^{***}$ | $50.28^{***}$ | $4.74^{***}$ | $35.45^{***}$ | $3.83^{***}$ | $32.55^{***}$ | $8.3^{***}$ |
| Breusch-Pagan | 0 | 0.001 | 0 | 0.038 | 0 | 0.026 | 0 | 0.015 | 0 | 0.077 |
| Observations | 17,109 | 4,735 | 13,382 | 3,547 | 12,103 | 3,640 | 8,383 | 2,458 | 6,216 | 2,001 |
| $R^2$ | 0.157 | 0.185 | 0.155 | 0.181 | 0.148 | 0.178 | 0.144 | 0.181 | 0.167 | 0.204 |
| Adjusted $R^2$ | 0.155 | 0.155 | 0.153 | 0.144 | 0.145 | 0.141 | 0.140 | 0.132 | 0.162 | 0.147 |
| Residual Std. Error | 0.246 | 0.252 | 0.247 | 0.255 | 0.272 | 0.279 | 0.275 | 0.281 | 0.296 | 0.295 |
| | ($df = 17069$) | ($df = 4565$) | ($df = 13343$) | ($df = 3393$) | ($df = 12065$) | ($df = 3482$) | ($df = 8346$) | ($df = 2318$) | ($df = 6180$) | ($df = 1866$) |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 and robust standard errors in parentheses. ✓ indicates that controls are included.

## 3.E Sanity checks

For several of the personal characteristic controls, the results align well with known effects from literature, passing sanity checks. The negative coefficient for females and married females shows the documented gender gap in income growth [45, 46]. The positive coefficient for having a native nationality coincides with migration literature [47]. The start age proxies for attained education and shows rising income up till the age of 23-24. This is the age that someone who followed a model trajectory for a master's degree typically has in their first working year, as per the known college wage premium [48]. The region coefficients reflect the known higher youth unemployment in the Walloon region and lower youth unemployment in the Flemish region relative to the Brussels-Capital Region [49]. And lastly, the start year dummies confirm the negative effect of the financial crisis on the labour market [50, 51], which made it more difficult for starters to find a job [52–54], which in turn has a negative influence on labour market outcomes [55–57]. Examples of the coefficient estimates for the basic model (Eq.1 in main text) can be found in Table 3.E.1 and Table 3.E.2. The coefficients retain their sign and magnitude throughout the regressions, which can be provided on request.

| | *Dependent variable:* | | |
|---|---|---|---|
| | log(income after 3y + 1) | log(income after 5y + 1) | log(income after 7y + 1) |
| | (1) | (2) | (3) |
| log(start income + 1) | $0.279^{***}$ | $0.241^{***}$ | $0.212^{***}$ |
| | (0.009) | (0.011) | (0.016) |
| log(start wealth + 1) | $0.012^{***}$ | $0.013^{***}$ | $0.011^{***}$ |
| | (0.001) | (0.001) | (0.002) |
| Start age 18 | 0.010 | 0.035 | 0.038 |
| | (0.030) | (0.033) | (0.051) |
| Start age 20 | 0.004 | 0.008 | −0.014 |
| | (0.011) | (0.013) | (0.020) |
| Start age 21 | $0.022^{**}$ | 0.007 | −0.006 |
| | (0.011) | (0.013) | (0.019) |
| Start age 22 | $0.050^{***}$ | $0.050^{***}$ | $0.034^{*}$ |
| | (0.010) | (0.013) | (0.018) |
| Start age 23 | $0.098^{***}$ | $0.095^{***}$ | $0.076^{***}$ |
| | (0.010) | (0.013) | (0.018) |
| Start age 24 | $0.113^{***}$ | $0.110^{***}$ | $0.112^{***}$ |
| | (0.010) | (0.013) | (0.018) |
| Start age 25 | $0.126^{***}$ | $0.102^{***}$ | $0.078^{***}$ |
| | (0.011) | (0.013) | (0.020) |
| Start age 26 | $0.088^{***}$ | $0.081^{***}$ | 0.025 |
| | (0.012) | (0.015) | (0.023) |
| Start age 27 | $0.073^{***}$ | $0.056^{***}$ | $0.068^{**}$ |
| | (0.014) | (0.019) | (0.028) |
| Start age 28 | $0.089^{***}$ | $0.077^{***}$ | −0.024 |
| | (0.017) | (0.022) | (0.030) |
| Start age 29 | $0.078^{***}$ | 0.047 | −0.010 |
| | (0.023) | (0.029) | (0.038) |
| Start age 30 | $0.087^{***}$ | $0.064^{**}$ | $-0.057^{*}$ |
| | (0.025) | (0.026) | (0.034) |
| Start year 2010 | 0.006 | $0.021^{***}$ | −0.005 |
| | (0.006) | (0.007) | (0.007) |
| Start year 2011 | $0.016^{***}$ | $0.018^{***}$ | |
| | (0.006) | (0.007) | |
| Start year 2012 | $0.013^{**}$ | 0.011 | |
| | (0.006) | (0.007) | |
| Start year 2013 | $0.024^{***}$ | | |
| | (0.007) | | |
| Start year 2014 | $0.019^{***}$ | | |
| | (0.007) | | |
| Married | $0.200^{***}$ | $0.225^{***}$ | $0.249^{***}$ |
| | (0.021) | (0.017) | (0.019) |
| Female | $-0.076^{***}$ | $-0.106^{***}$ | $-0.135^{***}$ |
| | (0.004) | (0.005) | (0.008) |
| Belgian | $0.049^{***}$ | $0.048^{***}$ | 0.027 |
| | (0.013) | (0.016) | (0.023) |
| Region Flanders | $0.036^{***}$ | $0.024^{**}$ | $0.027^{*}$ |
| | (0.008) | (0.010) | (0.015) |
| Region Wallonia | 0.002 | −0.014 | −0.019 |
| | (0.008) | (0.010) | (0.014) |
| Married:Female | $-0.151^{***}$ | $-0.146^{***}$ | $-0.203^{***}$ |
| | (0.026) | (0.021) | (0.025) |
| Constant | $7.053^{***}$ | $7.516^{***}$ | $7.948^{***}$ |
| | (0.083) | (0.105) | (0.151) |
| F Statistic | $107.7^{***}$ | $82.56^{***}$ | $55.65^{***}$ |
| Breusch-Pagan test (p-val) | 0 | 0 | 0 |
| Observations | 17,576 | 12,573 | 6,682 |
| $R^2$ | 0.155 | 0.147 | 0.161 |
| Adjusted $R^2$ | 0.154 | 0.145 | 0.158 |
| Residual Std. Error | $0.246(df = 17550)$ | $0.271(df = 12549)$ | $0.295(df = 6660)$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 and robust standard errors in parentheses. | | |

*Table 3.E.1: The effect of starting wealth on income performance three years (column 1), five years (column 2), and seven years (column 3) into a career using Ordinary Least Squares (OLS) regressions. The columns within each horizon have different levels of location controls. For start year, 2009 is the reference year. For start age, 19 is the reference age. For region, Brussels is the reference region.*

*Table 3.E.2: The effect of starting wealth on income performance five years into a career with the income growth trim (columns 1–3), the separate income trim (columns 4–6), and no trim (columns 7–9) using Ordinary Least Squares (OLS) regressions. The columns within each trim have different levels of location controls.*

| | *Dependent variable:* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | log(income after 3y + 1) | | | log(income after 5y + 1) | | | log(income after 7y + 1) | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| log(start income + 1) | 0.225*** | 0.230*** | 0.061 | 0.097* | 0.067 | −0.107 | 0.124* | 0.316** | 0.069 |
| | (0.041) | (0.088) | (0.208) | (0.054) | (0.098) | (0.246) | (0.068) | (0.151) | (0.275) |
| log(start wealth + 1) | 0.012*** | 0.013*** | 0.016*** | 0.013*** | 0.014*** | 0.018*** | 0.011*** | 0.015*** | −0.004 |
| | (0.001) | (0.002) | (0.004) | (0.001) | (0.003) | (0.005) | (0.002) | (0.003) | (0.009) |
| Start year 2010 | 0.005 | 0.010 | −0.042 | 0.019*** | 0.011 | 0.040 | −0.005 | −0.005 | 0.160* |
| | (0.006) | (0.012) | (0.048) | (0.006) | (0.014) | (0.057) | (0.007) | (0.016) | (0.086) |
| Start year 2011 | 0.016*** | 0.031** | 0.095* | 0.018*** | 0.022 | 0.142** | | | |
| | (0.006) | (0.012) | (0.055) | (0.007) | (0.015) | (0.071) | | | |
| Start year 2012 | 0.011* | 0.011 | 0.004 | 0.010 | 0.001 | −0.011 | | | |
| | (0.006) | (0.014) | (0.054) | (0.007) | (0.015) | (0.071) | | | |
| Start year 2013 | 0.024*** | 0.032** | 0.125 | | | | | | |
| | (0.007) | (0.016) | (0.085) | | | | | | |
| Start year 2014 | 0.018*** | 0.018 | −0.023 | | | | | | |
| | (0.007) | (0.015) | (0.061) | | | | | | |
| Married | 0.193*** | 0.185*** | 0.413*** | 0.220*** | 0.239*** | 0.200*** | 0.247*** | 0.245*** | 0.244*** |
| | (0.021) | (0.037) | (0.074) | (0.017) | (0.031) | (0.051) | (0.019) | (0.032) | (0.065) |
| Female | −0.077*** | −0.064*** | −0.043*** | −0.108*** | −0.086*** | −0.065*** | −0.136*** | −0.109*** | −0.075** |
| | (0.004) | (0.008) | (0.016) | (0.005) | (0.010) | (0.021) | (0.008) | (0.014) | (0.035) |
| Belgian | 0.044*** | 0.026 | −0.024 | 0.043*** | 0.038 | −0.008 | 0.023 | 0.026 | 0.018 |
| | (0.013) | (0.020) | (0.051) | (0.016) | (0.025) | (0.049) | (0.023) | (0.037) | (0.087) |
| Region Flanders | 0.035*** | | | 0.023** | | | 0.027* | | |
| | (0.008) | | | (0.010) | | | (0.014) | | |
| Region Wallonia | 0.001 | | | −0.015 | | | −0.019 | | |
| | (0.008) | | | (0.010) | | | (0.014) | | |
| Married:Female | −0.140*** | −0.156*** | −0.382*** | −0.139*** | −0.181*** | −0.148** | −0.202*** | −0.187*** | −0.308*** |
| | (0.026) | (0.043) | (0.088) | (0.021) | (0.039) | (0.075) | (0.025) | (0.044) | (0.094) |
| Constant | 7.575*** | 7.526*** | 9.269*** | 8.896*** | 9.085*** | 10.884*** | 8.794*** | 6.955*** | 9.242*** |
| | (0.395) | (0.836) | (1.984) | (0.515) | (0.932) | (2.351) | (0.651) | (1.436) | (2.618) |
| Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start income:Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Postal code | | ✓ | | | ✓ | | | ✓ | |
| Neighborhood code | | | ✓ | | | ✓ | | | ✓ |
| F Statistic | 76.73*** | 6*** | 5.37*** | 57.34*** | 5.13*** | 4.26*** | 36.55*** | 4.32*** | 7.91*** |
| Breusch-Pagan | 0 | 0.003 | 0 | 0 | 0.033 | 0 | 0 | 0.079 | 0 |
| Observations | 17,576 | 4,856 | 1,416 | 12,573 | 3,761 | 1,090 | 6,682 | 2,121 | 609 |
| $R^2$ | 0.161 | 0.187 | 0.435 | 0.153 | 0.183 | 0.377 | 0.166 | 0.199 | 0.400 |
| Adjusted $R^2$ | 0.159 | 0.158 | 0.195 | 0.150 | 0.148 | 0.102 | 0.161 | 0.146 | 0.086 |
| Residual Std. Error | 0.245 | 0.251 | 0.239 | 0.270 | 0.276 | 0.269 | 0.295 | 0.295 | 0.304 |
| | ($df$ = 17538) | ($df$ = 4688) | ($df$ = 994) | ($df$ = 12537) | ($df$ = 3605) | ($df$ = 755) | ($df$ = 6648) | ($df$ = 1988) | ($df$ = 399) |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 and robust standard errors in parentheses. ✓indicates that controls are included.

# 3.F Robustness checks

Here we check the robustness of our results against some of the choices that were made in the main text. We check different income trimmings and income definitions. We always show the furthest horizon available. We find that our results hold throughout.

## 3.F.1 Trimming

Here we show that outliers in the data are not driving the observed effect. We compare the results of three different trimming schemes:

- Income growth trim (used in main text): here we jointly exclude the individuals in the top 1% of wealth and those in the top 1% of income growth after the specified horizon.
- Separate income trim: here we jointly exclude the individuals in the top 1% of wealth, the top 1% of start income, and the top 1% of income after the specified horizon.
- No trim.

The results can be found in Table 3.F.1-3.F.4. We always show the furthest horizon available (seven years). Where neighbourhood controls are used however, we show the five year horizon. We do this for the same reason as given in the main text, namely sample size.

*Table 3.F.1: The effect of starting wealth on income performance seven years into a career with the income growth trim (columns 1–2), the separate income trim (columns 3–4), and no trim (columns 5–6) using Ordinary Least Squares (OLS) regressions. The columns within each trim have different levels of location controls.*

| | *Dependent variable: log(income after 7y + 1)* | | | | | |
|---|---|---|---|---|---|---|
| | Income growth trim | | Separate income trim | | No trim | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| log(start income + 1) | 0.124* | 0.316** | 0.114* | 0.289* | 0.086 | 0.289* |
| | (0.068) | (0.151) | (0.069) | (0.154) | (0.071) | (0.152) |
| log(start wealth + 1) | 0.011*** | 0.015*** | 0.012*** | 0.017*** | 0.013*** | 0.016*** |
| | (0.002) | (0.003) | (0.002) | (0.004) | (0.002) | (0.004) |
| Married | 0.247*** | 0.245*** | 0.232*** | 0.238*** | 0.260*** | 0.281*** |
| | (0.019) | (0.032) | (0.019) | (0.032) | (0.020) | (0.034) |
| Female | −0.136*** | −0.109*** | −0.135*** | −0.108*** | −0.146*** | −0.112*** |
| | (0.008) | (0.014) | (0.008) | (0.014) | (0.008) | (0.014) |
| Belgian | 0.023 | 0.026 | 0.026 | 0.016 | 0.003 | 0.012 |
| | (0.023) | (0.037) | (0.023) | (0.038) | (0.027) | (0.042) |
| Married:Female | −0.202*** | −0.187*** | −0.185*** | −0.169*** | −0.218*** | −0.223*** |
| | (0.025) | (0.044) | (0.025) | (0.045) | (0.025) | (0.046) |
| Constant | 8.794*** | 6.955*** | 8.887*** | 7.224*** | 9.168*** | 7.228*** |
| | (0.651) | (1.436) | (0.655) | (1.464) | (0.683) | (1.447) |
| Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start year | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start income:Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | ✓ | | ✓ | |
| Postal code | | ✓ | | ✓ | | ✓ |
| F Statistic | 36.55*** | 4.32*** | 30.67*** | 4.12*** | 35.91*** | 4.39*** |
| Breusch-Pagan | 0 | 0.079 | 0 | 0.073 | 0 | 0.067 |
| Observations | 6,682 | 2,121 | 6,622 | 2,105 | 6,816 | 2,197 |
| $R^2$ | 0.166 | 0.199 | 0.139 | 0.185 | 0.161 | 0.200 |
| Adjusted $R^2$ | 0.161 | 0.146 | 0.135 | 0.131 | 0.157 | 0.147 |
| Residual Std. Error | 0.295 | 0.295 | 0.295 | 0.297 | 0.307 | 0.308 |
| | ($df = 6648$) | ($df = 1988$) | ($df = 6588$) | ($df = 1972$) | ($df = 6782$) | ($df = 2061$) |

*Note:* *p<0.1; **p<0.05; ***p<0.01 and standard errors in parentheses. ✓indicates that controls are included.

*Table 3.F.2: The effect of starting wealth on income performance five years into a career with the income growth trim (columns 1–3), the separate income trim (columns 4–6), and no trim (columns 7–9) using Ordinary Least Squares (OLS) regressions. The columns within each trim have different levels of location controls.*

| | *Dependent variable: log(income after 5y + 1)* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Income growth trim | | | separate income trim | | | No trim | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| log(start income + 1) | 0.097* | 0.067 | −0.107 | 0.085 | 0.065 | −0.067 | 0.080 | 0.056 | −0.110 |
| | (0.054) | (0.098) | (0.246) | (0.054) | (0.098) | (0.244) | (0.055) | (0.098) | (0.245) |
| log(start wealth + 1) | 0.013*** | 0.014*** | 0.018*** | 0.014*** | 0.015*** | 0.019*** | 0.014*** | 0.014*** | 0.016*** |
| | (0.001) | (0.003) | (0.005) | (0.001) | (0.003) | (0.005) | (0.001) | (0.002) | (0.005) |
| Married | 0.220*** | 0.239*** | 0.200*** | 0.210*** | 0.225*** | 0.178*** | 0.248*** | 0.250*** | 0.197*** |
| | (0.017) | (0.031) | (0.051) | (0.017) | (0.030) | (0.052) | (0.018) | (0.030) | (0.049) |
| Female | −0.108*** | −0.086*** | −0.065*** | −0.107*** | −0.083*** | −0.066*** | −0.114*** | −0.089*** | −0.073*** |
| | (0.005) | (0.010) | (0.021) | (0.005) | (0.010) | (0.021) | (0.005) | (0.010) | (0.021) |
| Belgian | 0.043*** | 0.038 | −0.008 | 0.044*** | 0.046* | −0.010 | 0.037** | 0.036 | −0.009 |
| | (0.016) | (0.025) | (0.049) | (0.016) | (0.025) | (0.052) | (0.018) | (0.026) | (0.050) |
| Married:Female | −0.139*** | −0.181*** | −0.148** | −0.132*** | −0.165*** | −0.113 | −0.162*** | −0.170*** | −0.141* |
| | (0.021) | (0.039) | (0.075) | (0.021) | (0.039) | (0.077) | (0.022) | (0.041) | (0.073) |
| Constant | 8.896*** | 9.085*** | 10.884*** | 9.013*** | 9.088*** | 10.439*** | 9.056*** | 9.194*** | 10.922*** |
| | (0.515) | (0.932) | (2.351) | (0.515) | (0.929) | (2.334) | (0.522) | (0.932) | (2.344) |
| Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start year | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start income:Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | | ✓ | | | ✓ | | |
| Postal code | | ✓ | | | ✓ | | | ✓ | |
| Neighbourhood code | | | ✓ | | | ✓ | | | ✓ |
| F Statistic | 57.34*** | 5.13*** | 4.26*** | 46.81*** | 4.51*** | 3.29*** | 56.43*** | 5.1*** | 4.21*** |
| Breusch-Pagan | 0 | 0.033 | 0 | 0 | 0.03 | 0 | 0 | 0.019 | 0 |
| Observations | 12,573 | 3,761 | 1,090 | 12,467 | 3,635 | 1,060 | 12,828 | 3,871 | 1,123 |
| $R^2$ | 0.153 | 0.183 | 0.377 | 0.121 | 0.161 | 0.376 | 0.147 | 0.176 | 0.374 |
| Adjusted $R^2$ | 0.150 | 0.148 | 0.102 | 0.119 | 0.123 | 0.097 | 0.145 | 0.141 | 0.098 |
| Residual Std. Error | 0.270 | 0.276 | 0.269 | 0.269 | 0.273 | 0.263 | 0.281 | 0.286 | 0.273 |
| | ($df = 12537$) | ($df = 3605$) | ($df = 755$) | ($df = 12431$) | ($df = 3480$) | ($df = 731$) | ($df = 12792$) | ($df = 3712$) | ($df = 778$) |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 and standard errors in parentheses. ✓indicates that controls are included.

*Table 3.F.3: The effect of (+€5,000) transferred financial wealth on income performance seven years into a career with the income growth trim (columns 1–4), the separate income trim (columns 5–8), and no trim (columns 9–12) using OLS regressions. The effect is estimated in two ways: first, in the **IV columns** by instrumenting start wealth with three dummies indicating the presence of different kinds of financial wealth transfers (+€5k financial boost, €5-10k cash boost, and +€10k cash boost); secondly, in the **Dummy columns** by combining the aforementioned dummies into one Recipient dummy indicating at least one kind of received (+€5k) financial wealth transfer and replacing the start wealth parameter by it. The columns within each estimation model have different levels of location controls.*

| | *Dependent variable: log(income after 7y + 1)* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Income growth trim | | | | separate income trim | | | | No trim | | | |
| | *IV* | *IV* | *Dummy* | *Dummy* | *IV* | *IV* | *Dummy* | *Dummy* | *IV* | *IV* | *Dummy* | *Dummy* |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| log(start income + 1) | 0.124* | 0.301** | 0.133* | 0.336** | 0.112 | 0.268* | 0.124* | 0.310* | 0.086 | 0.273* | 0.096 | 0.309** |
| | (0.068) | (0.149) | (0.069) | (0.156) | (0.069) | (0.150) | (0.070) | (0.159) | (0.072) | (0.149) | (0.073) | (0.157) |
| log(start wealth + 1) | 0.011* | 0.026** | | | 0.014** | 0.033*** | | | 0.013** | 0.027*** | | |
| | (0.006) | (0.011) | | | (0.006) | (0.011) | | | (0.005) | (0.010) | | |
| (+5k) Recipient | | | 0.025** | 0.059*** | | | 0.030*** | 0.067*** | | | 0.030*** | 0.063*** |
| | | | (0.011) | (0.022) | | | (0.011) | (0.022) | | | (0.011) | (0.021) |
| Married | 0.247*** | 0.239*** | 0.251*** | 0.252*** | 0.231*** | 0.228*** | 0.236*** | 0.246*** | 0.260*** | 0.276*** | 0.264*** | 0.287*** |
| | (0.019) | (0.033) | (0.019) | (0.032) | (0.019) | (0.033) | (0.019) | (0.032) | (0.020) | (0.035) | (0.020) | (0.034) |
| Female | −0.136*** | −0.111*** | −0.134*** | −0.105*** | −0.135*** | −0.111*** | −0.133*** | −0.104*** | −0.146*** | −0.114*** | −0.144*** | −0.109*** |
| | (0.008) | (0.014) | (0.008) | (0.014) | (0.008) | (0.015) | (0.008) | (0.014) | (0.008) | (0.015) | (0.008) | (0.014) |
| Belgian | 0.023 | 0.027 | 0.024 | 0.023 | 0.025 | 0.018 | 0.026 | 0.012 | 0.003 | 0.012 | 0.005 | 0.008 |
| | (0.023) | (0.037) | (0.023) | (0.037) | (0.023) | (0.038) | (0.023) | (0.038) | (0.027) | (0.042) | (0.027) | (0.042) |
| Married:Female | −0.202*** | −0.180*** | −0.207*** | −0.195*** | −0.185*** | −0.159*** | −0.191*** | −0.179*** | −0.218*** | −0.219*** | −0.223*** | −0.230*** |
| | (0.025) | (0.044) | (0.025) | (0.044) | (0.025) | (0.045) | (0.025) | (0.044) | (0.025) | (0.047) | (0.025) | (0.046) |
| Constant | 8.793*** | 7.026*** | 8.772*** | 6.875*** | 8.894*** | 7.323*** | 8.867*** | 7.143*** | 9.169*** | 7.303*** | 9.156*** | 7.142*** |
| | (0.651) | (1.407) | (0.663) | (1.485) | (0.654) | (1.423) | (0.668) | (1.515) | (0.683) | (1.418) | (0.694) | (1.499) |
| Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start year | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start income:Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | |
| Postal code | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ |
| Observations | 6,682 | 2,121 | 6,682 | 2,121 | 6,622 | 2,105 | 6,622 | 2,105 | 6,816 | 2,197 | 6,816 | 2,197 |
| Breusch-Pagan | 0 | 0.079 | 0 | 0.06 | 0 | 0.073 | 0 | 0.084 | 0 | 0.067 | 0 | 0.054 |
| $R^2$ | 0.166 | 0.195 | 0.162 | 0.195 | 0.139 | 0.177 | 0.135 | 0.180 | 0.161 | 0.196 | 0.157 | 0.195 |
| Adjusted $R^2$ | 0.161 | 0.142 | 0.158 | 0.142 | 0.134 | 0.122 | 0.131 | 0.125 | 0.157 | 0.144 | 0.152 | 0.142 |
| Weak instruments | 0 | 0 | | | 0 | 0 | | | 0 | 0 | | |
| Wu-Hausman | 0.94 | 0.31 | | | 0.68 | 0.13 | | | 0.96 | 0.25 | | |
| Sargan | 0.13 | 0.1 | | | 0.21 | 0.1 | | | 0.4 | 0.22 | | |
| Residual Std. Error | 0.295 | 0.296 | 0.295 | 0.296 | 0.295 | 0.298 | 0.296 | 0.298 | 0.307 | 0.309 | 0.308 | 0.309 |
| | ($df$ = 6648) | ($df$ = 1988) | ($df$ = 6648) | ($df$ = 1988) | ($df$ = 6588) | ($df$ = 1972) | ($df$ = 6588) | ($df$ = 1972) | ($df$ = 6782) | ($df$ = 2061) | ($df$ = 6782) | ($df$ = 2061) |
| F Statistic | | | 36*** | 4.27*** | | | 29.86*** | 3.94*** | | | 34.98*** | 4.21*** |

*Note:* *p<0.1; **p<0.05; ***p<0.01 and robust standard errors in parentheses. ✓ indicates that controls are included.

*Table 3.F.4: The effect of (+€5,000) transferred financial wealth and thrift on income performance seven years into a career with the income growth trim (columns 1–2), the separate income trim (columns 3–4), and no trim (columns 5–6) using OLS regressions. Thrift is measured over the last two ($2y\ thrift$) non-working years. The columns within each estimation model have different levels of location controls.*

| | *Dependent variable: log(income after 7y + 1)* | | | | | |
|---|---|---|---|---|---|---|
| | Income growth trim | | Separate income trim | | No trim | |
| | $2y\ thrift$ | $2y\ thrift$ | $2y\ thrift$ | $2y\ thrift$ | $2y\ thrift$ | $2y\ thrift$ |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| log(start income + 1) | 0.101 | 0.381 | 0.081 | 0.318 | 0.073 | 0.323 |
| | (0.114) | (0.233) | (0.115) | (0.238) | (0.113) | (0.235) |
| (+5k) Recipient | $0.048^{**}$ | $0.097^{***}$ | $0.054^{***}$ | $0.100^{***}$ | $0.058^{***}$ | $0.100^{***}$ |
| | (0.019) | (0.037) | (0.019) | (0.038) | (0.020) | (0.038) |
| 2y thrift | $0.063^{***}$ | $0.100^{***}$ | $0.065^{***}$ | $0.104^{***}$ | $0.074^{***}$ | $0.109^{***}$ |
| | (0.016) | (0.032) | (0.017) | (0.033) | (0.017) | (0.033) |
| Married | $0.246^{***}$ | $0.245^{***}$ | $0.231^{***}$ | $0.240^{***}$ | $0.260^{***}$ | $0.282^{***}$ |
| | (0.019) | (0.033) | (0.020) | (0.033) | (0.020) | (0.035) |
| Female | $-0.134^{***}$ | $-0.107^{***}$ | $-0.134^{***}$ | $-0.106^{***}$ | $-0.145^{***}$ | $-0.111^{***}$ |
| | (0.008) | (0.015) | (0.008) | (0.015) | (0.008) | (0.015) |
| Belgian | 0.022 | 0.022 | 0.024 | 0.012 | 0.001 | 0.006 |
| | (0.023) | (0.038) | (0.023) | (0.038) | (0.028) | (0.043) |
| (+5k) Recipient :2y thrift | $-0.074^{**}$ | $-0.153^{**}$ | $-0.076^{**}$ | $-0.139^{**}$ | $-0.089^{***}$ | $-0.155^{**}$ |
| | (0.034) | (0.069) | (0.034) | (0.070) | (0.035) | (0.071) |
| Married:Female | $-0.205^{***}$ | $-0.200^{***}$ | $-0.189^{***}$ | $-0.180^{***}$ | $-0.222^{***}$ | $-0.236^{***}$ |
| | (0.025) | (0.045) | (0.025) | (0.045) | (0.026) | (0.047) |
| Constant | $9.107^{***}$ | $6.485^{***}$ | $9.298^{***}$ | $7.098^{***}$ | $9.391^{***}$ | $7.049^{***}$ |
| | (1.079) | (2.205) | (1.091) | (2.248) | (1.075) | (2.222) |
| Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start year | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start income:Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | ✓ | | ✓ | |
| Postal code | | ✓ | | ✓ | | ✓ |
| F Statistic | $32.55^{***}$ | $8.3^{***}$ | $27.16^{***}$ | $8.1^{***}$ | $31.94^{***}$ | $4.81^{***}$ |
| Breusch-Pagan | 0 | 0.077 | 0 | 0.075 | 0 | 0.047 |
| Observations | 6,216 | 2,001 | 6,154 | 1,989 | 6,343 | 2,072 |
| $R^2$ | 0.167 | 0.204 | 0.139 | 0.189 | 0.162 | 0.203 |
| Adjusted $R^2$ | 0.162 | 0.147 | 0.134 | 0.130 | 0.157 | 0.147 |
| Residual Std. Error | 0.296 | 0.295 | 0.296 | 0.298 | 0.309 | 0.310 |
| | ($df = 6180$) | ($df = 1866$) | ($df = 6118$) | ($df = 1854$) | ($df = 6307$) | ($df = 1934$) |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 and robust standard errors in parentheses. ✓indicates that controls are included.

### 3.F.2 Income definition

Here we show robustness of the results against the income definition. We compare the results for three different income definitions:

- supernarrow income (used in main text) = electronic transfers from €2,000-5,000 + electronic transfers from €1,000-2,000 + electronic transfers from €0-1,000
- narrow income = electronic transfers from €5,000-10,000 + electronic transfers from €2,000-5,000 + electronic transfers from €1,000-2,000 + electronic transfers from €0-1,000
- narrowplus income = cash deposits + deposits via cheques + electronic transfers from €5,000-10,000 + electronic transfers from €2,000-5,000 + electronic transfers from €1,000-2,000 + electronic transfers from €0-1,000

Because of the influence between the income definition, the starters detection, and the *Recipient* dummy variable, we have to vary both the income definition and the *Recipient* threshold simultaneously.

For instance, in the main text, the *Recipient* dummy variable was based on dummy variables signifying exogenous wealth transfers of €$5,000$ or more (*(+5k) Recipient*). With the *supernarrow* definition, income contains electronic transfers up to €$5,000$. This way there is no overlap between what is labeled income, and what categorizes a person as being a *Recipient*. However, if then only the income definition is changed to narrow or narrowplus, which include up to €$10,000$ electronic transfers, a transfer of €$5,000$ before the start of a career is also counted towards income. This only leaves €$1,000$ till the individual gets excluded from the career starters due to reaching the €$6,000$ euro income limit under which they need to fall pre-career. It thus makes sense to only use exogenous wealth transfers of €$10,000$ or more to construct the *Recipient* dummy (*(+10k) Recipient*) when using the narrow or narrowplus income definition. This high requirement of receiving over €$10,000$ in financial support decreases the overall sample of recipients by more than half for the five year horizon. Even with this decrease in sample size, the basic results, shown in Table 3.F.5-3.F.7 are in line with those of the main text.

*Table 3.F.5: The effect of starting wealth on income performance seven years into a career for the Supernarrow (columns 1–2), Narrow (columns 3–4), and Narrowplus (columns 5–6) income definition using Ordinary Least Squares (OLS) regressions. The columns within each horizon have different levels of location controls.*

| | *Dependent variable: log(income after 7y + 1)* | | | | | |
|---|---|---|---|---|---|---|
| | Supernarrow income | | Narrow income | | Narrowplus income | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| log(start income + 1) | $0.124^{*}$ | $0.316^{**}$ | 0.109 | $0.282^{**}$ | 0.076 | $0.296^{**}$ |
| | (0.068) | (0.151) | (0.067) | (0.137) | (0.072) | (0.145) |
| log(start wealth + 1) | $0.011^{***}$ | $0.015^{***}$ | $0.012^{***}$ | $0.020^{***}$ | $0.014^{***}$ | $0.025^{***}$ |
| | (0.002) | (0.003) | (0.002) | (0.004) | (0.002) | (0.004) |
| Married | $0.247^{***}$ | $0.245^{***}$ | $0.256^{***}$ | $0.277^{***}$ | $0.277^{***}$ | $0.292^{***}$ |
| | (0.019) | (0.032) | (0.020) | (0.032) | (0.022) | (0.041) |
| Female | $-0.136^{***}$ | $-0.109^{***}$ | $-0.136^{***}$ | $-0.116^{***}$ | $-0.135^{***}$ | $-0.102^{***}$ |
| | (0.008) | (0.014) | (0.008) | (0.015) | (0.009) | (0.018) |
| Belgian | 0.023 | 0.026 | 0.007 | −0.001 | −0.015 | −0.022 |
| | (0.023) | (0.037) | (0.024) | (0.039) | (0.027) | (0.052) |
| Married:Female | $-0.202^{***}$ | $-0.187^{***}$ | $-0.201^{***}$ | $-0.221^{***}$ | $-0.222^{***}$ | $-0.232^{***}$ |
| | (0.025) | (0.044) | (0.026) | (0.046) | (0.029) | (0.055) |
| Constant | $8.794^{***}$ | $6.955^{***}$ | $8.981^{***}$ | $7.335^{***}$ | $9.363^{***}$ | $7.200^{***}$ |
| | (0.651) | (1.436) | (0.639) | (1.313) | (0.688) | (1.387) |
| Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start year | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start income:Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | ✓ | | ✓ | |
| Postal code | | ✓ | | ✓ | | ✓ |
| F Statistic | $36.55^{***}$ | $4.32^{***}$ | $34.19^{***}$ | $4.67^{***}$ | $27.79^{***}$ | $4.06^{***}$ |
| Breusch-Pagan | 0 | 0.079 | 0 | 0.06 | 0 | 0.479 |
| Observations | 6,682 | 2,121 | 6,624 | 2,068 | 5,985 | 1,618 |
| $R^2$ | 0.166 | 0.199 | 0.153 | 0.207 | 0.143 | 0.202 |
| Adjusted $R^2$ | 0.161 | 0.146 | 0.149 | 0.154 | 0.138 | 0.143 |
| Residual Std. Error | 0.295 | 0.295 | 0.316 | 0.317 | 0.322 | 0.325 |
| | ($df = 6648$) | ($df = 1988$) | ($df = 6590$) | ($df = 1936$) | ($df = 5951$) | ($df = 1507$) |

*Note:* $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ and standard errors in parentheses. ✓indicates that controls are included.

*Table 3.F.6: The effect of starting wealth on income performance five years into a career for the Supernarrow (columns 1–3), Narrow (columns 4–6), and Narrowplus (columns 7–9) income definition using Ordinary Least Squares (OLS) regressions. The columns within each horizon have different levels of location controls.*

| | *Dependent variable: log(income after 5y + 1)* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Supernarrow income | | | Narrow income | | | Narrowplus income | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| log(start income + 1) | 0.097* | 0.067 | −0.107 | 0.061 | 0.160 | −0.321 | 0.122** | 0.107 | 0.081 |
| | (0.054) | (0.098) | (0.246) | (0.053) | (0.098) | (0.287) | (0.050) | (0.104) | (0.248) |
| log(start wealth + 1) | 0.013*** | 0.014*** | 0.018*** | 0.015*** | 0.018*** | 0.013** | 0.014*** | 0.018*** | 0.017** |
| | (0.001) | (0.003) | (0.005) | (0.001) | (0.003) | (0.006) | (0.002) | (0.003) | (0.007) |
| Married | 0.220*** | 0.239*** | 0.200*** | 0.225*** | 0.252*** | 0.192*** | 0.251*** | 0.250*** | 0.188*** |
| | (0.017) | (0.031) | (0.051) | (0.018) | (0.034) | (0.052) | (0.020) | (0.039) | (0.067) |
| Female | −0.108*** | −0.086*** | −0.065*** | −0.110*** | −0.088*** | −0.091*** | −0.106*** | −0.088*** | −0.143*** |
| | (0.005) | (0.010) | (0.021) | (0.005) | (0.010) | (0.025) | (0.006) | (0.013) | (0.027) |
| Belgian | 0.043*** | 0.038 | −0.008 | 0.035* | 0.026 | −0.050 | 0.055*** | 0.030 | −0.012 |
| | (0.016) | (0.025) | (0.049) | (0.018) | (0.029) | (0.060) | (0.021) | (0.035) | (0.079) |
| Married:Female | −0.139*** | −0.181*** | −0.148** | −0.155*** | −0.207*** | −0.057 | −0.186*** | −0.209*** | −0.066 |
| | (0.021) | (0.039) | (0.075) | (0.023) | (0.044) | (0.083) | (0.026) | (0.052) | (0.100) |
| Constant | 8.896*** | 9.085*** | 10.884*** | 9.274*** | 8.245*** | 13.396*** | 8.705*** | 8.753*** | 9.328*** |
| | (0.515) | (0.932) | (2.351) | (0.512) | (0.937) | (2.734) | (0.477) | (0.997) | (2.401) |
| Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start year | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start income:Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | | ✓ | | | ✓ | | |
| Postal code | | ✓ | | | ✓ | | | ✓ | |
| Neighbourhood code | | | ✓ | | | ✓ | | | ✓ |
| F Statistic | 57.34*** | 5.13*** | 4.26*** | 50.32*** | 5.22*** | 4.19*** | 39.25*** | 4.35*** | 4.27*** |
| Breusch-Pagan | 0 | 0.033 | 0 | 0 | 0.002 | 0 | 0 | 0.003 | 0 |
| Observations | 12,573 | 3,761 | 1,090 | 12,320 | 3,579 | 970 | 11,146 | 2,738 | 767 |
| $R^2$ | 0.153 | 0.183 | 0.377 | 0.140 | 0.177 | 0.392 | 0.124 | 0.171 | 0.404 |
| Adjusted $R^2$ | 0.150 | 0.148 | 0.102 | 0.138 | 0.141 | 0.109 | 0.121 | 0.130 | 0.106 |
| Residual Std. Error | 0.270 | 0.276 | 0.269 | 0.289 | 0.295 | 0.289 | 0.299 | 0.305 | 0.292 |
| | ($df$ = 12537) | ($df$ = 3605) | ($df$ = 755) | ($df$ = 12284) | ($df$ = 3427) | ($df$ = 661) | ($df$ = 11110) | ($df$ = 2610) | ($df$ = 510) |

*Note:* $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ and standard errors in parentheses. ✓ indicates that controls are included.

*Table 3.F.7: The effect of (+€10,000) transferred financial wealth on income performance seven years into a career for the Supernarrow (columns 1–4), Narrow (columns 5–8), and Narrowplus (columns 9–12) income definition using OLS regressions. The effect is estimated in two ways: first, in the* ***IV columns*** *by instrumenting start wealth with two dummies indicating the presence of different kinds of financial wealth transfers (+€10k financial boost, and +€10k cash boost); secondly, in the* ***Dummy columns*** *by combining the aforementioned dummies into one Recipient dummy indicating at least one kind of received (+€10k) financial wealth transfer and replacing the start wealth parameter by it. The columns within each estimation model have different levels of location controls.*

| | *Dependent variable: log(income after 3y + 1)* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Supernarrow income | | | | Narrow income | | | | Narrowplus income | | | |
| | *IV* | *IV* | *Dummy* | *Dummy* | *IV* | *IV* | *Dummy* | *Dummy* | *IV* | *IV* | *Dummy* | *Dummy* |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| log(start income + 1) | 0.219*** | 0.226** | 0.228*** | 0.232*** | 0.136*** | 0.178** | 0.143*** | 0.168** | 0.176*** | 0.235** | 0.176*** | 0.223** |
| | (0.041) | (0.089) | (0.042) | (0.087) | (0.041) | (0.079) | (0.040) | (0.077) | (0.042) | (0.096) | (0.042) | (0.094) |
| log(start wealth + 1) | 0.022*** | 0.031*** | | | 0.021*** | 0.024** | | | 0.013*** | 0.026** | | |
| | (0.004) | (0.008) | | | (0.005) | (0.010) | | | (0.005) | (0.011) | | |
| (+10k) Recipient | | | 0.048*** | 0.058*** | | | 0.048*** | 0.047** | | | 0.032*** | 0.059** |
| | | | (0.008) | (0.017) | | | (0.009) | (0.021) | | | (0.011) | (0.027) |
| Married | 0.192*** | 0.177*** | 0.195*** | 0.188*** | 0.191*** | 0.220*** | 0.192*** | 0.226*** | 0.252*** | 0.258*** | 0.253*** | 0.266*** |
| | (0.021) | (0.037) | (0.021) | (0.037) | (0.023) | (0.045) | (0.023) | (0.045) | (0.028) | (0.055) | (0.028) | (0.055) |
| Female | −0.079*** | −0.068*** | −0.075*** | −0.061*** | −0.086*** | −0.072*** | −0.082*** | −0.067*** | −0.081*** | −0.064*** | −0.079*** | −0.059*** |
| | (0.004) | (0.008) | (0.004) | (0.008) | (0.004) | (0.009) | (0.004) | (0.008) | (0.005) | (0.010) | (0.005) | (0.010) |
| Belgian | 0.041*** | 0.022 | 0.045*** | 0.026 | 0.043*** | 0.027 | 0.047*** | 0.028 | 0.027 | 0.011 | 0.031* | 0.011 |
| | (0.014) | (0.021) | (0.013) | (0.020) | (0.015) | (0.025) | (0.015) | (0.025) | (0.018) | (0.031) | (0.018) | (0.031) |
| Married:Female | −0.136*** | −0.144*** | −0.145*** | −0.163*** | −0.118*** | −0.205*** | −0.124*** | −0.214*** | −0.146*** | −0.242*** | −0.151*** | −0.257*** |
| | (0.026) | (0.043) | (0.026) | (0.043) | (0.028) | (0.051) | (0.028) | (0.051) | (0.034) | (0.064) | (0.034) | (0.064) |
| Constant | 7.565*** | 7.442*** | 7.621*** | 7.586*** | 8.387*** | 7.940*** | 8.451*** | 8.186*** | 8.101*** | 7.417*** | 8.180*** | 7.689*** |
| | (0.394) | (0.851) | (0.397) | (0.833) | (0.389) | (0.761) | (0.387) | (0.735) | (0.404) | (0.926) | (0.401) | (0.893) |
| Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start year | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start income:Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | |
| Postal code | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ |
| Observations | 17,576 | 4,856 | 17,576 | 4,856 | 17,138 | 4,582 | 17,138 | 4,582 | 15,447 | 3,358 | 15,447 | 3,358 |
| Breusch-Pagan | 0 | 0.003 | 0 | 0.003 | 0 | 0 | 0 | 0 | 0 | 0.041 | 0 | 0.037 |
| $R^2$ | 0.156 | 0.173 | 0.156 | 0.182 | 0.146 | 0.172 | 0.145 | 0.172 | 0.138 | 0.180 | 0.132 | 0.180 |
| Adjusted $R^2$ | 0.154 | 0.143 | 0.154 | 0.153 | 0.144 | 0.141 | 0.143 | 0.142 | 0.136 | 0.145 | 0.130 | 0.145 |
| Weak instruments | 0 | 0 | | | 0 | 0 | | | 0 | 0 | | |
| Wu-Hausman | 0 | 0.01 | | | 0.03 | 0.19 | | | 0.96 | 0.26 | | |
| Sargan | 0.31 | 0.41 | | | 0.03 | 0.91 | | | 0.01 | 0.74 | | |
| Residual Std. Error | 0.246 | 0.253 | 0.246 | 0.252 | 0.263 | 0.271 | 0.264 | 0.271 | 0.273 | 0.277 | 0.274 | 0.277 |
| | ($df = 17538$) | ($df = 4688$) | ($df = 17538$) | ($df = 4688$) | ($df = 17100$) | ($df = 4419$) | ($df = 17100$) | ($df = 4419$) | ($df = 15409$) | ($df = 3221$) | ($df = 15409$) | ($df = 3221$) |
| F Statistic | | | 73.29*** | 5.67*** | | | 66.02*** | 5.37*** | | | 51.36*** | 5.4*** |

*Note:* *p<0.1; **p<0.05; ***p<0.01 and robust standard errors in parentheses. ✓indicates that controls are included.

*Table 3.F.8: The effect of (+€10, 000) transferred financial wealth and thrift on income performance three years into a career with the Supernarrow (columns 1–4), Narrow (columns 5–8), and Narrowplus (columns 9–12) income definition using OLS regressions. Thrift is measured over the last two (2y thrift) or over the last three non-working years (3y thrift). The columns within each estimation model have different levels of location controls.*

| | *Dependent variable: log(income after 3y + 1)* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Supernarrow income | | | | Narrow income | | | | Narrowplus income | | | |
| | *2y thrift* | *2y thrift* | *3y thrift* | *3y thrift* | *2y thrift* | *2y thrift* | *3y thrift* | *3y thrift* | *2y thrift* | *2y thrift* | *3y thrift* | *3y thrift* |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| log(start income + 1) | 0.219$^{***}$ | 0.286$^{***}$ | 0.216$^{***}$ | 0.358$^{***}$ | 0.127$^{***}$ | 0.164$^{*}$ | 0.113$^{*}$ | 0.190 | 0.185$^{***}$ | 0.194 | 0.219$^{***}$ | 0.228 |
| | (0.049) | (0.105) | (0.059) | (0.126) | (0.048) | (0.094) | (0.059) | (0.125) | (0.048) | (0.120) | (0.062) | (0.194) |
| (+10k) Recipient | 0.055$^{***}$ | 0.067$^{***}$ | 0.051$^{***}$ | 0.078$^{***}$ | 0.079$^{***}$ | 0.062$^{**}$ | 0.063$^{***}$ | 0.063$^{**}$ | 0.074$^{***}$ | 0.080$^{***}$ | 0.057$^{***}$ | 0.077$^{**}$ |
| | (0.015) | (0.025) | (0.014) | (0.024) | (0.015) | (0.026) | (0.017) | (0.029) | (0.015) | (0.029) | (0.022) | (0.038) |
| 2y thrift | 0.047$^{***}$ | 0.069$^{***}$ | | | 0.050$^{***}$ | 0.063$^{***}$ | | | 0.060$^{***}$ | 0.066$^{***}$ | | |
| | (0.007) | (0.015) | | | (0.008) | (0.016) | | | (0.009) | (0.020) | | |
| 3y thrift | | | 0.045$^{***}$ | 0.074$^{***}$ | | | 0.051$^{***}$ | 0.065$^{***}$ | | | 0.065$^{***}$ | 0.079$^{***}$ |
| | | | (0.009) | (0.019) | | | (0.010) | (0.021) | | | (0.011) | (0.026) |
| Married | 0.198$^{***}$ | 0.186$^{***}$ | 0.181$^{***}$ | 0.116$^{**}$ | 0.197$^{***}$ | 0.224$^{***}$ | 0.170$^{***}$ | 0.149$^{***}$ | 0.260$^{***}$ | 0.266$^{***}$ | 0.250$^{***}$ | 0.180$^{***}$ |
| | (0.021) | (0.037) | (0.025) | (0.046) | (0.023) | (0.045) | (0.026) | (0.055) | (0.028) | (0.055) | (0.032) | (0.064) |
| Female | −0.074$^{***}$ | −0.061$^{***}$ | −0.075$^{***}$ | −0.063$^{***}$ | −0.081$^{***}$ | −0.067$^{***}$ | −0.082$^{***}$ | −0.062$^{***}$ | −0.078$^{***}$ | −0.058$^{***}$ | −0.077$^{***}$ | −0.053$^{***}$ |
| | (0.004) | (0.008) | (0.004) | (0.009) | (0.004) | (0.009) | (0.005) | (0.010) | (0.005) | (0.010) | (0.005) | (0.012) |
| Belgian | 0.043$^{***}$ | 0.022 | 0.030$^{*}$ | 0.018 | 0.044$^{***}$ | 0.024 | 0.041$^{**}$ | 0.033 | 0.027 | 0.0005 | 0.030 | 0.020 |
| | (0.014) | (0.021) | (0.016) | (0.024) | (0.015) | (0.025) | (0.017) | (0.028) | (0.018) | (0.032) | (0.021) | (0.037) |
| (+10k) Recipient :2y thrift | −0.037 | −0.055 | | | −0.085$^{***}$ | −0.063$^{**}$ | | | −0.107$^{***}$ | −0.085$^{***}$ | | |
| | (0.024) | (0.034) | | | (0.023) | (0.030) | | | (0.020) | (0.025) | | |
| (+10k) Recipient :3y thrift | | | −0.025 | −0.058$^{**}$ | | | −0.057$^{**}$ | −0.047 | | | −0.088$^{**}$ | −0.070$^{*}$ |
| | | | (0.022) | (0.030) | | | (0.027) | (0.033) | | | (0.035) | (0.038) |
| Married:Female | −0.147$^{***}$ | −0.161$^{***}$ | −0.124$^{***}$ | −0.095$^{*}$ | −0.127$^{***}$ | −0.213$^{***}$ | −0.098$^{***}$ | −0.131$^{**}$ | −0.158$^{***}$ | −0.261$^{***}$ | −0.145$^{***}$ | −0.176$^{**}$ |
| | (0.026) | (0.043) | (0.030) | (0.053) | (0.028) | (0.051) | (0.033) | (0.063) | (0.034) | (0.065) | (0.040) | (0.075) |
| Constant | 7.702$^{***}$ | 7.082$^{***}$ | 7.746$^{***}$ | 6.351$^{***}$ | 8.601$^{***}$ | 8.231$^{***}$ | 8.752$^{***}$ | 7.944$^{***}$ | 8.091$^{***}$ | 7.991$^{***}$ | 7.777$^{***}$ | 7.602$^{***}$ |
| | (0.470) | (1.003) | (0.558) | (1.199) | (0.463) | (0.899) | (0.567) | (1.194) | (0.461) | (1.146) | (0.594) | (1.840) |
| Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start year | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Start income:Start age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | |
| Postal code | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ |
| F Statistic | 68.07$^{***}$ | 5.8$^{***}$ | 53.37$^{***}$ | 4.77$^{***}$ | 61.79$^{***}$ | 5.35$^{***}$ | 48.93$^{***}$ | 4.14$^{***}$ | 48.92$^{***}$ | 5.42$^{***}$ | 38.69$^{***}$ | 4.01$^{***}$ |
| Breusch-Pagan | 0 | 0.002 | 0 | 0.053 | 0 | 0 | 0 | 0.022 | 0 | 0.058 | 0 | 0.275 |
| Observations | 17,109 | 4,735 | 13,382 | 3,547 | 16,665 | 4,473 | 12,938 | 3,319 | 14,946 | 3,253 | 11,596 | 2,332 |
| $R^2$ | 0.157 | 0.186 | 0.155 | 0.182 | 0.146 | 0.173 | 0.147 | 0.165 | 0.136 | 0.181 | 0.134 | 0.175 |
| Adjusted $R^2$ | 0.155 | 0.155 | 0.153 | 0.145 | 0.144 | 0.142 | 0.144 | 0.126 | 0.133 | 0.145 | 0.131 | 0.127 |
| Residual Std. Error | 0.246 | 0.252 | 0.247 | 0.255 | 0.264 | 0.272 | 0.265 | 0.274 | 0.274 | 0.278 | 0.275 | 0.279 |
| | ($df$ = 17069) | ($df$ = 4565) | ($df$ = 13343) | ($df$ = 3393) | ($df$ = 16625) | ($df$ = 4308) | ($df$ = 12899) | ($df$ = 3171) | ($df$ = 14906) | ($df$ = 3114) | ($df$ = 11557) | ($df$ = 2203) |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 and robust standard errors in parentheses. ✓ indicates that controls are included.

# References


[1] Thomas Piketty and Emmanuel Saez. *Inequality in the long run*. Science, 344(6186):838–843, 2014.

[2] Thomas Piketty. *About capital in the twenty-first century*. American Economic Review, 105(5):48–53, 2015.

[3] David H. Autor. *Skills, education, and the rise of earnings inequality among the "other 99 percent"*. Science, 344(6186):843–851, 2014.

[4] Richard D. Wilkinson and Kate Pickett. *The spirit level: Why more equal societies almost always do better.* Bloomsbury Publishing, New York, NY, US, 2009.

[5] Johannes Haushofer and Ernst Fehr. *On the psychology of poverty*. Science, 344(6186):862–867, 2014.

[6] Mr Jonathan David Ostry, Mr Andrew Berg, and Mr Charalambos G Tsangarides. *Redistribution, inequality, and growth*. International Monetary Fund, 2014.

[7] Alex Bell, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. *Who Becomes an Inventor in America? The Importance of Exposure to Innovation*. The Quarterly Journal of Economics, nov 2018.

[8] Ambar Narayan, Roy Van der Weide, Alexandru Cojocaru, Christoph Lakner, Silvia Redaelli, Daniel Gerszon Mahler, Rakesh Gupta N Ramasubbaiah, and Stefan Thewissen. *Fair Progress?: Economic Mobility across Generations around the World*, 2018.

[9] Barack Obama. *Remarks by the president on economic mobility*. The White House: Office of the Press Secretary Washington, DC Retrieved from https://obamawhitehouse.archives.gov/the-press-office/2013/12/04/remarks-president-economic-mobility, 2013.

[10] Miles Corak. *Income Inequality, Equality of Opportunity, and Intergenerational Mobility*. Journal of Economic Perspectives, 27(3):79–102, September 2013.

[11] Steven N. Durlauf and Ananth Seshadri. *Understanding the Great Gatsby Curve*. NBER Macroeconomics Annual, 32(1):333–393, 2018.

[12] James J. Heckman and Stefano Mosso. *The Economics of Human Development and Social Mobility*. Annual Review of Economics, 6(1):689–733, 2014.

[13] Raj Chetty, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang. *The fading American dream: Trends in absolute income mobility since 1940*. Science, 356(6336):398–406, 2017.

[14] Bhashkar Mazumder. *Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data*. The Review of Economics and Statistics, 87(2):235–255, 2005.

[15] Joe Blanden, Paul Gregg, and Stephen Machin. *Intergenerational mobility in europe and north america*. Technical report, Report supported by the Sutton Trust, Centre for Economic Performance, London School of Economics, 2005.

[16] Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. *Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States* *. The Quarterly Journal of Economics, 129(4):1553–1623, 2014.

[17] Pablo A Mitnik, Victoria Bryant, Michael Weber, and David B Grusky. *New estimates of intergenerational mobility using administrative data*. Technical report, Stanford Center on Poverty and Inequality; Internal Revenue Service, 2015.

[18] Anna Aizer and Janet Currie. *The intergenerational transmission of inequality: Maternal disadvantage and health at birth*. Science, 344(6186):856–861, 2014.

[19] Raj Chetty and Nathaniel Hendren. *The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects**. The Quarterly Journal of Economics, 133(3):1107–1162, 2018.

[20] Raj Chetty, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. *How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star* *. The Quarterly Journal of Economics, 126(4):1593–1660, 2011.

[21] Raj Chetty, John N Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. *Mobility Report Cards: The Role of Colleges in Intergenerational Mobility*. Working Paper 23618, National Bureau of Economic Research, July 2017.

[22] Seymour Spilerman. *Wealth and stratification processes*. Annual Review of Sociology, 26(1):497–524, 2000.

[23] Simon Halphen Boserup, Claus Thustrup Kreiner, and Wojciech Kopczuk. *Stability and persistence of intergenerational wealth formation: Evidence from Danish wealth records of three generations*. Working paper, University of Copenhagen, 2014. JEL classification: D31.

[24] Edward N. Wolff. *Inheriting Wealth in America*. OUP USA, 2015.

[25] Richard A Benton and Lisa A Keister. *The lasting effect of intergenerational wealth transfers: Human capital, family formation, and wealth*. Social science research, 68:1–14, 2017.

[26] Simon Halphen Boserup, Wojciech Kopczuk, and Claus Thustrup Kreiner. *Born with a silver spoon? Danish evidence on wealth inequality in childhood*. The Economic Journal, 128(612):F514–F544, 2018.

[27] Richard V Burkhauser, Brian Nolan, and Kenneth A Couch. *Intragenerational Inequality and Intertemporal Mobility*. In The Oxford Handbook of Economic Inequality. Oxford University Press, 2009.

[28] Mark Huggett, Gustavo Ventura, and Amir Yaron. *Sources of Lifetime Inequality*. American Economic Review, 101(7):2923–54, December 2011.

[29] Silvio Rendon. *Job search and asset accumulation under borrowing constraints*. International Economic Review, 47(1):233–263, 2006.

[30] Stijn Baert, Bart Cockx, and Dieter Verhaest. *Overeducation at the start of the career: Stepping stone or trap?* Labour Economics, 25:123–140, dec 2013.

[31] Muriel Niederle and Alvin E Roth. *Unraveling reduces mobility in a labor market: Gastroenterology with and without a centralized match*. Journal of political Economy, 111(6):1342–1352, 2003.

[32] Gary S. Becker. *Human Capital*. University of Chicago Pr., 3 edition, 2009.

[33] Max Lawson and Matthew Martin. *The Commitment to Reducing Inequality Index: A new global ranking of governments based on what they are doing to tackle the gap between rich and poor*. Technical report, Oxfam, jul 2017.

[34] Daniela Vandone. *Consumer Credit in Europe*. Physica-Verlag HD, 2009.

[35] Pierre Jamagne. *Vademecum Statistische sectoren*. Statbel - Belgian Government, 2018.

[36] Brnice Storms and Karel Van den Bosch. *Wat heeft een gezin minimaal nodig? Een budgetstandaard voor Vlaanderen*. Acco, 2009.

[37] Vincent Blondel, Gautier Krings, and Isabelle Thomas. *Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone*. Brussels Studies, oct 2010.

[38] Mark S. Granovetter. *The Strength of Weak Ties11This paper originated in discussions with Harrison White, to whom I am indebted for many suggestions and ideas. Earlier drafts were read by Ivan Chase, James Davis, William Michelson, Nancy Lee, Peter Rossi, Charles Tilly, and an anonymous referee; their criticisms resulted in significant improvements.* In Samuel Leinhardt, editor, Social Networks, pages 347 – 367. Academic Press, 1977.

[39] Cecilia Henning and Mats Lieberg. *Strong ties or weak ties? Neighbourhood networks in a new perspective*. Scandinavian Housing and Planning Research, 13(1):3–26, 1996.

[40] Elena G. F. Stancanelli. *Do the Rich Stay Unemployed Longer? An Empirical Study for the UK*. Oxford Bulletin of Economics and Statistics, 61(3):295–314, 1999.

[41] Hans G. Bloemen and Elena G. F. Stancanelli. *Individual Wealth, Reservation Wages, and Transitions into Employment*. Journal of Labor Economics, 19(2):400–439, 2001.

[42] Nezih Guner. *GARY BECKERS LEGACY ON INTERGENERATIONAL MOBILITY*. Journal of Demographic Economics, 81(1):3343, 2015.

[43] Hao Li and Sherwin Rosen. *Unraveling in matching markets*. American Economic Review, pages 371–387, 1998.

[44] VDAB. *Data from Arvastat on "Detailed unemployment statistics"*. `https://arvastat.vdab.be/arvastat_detailtabellen_werkloosheid.html`, 2018. Retrieved on 07.Nov.2018.

[45] Alan Manning and Joanna Swaffield. *The gender gap in early-career wage growth*. The Economic Journal, 118(530):983–1024, 2008.

[46] Greg Hundley. *Male/Female Earnings Differences in Self-Employment: The Effects of Marriage, Children, and the Household Division of Labor*. ILR Review, 54(1):95–114, 2000.

[47] Florian Lehmer and Johannes Ludsteck. *The Immigrant Wage Gap in Germany: Are East Europeans Worse Off?* International Migration Review, 45(4):872–906, 2011.

[48] Ernest T. Pascarella, Patrick T. Terenzini, Nicholas A. Bowman, Tricia A. Seifert, and Matthew J. Mayhew. *How College Affects Students*. John Wiley & Sons Inc, 2016.

[49] StatBel. *Data from StatBel on "Regional unemployment statistics"*. https://statbel.fgov.be/nl/themas/werk-opleiding/arbeidsmarkt/werkgelegenheid-en-werkloosheid#figures, 2018. Last accessed on 07.Nov.2018.

[50] Werner Eichhorst, Veronica Escudero, Paul Marx, and Steven Tobin. *The Impact of the Crisis on Employment and the Role of Labour Market Institutions*. IZA Discussion Papers 5320, Institute for the Study of Labor (IZA), November 2010.

[51] J. De Mulder and M. Druant. *The Belgian labour market during and after the crisis*. Economic Review, (i):89–104, June 2011.

[52] Sher Verick. *Who Is Hit Hardest during a Financial Crisis? The Vulnerability of Young Men and Women to Unemployment in an Economic Downturn*. IZA Discussion Papers 4359, Institute for the Study of Labor (IZA), August 2009.

[53] Misbah Tanveer Choudhry, Enrico Marelli, and Marcello Signorelli. *Youth unemployment rate and impact of financial crises*. International Journal of Manpower, 33(1):76–95, 2012.

[54] Niall O'Higgins. *This Time It's Different? Youth Labour Markets during 'The Great Recession'*. Comparative Economic Studies, 54(2):395–412, Jun 2012.

[55] Paul Gregg and Emma Tominey. *The wage scar from male youth unemployment*. Labour Economics, 12(4):487 – 509, 2005. European Association of Labour Economists 16th Annual Conference, Universidade Nova de Lisboa, Lisbon, 9th 11th Sepetember, 2004.

[56] Marie Gartell. *Unemployment and subsequent earnings for Swedish college graduates: A study of scarring effects*. Working paper 2009:10, IFAU - Institute for Labour Market Policy Evaluation, Uppsala, 2009.

[57] Bart Cockx and Corinna Ghirelli. *Scars of recessions in a rigid labor market*. Labour Economics, 41:162 – 176, 2016. SOLE/EALE conference issue 2015.

# 4

# Liquid wealth heterogeneity, asymmetric consumption dynamics, and myopic loss aversion

Co-authored with: Benjamin Vandermarliere, and Koen Schoors.

## 4.1 Schematic overview

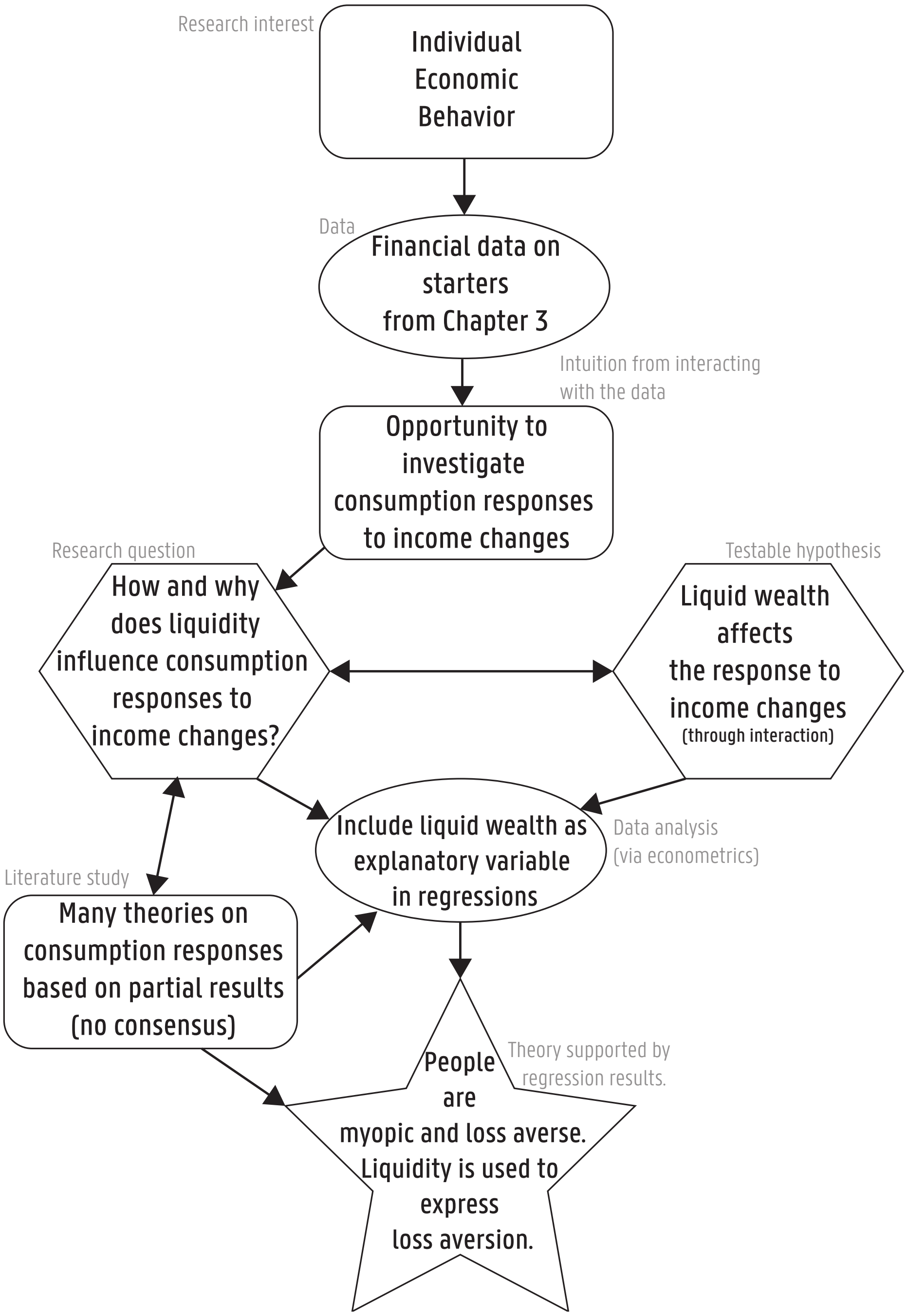



*Figure 4.1: Schematic overview of the workflow of the project on liquid wealth heterogeneity and the asymmetry in consumption dynamics.*

## 4.2 abstract

The consumption reactions to income changes of individuals is an important transmission channel of monetary and fiscal policies to the economy at large. For this reason, insights into such reactions are crucial for evaluating and designing efficient monetary and fiscal policies. Here, we add to the literature by using transaction-level bank account data on Belgian career starters to empirically study the effect of liquid wealth on consumption dynamics in the absence of both illiquid wealth and debt. We find an asymmetric consumption response to anticipated income changes, with a stronger response to income increases than to decreases. This asymmetry in consumption responses is shown to originate from the asymmetric consumption smoothing effect of liquid wealth. Traditional rational models of consumption are unable to fully explain the results. These results are consistent, however, with the predictions of the behavioural model of myopic loss aversion. Early in a career, individuals thus exhibit a combination of greater sensitivity to losses than to gains and a tendency to evaluate outcomes frequently. This hints that interventions such as those from Kenynesian fiscal policies are more effective in younger economies with lower levels of liquid wealth than in aging economies with high levels of liquid wealth.

## 4.3 Introduction

Understanding individuals' consumption responses to anticipated income increases and decreases [1] plays an important role in designing and evaluating monetary and fiscal policies that influence income. Policies such as tax and labour market reforms or stabilisation, and income support policies, affect the economy at large through their effect on individuals' consumption. How will consumption evolve when a policy is expected to affect income? Will an income increase have the same effect on consumption as a decrease? Does this reaction differ throughout the wealth distribution? If so, what drives this? Answering these question precisely has been difficult due to limited empirical data.

Prior work has mostly relied on data from small, survey-based sources [1, 2] facing known issues such as measurement error and recollection bias [3]. Recently, detailed financial trace data, originating from the usage of financial services and products, has become available for research [4, 5]. One trend has been to use data from financial aggregator platforms (FAP) [6–8]. Although the nature of FAP data solves recollection biases and eliminates any potential motivations to misreport,

[1] In this paper we focus on anticipated income changes. This has the advantage of not making any assumption about the income process, with income shocks and the error term being modelled jointly. In the remainder of the paper, when we refer to income increases or decreases, we always imply anticipated income decreases or increases.

concerns remain about selection bias and the salience effect of platform usage on financial behaviour [5]. Recent work by *Ganong and Noel* [9] has addressed these concerns by employing representative de-identified bank account data, which suffer less from selection bias in financial literacy [2]. To the best of our knowledge, such granular trace data have only been used to investigate consumption responses to one type of anticipated income change (increase or decrease) at a time. This paper seeks to fill the gap by investigating the consumption responses to both anticipated income increases and decreases in the presence or absence of liquid wealth.

Theories developed to explain consumption responses to income changes can be roughly divided into two categories: forward-looking, (near-)rational theories and behavioural theories. Rational theories that received much attention are the well-know Life Cycle/Permanent Income Hypothesis (LCH/PIH) [10, 11], and the models extending LCH/PIH with credit- and/or liquidity-constraints [12]. Widely used behavioural models are the myopic agent model [13, 14], and the loss-averse agent model [15, 16]. These various models entail different expected reaction patterns and can thus be tested in an empirical setting.

The LCH/PIH posit that people use savings to smooth fluctuations in income and that they shouldn't respond to anticipated income changes. It predicts no consumption response to anticipated income changes and has been rejected in all recent empirical research i.a. [6, 17–20]. Enriching the LCH/PIH with credit and liquidity constraints allows for the violation of LCH/PIH in the case of binding budget constraints. It states that credit-constrained individuals cannot (fully) borrow against expected future income increases and will thus only react once the increase materialises. Empirical research on anticipated income increases (e.g. tax rebates, temporary government shutdown) [8, 21, 22] has validated this predicted behaviour. In contrast, for anticipated income decreases, the liquidity- and credit-constrained agent model does not predict a reaction, since there is no reason why a liquidity-constrained individual should fail to save in anticipation of an income decrease.

Among the behavioural models, the myopic agent model does predict a consumption reaction to both expected income increases and decreases. It suggests that an agent has the tendency to evaluate outcomes frequently which leads to present-bias and thus to an individual's decision-making focusing around the now (e.g. reaction are based on the short-term). The model allows individuals to have different levels of present-bias or myopia towards consumption, leading to consumption reactions when the income change occurs. For anticipated income decreases, this provides a differentiating prediction between the credit- and liquidity-constrained agent model (no consumption reaction) and the myopic agent model

[2]Financially literate individuals will be more likely make use of FAP, whereas this is not the case for regular bank accounts.

(sudden consumption decrease). Using data on the phasing out of unemployment benefits, an expected income decrease, *Ganong and Noel* [9] found evidence of sudden consumption decrease after the last month of receiving benefits, in support of the myopic agent model. Myopia can thus explain the response to both anticipated income increases and decreases, in line with much of the previous mentioned findings.

Throughout the literature, a large range of estimated response values can be found [23]. The different data collection methods and circumstances make it hard to compare consumption reactions across studies. Because of this, several aggregate and survey studies [18, 24] have studied income increases and decreases in the same framework. These studies find an asymmetric consumption response, with higher reactions to income decreases than increases, something that myopia in itself is unable to explain. They have in turn proposed the (forward-looking but) loss-averse agent model as an explanation [16, 24]. Herein an individual cares much more about losses than about gains, relative to their reference point. This idea that individuals might tenaciously "hang on" to their former living standards was already noted in 1938 by *Stone and Stone* [25].

We conclude that the empirical work on the subject has led to a host of often contrasting theories. With most studies having some sort of data on credit constraints or liquidity, one consistent observation however has been that individuals with higher levels of liquidity have lower consumption reactions [2, 12]. This makes sense from a consumption smoothing perspective. Liquidity enables individuals to quickly react to income changes of any sorts. In light hereof, we home in on liquidity –in the form of liquid wealth– to investigate consumption reactions to anticipated income changes.

In modern economies, the start of a professional career stands as a major life event and offers unique opportunities to study how wealth inequality interacts with consumption dynamics. In Belgium, career starters hardly have any debt [26] [3], hold most of their wealth as liquid wealth, and are all in the same phase of their life. This makes the population particularly well suited to tease out the consumption response heterogeneity to anticipated income changes in the presence, or absence, of liquid wealth. The homogeneity with regards to debt in this population reduces the limitations we face caused by not having data on the individual debt positions. We combine data from all financial transactions, financial wealth, and demographic characteristics of several million Belgian clients obtained from a large European bank between 2006 and 2016. In the data, we identify career starters as individuals exhibiting a persistent transition to a wage income. We then create an annual account of the financial progress of each career starter from a year before their

[3] Belgium has no tradition of student loans. Tuition fees are low and there are several options for lower income households to get financial support and/or reduced housing and tuition rates from the government.

start to six years into their careers. This data set has the same benefits as the bank data utilised by [9], namely that it eliminates recollection bias, introduces a high level of granularity, but doesn't suffer from the same selection bias as data from specialized financial service providers.[4]. Furthermore, because 6 years of data is available, we are able to eliminate time-fixed unobserved heterogeneity, such as education, risk aversion, innate ability, social capital, and many more, from the regressions.

We find that individuals without liquid wealth or debt exhibit a marginal propensity to consume, or $MPC$ ($= \Delta Consumption/\Delta Income$), of (approximately) 1 to both anticipated income increases ($MPC^{+}$) and decreases ($MPC^{-}$) early in their career. This is in line with the predictions of myopia or present-bias. Further, in line with what has been consistently observed in literature, we find that higher levels of liquid wealth lead to lower consumption reactions. This indicates that having liquid wealth enables career starters to smooth their consumption in anticipation to income changes. The consumption smoothing is however not symmetric. Those with liquid wealth are found to smooth more for income decreases than for increases, as one would expect from loss-averse individuals. Combined, these findings lead to an average $MPC^{+}$ of 0.7, and $MPC^{-}$ of 0.4.

Overall our results indicate that consumer dynamics early in a career are driven by a combination of both myopia and loss aversion. Myopia gives individuals the tendency to evaluate outcome frequently and become less forward-looking than the rational models suggest. This causes them to react to income changes when they materialise, even if they were anticipated. At the same time, the career starters seem to be loss-averse and resist reducing their consumption below their reference point. The same combination of behaviors –greater sensitivity to losses than to gains and a tendency to evaluate outcomes frequently– has also been observed when individuals are faced with investment choices [27–29] [5]. Furthermore, our results are in line with the predictions of Kőszegi and Rabin's model of reference-dependent preferences [30, 31] which formalizes a model with the same intuition as myopic loss aversion.

While the data used in this paper is limited to individuals early in their careers, the resulting finding of myopic loss aversion seems to be in line with much of the partial findings in the MPC-literature. This may indicate that our results extend beyond our sample of career starters. A better understanding of the mechanisms driving consumption decisions in the face of income changes may help us evaluate how the effectiveness of policy interventions is influenced by population characteristics and adapt accordingly. A young economy with a lower average level of liquid

[4]In Belgium, government benefits are, as a rule, paid on a bank account. Furthermore, Europeans have a right to a basic payment account. Because of this, it is a rare exception for a Belgian not to have a bank account. There is thus no financial literacy selection bias.

[5]Strong experimental evidence has been found for myopic loss aversion to explain the equity premium puzzle [28, 29].

wealth might, for instance, be better suited for Keynesian fiscal policy strategies than aging, high wealth economies.

The remainder of the paper is organised as follows. Section 4.4 discusses the data and empirical strategy, section 4.5 presents the results, and section 4.6 discusses the results.

## 4.4 Data and empirical strategy

We use de-identified data on more than three million clients of a European bank in Belgium. The data covers the period of 2006-2016 and included information on personal characteristics, individual incoming and outgoing monetary transactions, and financial portfolio compositions. We deflated all monetary values to 2006 values. The variables extracted from the data and their definitions are given in Table 3.1. For income, electronic transfers above €$10,000$ and thus windfalls above this size are excluded. This ensures that the results are not driven by such outliers.

| Variable | Definition |
|---|---|
| Consumption (Yearly) | Debit card payments at P.O.S., cash withdrawals, written cheques, credit card payments, outgoing electronic transfers below €$10,000$ single transaction value. |
| Income (Yearly) | Cash deposits, consumer cheque deposits, and incoming electronic transfers below a single transaction threshold of €$10,000$. |
| Liquid wealth (Yearly) | Saving deposits, checking deposits, trading account deposits, pension savings, and financial insurances. |
| Married | Dummy variable equal to 1 if an individual is married. |
| Female | Dummy variable equal to 1 if the individual is female. |
| Belgian | Dummy variable equal to 1 if the individual is Belgian. |

*Table 4.1: Overview of all variables extracted from the data and their definition.*

We are interested and therefore extract the clients who remain active over the full time period and start their careers between 2006 and 2016. Below a brief overview of the identification procedure is given. A detailed description can be found in the Supplementary Information (S.I.) of Chapter 3, Appendix 3.A.

We identify $852,331$ individuals in the data that are work-eligible in the full sample period. We show the representativeness of the work-eligible sample in 4.A. Next, we identify career starters and homogenize their timelines to a consistent working years timeline. Our identification of career starters uses a series of thresholds that are based on cost of living statistics in Belgium [32]. From the need to avoid false positives, we consistently remain conservative in our cleaning decisions at the potential cost of losing false negatives.

We distinguish between a working and non-working year $t$ for an active client $i$ using the following conditions: an 'active and not working' individual $i$ in year $t$ has consumption $C_{i,t}$ >€300 and income $Inc_{i,t}$ <€$6,000$; an 'active and working' individual $i$ in year $t$ has $C_{i,t}$ >€$2,400$ and $Inc_{i,t} \geq$ €$10,000$. The upper income bound (€$6,000$) for not working is aimed to account for student worker earning limits in Belgium. The lower income bound (€$10,000$) to classify individuals as working is loosely based on minimum wage.

We identify a career starter as an individual having exactly one persistent switch from 'active and not working' to 'active and working'. The calendar year in which this switch occurs is then categorized as their starting year. To further reduce the chances of false positives, we only focus on individuals starting their careers between 2009 and 2011. Doing so ensures that we observe at least three consecutive not working years. It also ensures that our panel regressions can be run on a time series of six working years for all starters. Since individuals have different starting years, we synchronize them on a working-years timeline. Finally, we restrict the age of career start to be between 18 and 30 years old. Individuals completing the standard trajectory of basic or higher education in Belgium are included in this age interval. Summary statistics of the selected sample are reported in Table 3.B.1.

Our final dataset consists of $8,940$ career starters which we observe three years prior to starting their career up to and including their sixth working years. While the reduction in size from the original dataset is significant, the final population sample contains a unique level of granularity and reliability to study the effects of liquid wealth on consumption dynamics in the first years of a career. By focusing on this group, we try to minimize the effect of the observable debt-position, which through credit constraints could affect the consumption reaction.

To unpack the relationship between the consumption reaction to both positive ($\beta_x^+$) and negative ($\beta_x^-$) income changes, and liquid wealth, we use the following

| variable | mean (1) | std (2) | min (3) | median (4) | max (5) |
|---|---|---|---|---|---|
| Start income | 15,537.69 | 5,529.19 | 10,000.09 | 17,721.05 | 150,393.89 |
| Income after 6y | 26,346.38 | 10,190.07 | 10,018.81 | 23,990.78 | 99,474.29 |
| Start consumption | 12,110.72 | 5,458.95 | 2,402.29 | 11,438.32 | 149,912.48 |
| Consumption after 6y | 23,526.05 | 10,125.44 | 2,678.17 | 21,514.38 | 174,543.82 |
| Start wealth | 4,438.20 | 10,443.91 | 0.00 | 1,394.98 | 380,239.32 |
| Wealth after 6y | 14,609.42 | 23,911.55 | 0.00 | 5,718.57 | 385,841.21 |
| Female | 0.48 | 0.50 | 0 | 0 | 1 |
| Married at start | 0.01 | 0.09 | 0 | 0 | 1 |
| Married after 6y | 0.09 | 0.29 | 0 | 0 | 1 |
| Belgian at start | 0.97 | 0.16 | 0 | 1 | 1 |
| Belgian after 6y | 0.98 | 0.14 | 0 | 1 | 1 |
| Start year | 2009.98 | 0.81 | 2009 | 2010 | 2011 |
| Start age | 22.8 | 2.23 | 18 | 23 | 30 |

*Table 4.2: Summary statistics for all career starters that have worked for* 6 *years within the data set time frame. The variables are always measured at the beginning of the calendar year.*

estimation equation:

$$\begin{aligned}\Delta \mathrm{C}_{i,t} &= \beta_1^+ |\Delta \mathrm{Inc}_{i,t}| + \beta_2^- |\Delta \mathrm{Inc}_{i,t}| + \beta_3^+ |\Delta \mathrm{Inc}_{i,t}| \times \log(1 + \mathrm{W}_{i,t-2}) \\ &+ \beta_4^- |\Delta \mathrm{Inc}_{i,t}| \times \log(1 + \mathrm{W}_{i,t-2}) + \gamma \log(1 + \mathrm{W}_{i,t-2}) + \lambda \mathrm{X}_{i,t-1} + \eta_i + \epsilon_{i,t}.\end{aligned} \tag{4.1}$$

Herein, $\Delta \mathrm{C}_{i,t} = \mathrm{C}_{i,t} - \mathrm{C}_{i,t-1}$, and $\mathrm{C}_{i,t}$ is the total consumption during working year $t$ ($\in [2, 6]$) of individual $i$. The distribution of $\Delta \mathrm{C}_{i,t}$ over all starters is shown in Fig. 4.1. $|\Delta \mathrm{Inc}_{i,t}| = |\mathrm{Inc}_{i,t} - \mathrm{Inc}_{i,t-1}|$, and $\mathrm{Inc}_{i,t}$ is the total income during working year $t$ ($\in [2, 6]$) of individual $i$. The distribution of $\Delta \mathrm{Inc}_{i,t}$ over all starters is shown in Fig. 4.2. $\beta_1^+$ and $\beta_2^-$ thus give us the proportional consumption reaction to, respectively positive and negative, anticipated income changes [6]. $\mathrm{W}_{i,t-2}$ is the amount of liquid wealth individual $i$ has at the start of calendar year $t-2$. Each individual's time series thus starts with the wealth one year before the start of their careers, and income and consumption in their first working year. $\gamma$ indicates the effect level of liquid wealth has on consumption. $\beta_3^+$ and $\beta_4^-$ give us the effect liquid wealth has on the consumption reaction to income changes (resp. positive and negative). We expect the signs to be negative, indicating a consumption smoothing effect. Lastly, $X_{i,t-1}$ includes a set of dummy variables for civil state (Married), gender interacted with civil state (Married:Female), calendar year (one dummy for each year), and nationality (non-Belgian). $\mathrm{X}_{i,t-1}$ are always a snapshot of the state of the control variables at the start of calendar year $t-1$. $\eta_i$

[6] We note that there is an implicit assumption of complete information in this model, where individuals can perfectly anticipate income changes. This assumption can be weakened by instrumenting the income change at time $t$ with the change in time $t-1$. Due to data access restrictions, we can currently not perform this instrumentation.

are person-fixed effects and $\epsilon_{i,t}$ are the errors, which are clustered at the individual level.

Next to using year-on-year differences in income, further endogeneity concerns between income and liquid wealth are addressed in three ways. Firstly, by explicitly excluding all types of rents and financial income from our income measure, we avoid wealth from generating income. Secondly, by lagging liquid wealth by a full year, we avoid changes in income to be included in liquid wealth. Lastly, by using person-fixed effects ($\eta_i$), we control for any individual time-constant unobserved heterogeneity such as education, innate abilities, risk aversion, and others that might depend on wealth.

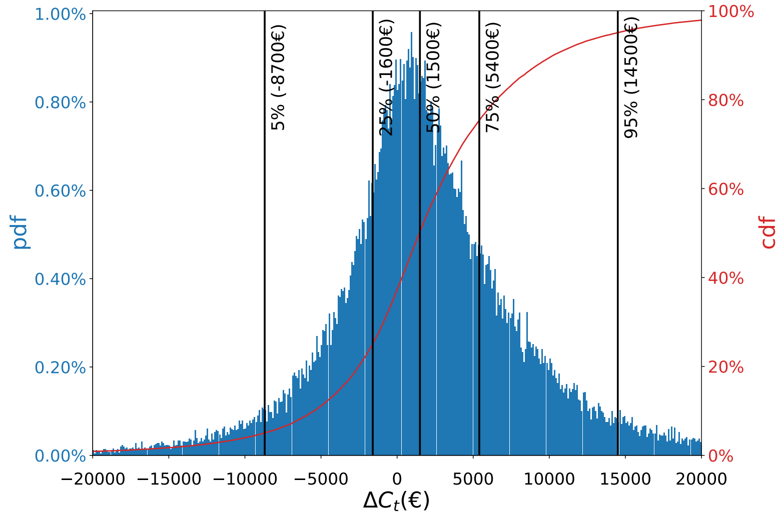



*Figure 4.1: Distribution of the year on year consumption changes for the full Belgian career starting sample over their first* 6 *working years. In blue the density plot is shown. In red the cumulative distribution function (cdf) is shown. The black vertical lines show the* 5*th,* 25*th,* 50*th,* 75*th, and* 95*th percentiles values of the cdf.*

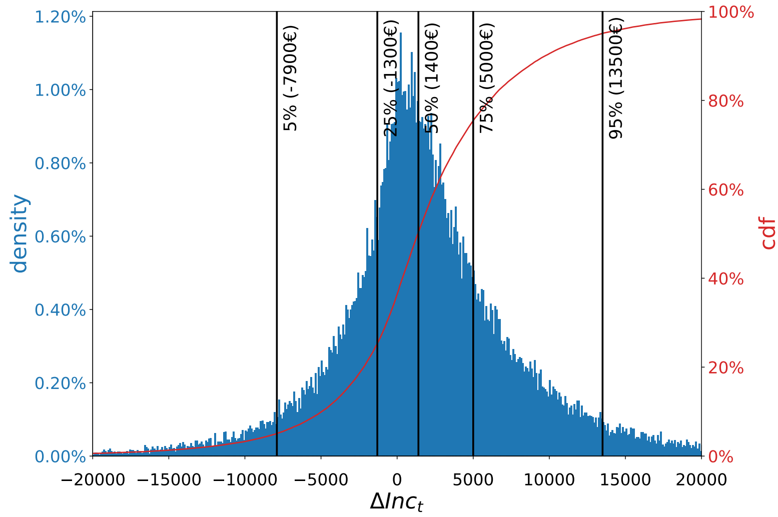



*Figure 4.2: Distribution of the year on year income changes for the full Belgian career starting sample over their first* 6 *working years. In blue the density plot is shown. In red the cumulative distribution function (cdf) is shown. The black vertical lines show the* 5*th,* 25*th,* 50*th,* 75*th, and* 95*th percentiles values of the cdf.*

## 4.5 Results

We gradually extend the model towards the specification in Eq. 4.1 and show the results in Table 4.3. The first two columns of Table 4.3 provide insights into the overall average consumption response to anticipated income increases and decreases. We find a higher consumption response to anticipated income increases than to income decreases, respectively $0.7$, and $0.4$. Adding the level of liquid wealth (and the other controls) does not explain away this asymmetry but does show that liquid wealth decreases the year-on-year consumption changes.

To see how liquid wealth influences this asymmetric consumption response, we add interaction terms between liquid wealth and the income changes in the third column of Table 4.3. We find that career starters without liquid wealth - the poor hand-to-mouth individuals [12] - adjust their consumption entirely to a change in income, both for anticipated income increases and decreases. This is in line with the hypothesised reaction of the myopic or present-bias agent model [13, 14]. An agent who is focused on the short term, will only react once the change has materialized. These observations firmly reject the mentioned rational, forward-looking agent models for career starters. For the LCH/PIH model, we reject the hypothesis of insensitivity to anticipated income changes. For the credit-constrained agent model we reject the hypothesis of exclusive sensitivity to anticipated income increases.

For the career starters with liquid wealth, we see that increasing levels of liquid wealth decreases the consumption response to both signs of income changes. While not significant at the $5\%$ level, the size of the consumption smoothing effect of liquid wealth seems larger for anticipated income decreases than for increases. Following *Jappelli and Pistaferri* [1], who argue that individuals might not react to very small anticipated income changes, we exclude the income changes below $10\%$ in the fourth column of Table 4.3. We find that the asymmetry becomes larger and the smoothing effect of liquid wealth now significantly differs between anticipated income increases and decreases.

Career starters are thus not observed to behave as (near-)rational forward-looking agents but rather as myopic agents. But, while myopia can explain the consumption response to both signs of income changes, it alone can not explain why liquid wealth asymmetrically smooths consumption.

| | *Dependent variable:* $\Delta C_t$ | | | |
|---|---|---|---|---|
| | basic | all controls | interaction | |
| | | | exclude $\Delta Inc_t$ | |
| | | | none | <10% |
| | (1) | (2) | (3) | (4) |
| $\Delta Inc_{t,+}$ | $0.718^{***}$ | $0.700^{***}$ | $1.021^{***}$ | $0.958^{***}$ |
| | (0.048) | (0.052) | (0.092) | (0.120) |
| $\Delta Inc_{t,-}$ | $0.411^{***}$ | $0.392^{***}$ | $0.949^{***}$ | $1.017^{***}$ |
| | (0.043) | (0.044) | (0.188) | (0.178) |
| $\log(1 + W_{t-2})$ | | $-373.775^{***}$ | $-281.524^{***}$ | $-387.845^{***}$ |
| | | (43.526) | (58.409) | (92.552) |
| $\Delta Inc_{t,+} \times \log(1 + W_{t-2})$ | | | $-0.039^{***}$ | $-0.027^{*}$ |
| | | | (0.012) | (0.016) |
| $\Delta Inc_{t,-} \times \log(1 + W_{t-2})$ | | | $-0.065^{***}$ | $-0.080^{***}$ |
| | | | (0.023) | (0.021) |
| Controls | | ✓ | ✓ | ✓ |
| Fixed effects | Person | Person | Person | Person |
| F Statistic | $277.72^{***}$ | $322.61^{***}$ | $312.95^{***}$ | $210.77^{***}$ |
| Between Adj $R^2$ | 0.508 | 0.515 | 0.520 | 0.503 |
| Within Adj $R^2$ | 0.162 | 0.174 | 0.182 | 0.150 |
| Overall Adj $R^2$ | 0.332 | 0.338 | 0.345 | 0.398 |
| Observations | 50,807 | 50,807 | 50,807 | 32,605 |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | | |

*Table 4.3: Estimates of the consumption response to positive and negative income changes without (columns 1-2) and with (column 3-4) liquid wealth interaction using person-fixed effects regressions over the first 6 working years of the career starters. Reported errors are person-clustered.*

Behavioral economics has however shown that individuals often exhibit loss aversion when faced with decisions under risk [15]. Herein an individual exhibits a greater sensitivity to losses than to gains.

In a consumption context, this would imply that individuals should be more sensitive to having to decrease consumption than to having to increase it. Furthermore, one would expect loss aversion to become more visible for larger changes, since these entail a higher loss and more psychological stress. To test this hypothesis, we exclude the absolute income changes below the $5_{th}$ ($P_5$), $25_{th}$ ($P_{25}$), $50_{th}$ ($P_{50}$), and $75_{th}$ ($P_{75}$) percentiles in Table 4.4.[7] The distribution of absolute income changes together with the position of the different percentiles are shown in Fig. 4.3. We find that the asymmetry does become larger for larger income changes, in line with the loss aversion theory. We perform a robustness analysis in the S.I., Appendix 4.B by excluding outliers in MPC values. The results are in line with the main results. Career starters are thus found to behave as myopic loss-averse agents who resist consumption decreases more than increases via liquid wealth.

[7] Note that this can also be interpreted as a robustness check for the 10% threshold in column (4) of Table 4.3.

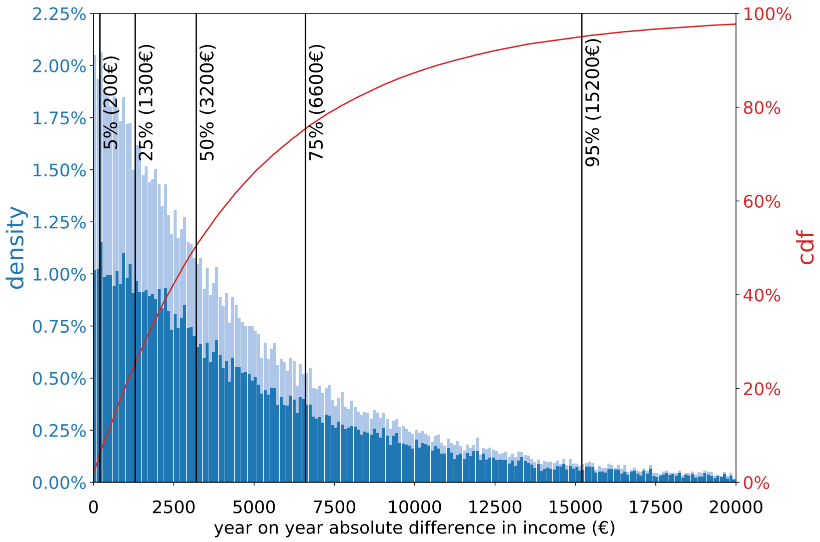



*Figure 4.3: Distribution of the year on year absolute income changes for the full Belgian career starting sample over their first* 6 *working years. In light blue the density plot is shown for the income decreases, in dark blue the same is shown for the income increases. In red the cumulative distribution function (cdf) is shown. The black vertical lines show the* 5*th,* 25*th,* 50*th,* 75*th, and* 95*th percentiles values of the cdf.*

| | *Dependent variable:* $\Delta C_t$ | | | |
|---|---|---|---|---|
| | exclude $\lvert\Delta Inc_t\rvert$ | | | |
| | $<P_5$ | $<P_{25}$ | $<P_{50}$ | $<P_{75}$ |
| | (1) | (2) | (3) | (4) |
| $\Delta Inc_{t,+}$ | $1.033^{***}$ | $1.022^{***}$ | $0.998^{***}$ | $0.844^{***}$ |
| | (0.093) | (0.106) | (0.130) | (0.182) |
| $\Delta Inc_{t,-}$ | $0.939^{***}$ | $0.951^{***}$ | $0.975^{***}$ | $1.108^{***}$ |
| | (0.189) | (0.188) | (0.197) | (0.264) |
| $\log(1+W_{t-2})$ | $-250.304^{***}$ | $-309.509^{***}$ | $-414.795^{***}$ | $-844.034^{***}$ |
| | (61.581) | (80.145) | (126.957) | (274.902) |
| $\Delta Inc_{t,+} \times \log(1+W_{t-2})$ | $-0.040^{***}$ | $-0.036^{***}$ | $-0.030^{*}$ | $-0.006$ |
| | (0.012) | (0.014) | (0.016) | (0.022) |
| $\Delta Inc_{t,-} \times \log(1+W_{t-2})$ | $-0.064^{***}$ | $-0.069^{***}$ | $-0.077^{***}$ | $-0.104^{***}$ |
| | (0.023) | (0.023) | (0.024) | (0.030) |
| Controls | ✓ | ✓ | ✓ | ✓ |
| Fixed effects | Person | Person | Person | Person |
| F Statistic | $302.17^{***}$ | $267.89^{***}$ | $198.21^{***}$ | $114.94^{***}$ |
| Between Adj $R^2$ | 0.485 | 0.5 | 0.518 | 0.55 |
| Within Adj $R^2$ | 0.179 | 0.163 | 0.099 | $-0.077$ |
| Overall Adj $R^2$ | 0.35 | 0.372 | 0.409 | 0.455 |
| Observations | 48,272 | 38,111 | 25,405 | 12,708 |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

*Table 4.4: Estimates of the consumption response to positive and negative income changes excluding the bottom* 5 *(column 1),* 25 *(column 2),* 50 *(column 3), and* 75 *(column 4) percentiles of absolute income changes. Person-fixed effects and person-clustered errors are used.*

## 4.6 Discussion

The major contribution of this paper is to show evidence for myopic loss aversion in the consumption dynamics during the first years of an individual's career. Specifically, we find an asymmetric consumption response to anticipated income changes, with greater sensitivity to income increases than to decreases.

We further find that this asymmetry can be explained by an asymmetric smoothing effect of liquid wealth on the consumption responses, with a stronger smoothing for income decreases than increases. Individuals without liquid wealth exhibit approximately a one-to-one reaction to both negative and positive income changes, which is in line with myopic or present-bias behavior. Liquid wealth, by enabling an individual to insulate their consumption from income decreases, allows them to express their loss aversion.

Our findings are in line with the predictions made by the reference-dependent utility model that posits that people evaluate outcomes relative to a reference point, are loss-averse, and feel news about imminent consumption more heavily than news about distance consumption [31]. The presence of myopic loss aversion has also been firmly established in investor behavior [27–29]. Furthermore, it is able to explain many of the partial findings in the empirical literature on consumption dynamics. In our paper, it explains why individuals react to both positive and negative income changes and why individuals with higher liquidity exhibit lower consumption responses to anticipated income changes.

An important avenue for further research is to test whether our findings extend beyond our sample of career starters in a mature and relatively egalitarian economy. Specifically we want to make sure that our results also hold for more mature age cohorts and for emerging market economies. As the results by *Japelli and Pistaferri* [2] indicate, we expect more mature age cohorts to have lower total consumption responses. It remains an open question if this is mainly mediated by a built-up of wealth or lower levels of myopia and/or loss aversion or something else.

Finally our results suggest there may be a link between a country's population structure and the effectiveness of its macroeconomic policies. Indeed, in ageing economies, where most people have liquid wealth and starters constitute only a small part of the total population, our results suggest that consumption may on average be less elastic to income. In young economies, in contrast, starters constitute a more substantial part of the population and the average level of liquid wealth is expected to be lower, suggesting that this country's consumption may on average respond stronger to changes in income. This hints to the possibility that Keynesian fiscal policies, that depend on fiscal multipliers and hence consumption dynamics, may be more effective in less mature and younger economies than in more mature and ageing economies. We defer the further analysis of this to future research.

# Appendix

## 4.A Representativeness of the data

Because few official statistics exist on career starters, we also compare key figures of our work-eligible sample to those of the general Belgian work-eligible population.

- The median net wage in Belgium lies around €$2,000$ [33]. Our sample's median income, which is broader than wage alone, is €$2,126.86$.

- The mean age of a working Belgian was $41.6$ in 2013 [34]. The mean age in our sample is $41.63$.

- The gender distribution, between the ages 18-64, in $2009$ for Belgium was $49.78\%$ female and $50.22\%$ male [33]. In our sample, with ages between 15-60, the gender distribution is $49.4\%$ female and $50.6\%$ male

- The distribution of starters and the work-eligible Belgians among the provinces in our sample is also largely in line with national statistics [33, 35]. Due to confidentiality however, we can not explicitly release these distributions.

We find that the key figures in our work-eligible sample are largely in line with those of the work-eligible Belgian population.

## 4.B Robustness check

We perform a robustness check against outliers in Table 4.B.1 by excluding extreme MPC-values. Extreme MPC-values might originate from numeric artefacts such as a $10$ $cent$ increase in income and a €$10$ increase in consumption leading to an MPC of $100$. For outliers in either consumption or income, these extreme MPC-values might also occur. To avoid this we calculate a yearly MPC per individual and exclude observations in the top and bottom $5\%$ (column $1$) and $15\%$ (column $2$). The results are in line with our earlier findings.

| | *Dependent variable: $\Delta C_t$* | |
|---|---|---|
| | exclude extreme MPC (bothways) | |
| | 5% | 15% |
| | (1) | (2) |
| $\Delta Inc_{t,+}$ | $0.991^{***}$ | $1.051^{***}$ |
| | (0.076) | (0.063) |
| $\Delta Inc_{t,-}$ | $1.072^{***}$ | $0.962^{***}$ |
| | (0.202) | (0.106) |
| $\log(1 + W_{t-2})$ | $-371.303^{***}$ | $-315.510^{***}$ |
| | (78.056) | (59.283) |
| $\Delta Inc_{t,+} \times \log(1 + W_{t-2})$ | $-0.023^{**}$ | $-0.020^{**}$ |
| | (0.010) | (0.009) |
| $\Delta Inc_{t,-} \times \log(1 + W_{t-2})$ | $-0.074^{***}$ | $-0.041^{***}$ |
| | (0.025) | (0.014) |
| Controls | ✓ | ✓ |
| Fixed effects | Person | Person |
| F Statistic | $376.81^{***}$ | $569.93^{***}$ |
| Between Adj $R^2$ | 0.586 | 0.708 |
| Within Adj $R^2$ | 0.289 | 0.455 |
| Overall Adj $R^2$ | 0.437 | 0.586 |
| Observations | 45,217 | 35,057 |
| *Note:* | $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ | |

*Table 4.B.1: Robustness checks of the regression reported in column (3) of Table 4.3 for different MPC trimming schemes. The table excludes the top and bottom 5% (column 1), and 15% (column 2) of MPC-value observations. Person-fixed effects and person-clustered errors are used.*

# References


[1] Tullio Jappelli and Luigi Pistaferri. *The Consumption Response to Income Changes*. Annual Review of Economics, 2(1):479–506, 2010.

[2] Tullio Jappelli and Luigi Pistaferri. *Reported MPC and Unobserved Heterogeneity*. CSEF Working Papers 505, Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy, June 2018.

[3] Jeffrey C. Moore, Linda L. Stinson, and Edward J. Welniak. *Income Measurement Error in Surveys: A Review*. Journal of Official Statistics, 16(4):331–361, 2000.

[4] Michael Gelman, Shachar Kariv, Matthew D. Shapiro, Dan Silverman, and Steven Tadelis. *Harnessing naturally occurring data to measure the response of spending to income*. Science, 345(6193):212–215, 2014.

[5] Scott R. Baker. *Debt and the Response to Household Income Shocks: Validation and Application of Linked Financial Account Data*. Journal of Political Economy, 126(4):1504–1557, 2018.

[6] Scott R. Baker and Constantine Yannelis. *Income changes and consumption: Evidence from the 2013 federal government shutdown*. Review of Economic Dynamics, 23:99 – 124, 2017.

[7] Arna Olafsson and Michaela Pagel. *The Liquid Hand-to-Mouth: Evidence from Personal Finance Management Software*. The Review of Financial Studies, 31(11):4398–4446, 2018.

[8] Michael Gelman, Shachar Kariv, Matthew D. Shapiro, Dan Silverman, and Steven Tadelis. *How individuals respond to a liquidity shock: Evidence from the 2013 government shutdown*. Journal of Public Economics, 2018.

[9] Peter Ganong and Pascal J Noel. *Consumer Spending During Unemployment: Positive and Normative Implications*. Working Paper 25417, National Bureau of Economic Research, January 2019.

[10] Milton Friedman. *The permanent income hypothesis*. In A theory of the consumption function, pages 20–37. Princeton University Press, 1957.

[11] Richard Brumberg and Franco Modigliani. *Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data*. In Kenneth Kurihara, editor, Post-Keynesian Economics. Rutgers University Press, New Brunswick, NJ, 1954.

[12] Greg Kaplan, Giovanni L Violante, and Justin Weidner. *The Wealthy Hand-to-Mouth*. Working Paper 20073, National Bureau of Economic Research, April 2014.

[13] D. Laibson. *Golden Eggs and Hyperbolic Discounting*. The Quarterly Journal of Economics, 112(2):443–478, may 1997.

[14] John Y. Campbell and N. Gregory Mankiw. *Consumption, Income, and Interest Rates: Reinterpreting the Time Series Evidence*. NBER Macroeconomics Annual, 4:185–216, 1989.

[15] Daniel Kahneman and Amos Tversky. *Prospect Theory: An Analysis of Decision under Risk*. Econometrica, 47(2):263–91, 1979.

[16] David Bowman, Deborah Minehart, and Matthew Rabin. *Loss aversion in a consumptionsavings model*. Journal of Economic Behavior & Organization, 38(2):155 – 178, 1999.

[17] Dimitris Christelis, Dimitris Georgarakos, Tullio Jappelli, Luigi Pistaferri, and Maarten van Rooij. *Asymmetric Consumption Effects of Transitory Income Shocks*. CSEF Working Papers 467, Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy, February 2017.

[18] Andreas Fuster, Greg Kaplan, and Basit Zafar. *What Would You Do With $500? Spending Responses to Gains, Losses, News and Loans*. Working Paper 24386, National Bureau of Economic Research, March 2018.

[19] Philip Bunn, Jeanne Le Roux, Kate Reinold, and Paolo Surico. *The consumption response to positive and negative income shocks*. Journal of Monetary Economics, 96:1 – 15, 2018.

[20] Bruno Albuquerque. *Household heterogeneity and consumption dynamics in the presence of borrowing and liquidity constraints*. Applied Economics Letters, 26(6):454–459, 2019.

[21] Nicholas S. Souleles. *The Response of Household Consumption to Income Tax Refunds*. The American Economic Review, 89(4):947–958, 1999.

[22] Matthew D. Shapiro and Joel Slemrod. *Did the 2008 Tax Rebates Stimulate Spending?* American Economic Review, 99(2):374–79, May 2009.

[23] Christopher Carroll, Jiri Slacalek, Kiichi Tokuoka, and Matthew N. White. *The distribution of wealth and the marginal propensity to consume*. Quantitative Economics, 8(3):977–1020, 2017.

[24] John Shea. *Myopia, Liquidity Constraints, and Aggregate Consumption: A Simple Test*. Journal of Money, Credit and Banking, 27(3):798–805, 1995.

[25] Richard Stone and W. M. Stone. *The Marginal Propensity to Consume and the Multiplier: A Statistical Investigation*. The Review of Economic Studies, 6(1):1–24, 10 1938.

[26] Daniela Vandone. *Household Consumer Credit Demand*. Physica-Verlag HD, Heidelberg, 2009.

[27] Daniel Kahneman and Dan Lovallo. *Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking*. Management Science, 39(1):17–31, 1993.

[28] Richard H. Thaler, Amos Tversky, Daniel Kahneman, and Alan Schwartz. *The Effect of Myopia and Loss Aversion on Risk Taking: An Experimental Test**. The Quarterly Journal of Economics, 112(2):647–661, 05 1997.

[29] Francis Larson, John A List, and Robert D Metcalfe. *Can Myopic Loss Aversion Explain the Equity Premium Puzzle? Evidence from a Natural Field Experiment with Professional Traders*. Working Paper 22605, National Bureau of Economic Research, September 2016.

[30] Botond Kszegi and Matthew Rabin. *A Model of Reference-Dependent Preferences**. The Quarterly Journal of Economics, 121(4):1133–1165, 11 2006.

[31] Botond Kszegi and Matthew Rabin. *Reference-Dependent Consumption Plans*. American Economic Review, 99(3):909–36, June 2009.

[32] Brnice Storms, Karel (role)edt Van den Bosch, and 1956-(viaf)276304469 Cantillon, Bea. *Wat heeft een gezin minimaal nodig? : een budgetstandaard voor Vlaanderen*. Leuven : Acco, 2009.

[33] StatBel. *National Statistics.* `https://statbel.fgov.be/en/themes`, 2018. Accessed on 18.05.2019.

[34] Acerta. *Average age of Belgian employee.* acerta-website, 2018. Accessed on 18.05.2019.

[35] VDAB. *Data from Arvastat on "Detailed unemployment statistics"*. `https://arvastat.vdab.be/arvastat_detailtabellen_werkloosheid.html`, 2018. Retrieved on 07.Nov.2018.

# Part III

# A multi-faceted view on behavior in a virtual world

# 5

# Revealed behavior in a virtual world: the case of EVE Online

The work presented here on the collection and processing of the EVE Online data was made possible and done in collaboration with the team members of the LAB-M project, led by Koen Schoors and Jan Ryckebusch. Special thanks to Kevin Hoefman for always providing expert knowledge on the internal game dynamics and Andres Maria Belaza Vallejo for his help in processing the raw data. Without their efforts several, if not all, of these projects would not have been possible.

## 5.1 Schematic overview

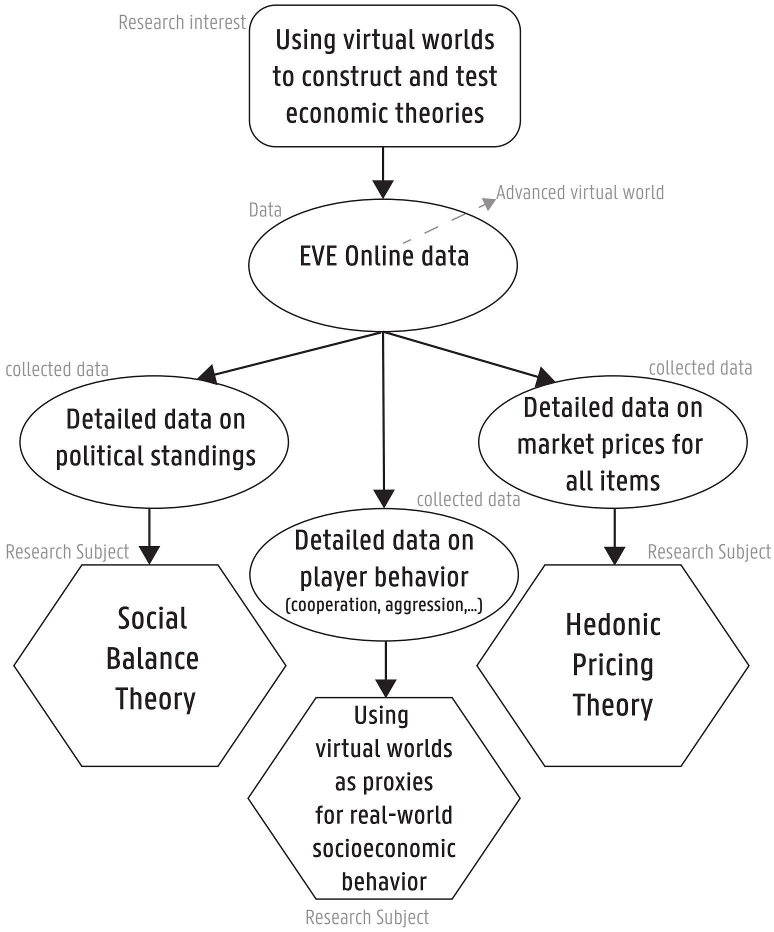



*Figure 5.1: Schematic overview of the workflow leading to the research subjects presented in Chapter 5.*

## 5.2 A live feed into socioeconomic behavior

As mentioned in the introduction, even when researchers can get access to the data of multinational companies with millions of clients, it is often only data on one small aspect of their clients' lives. This narrow view limits the possible questions that can be posed and insights that can be gathered. Being able to track, in detail, multiple aspects of an individual's behavior would enable researchers to ask a wider variety of questions, and to better capture the complexity of socioeconomic behavior. Currently, Virtual worlds, defined as: *"computer-based online community environments that are designed and shared by individuals so that they can interact in a custom-built, simulated world"*, are one of the best possible sources of such diverse trace data. Since every action inside it is logged, it can be seen as a live feed into the socioeconomic behavior of hundreds of thousands of people.

Virtual worlds come in all shapes and sizes, all with their own setting and goals. For this reason it is important for researchers to have an as complete as possible understanding of the experiences of participants before constructing possible research questions on the data. Finding relevant answers to enduring social science questions requires the motivations of the individuals inside the virtual world to align with the motivations in the real world. For instance, many virtual worlds exist in which there are one or several fixed story lines through which the players can choose to play, getting rewarded by the game along the way. Players can then choose to do this alone or in collaboration with other players but they almost always play against the game (e.g. players have a common, computer-driven enemy or developer-designed goal). While such a game could be useful to investigate collaboration, importance of team composition, leadership, and other social factors, it often lacks the driving forces to investigate economic questions such as market efficiency, price setting, etc.

Partly because of this lacking alignment in economic motivations, not much economic research has been done using virtual worlds. Recent advances in computer power and internet usage have however brought about more opportunities. Indeed, as *Knowles and Castronova* explain in [1], together with growing global access to the internet and increasing available computer power, the complexity and populations of virtual worlds also experienced an explosive growth. This brought with it more complex virtual economies which present a unique opportunity for socioeconomic research.

Recent literature employing these worlds for research ranges from: the general usefulness of virtual worlds as a research tool [2, 3], their economic organisation [1, 4], war and conflict economics within them (e.g. drivers of military success [5], conflict kindling institutions [6], etc.), and even inequality [7]. A lot of questions, originating from a range of social science fields, however remain. Our group already investigated a few of these as will be described below.

The rest of this chapter is organised as follows. In section 5.3 I first explain what EVE Online, the virtual world we study, is and why it is fit for economic research. Then, in section 5.4, I describe what data we collected and what we did to get it. Finally, I end this chapter in section 5.5 with a short description of what research this data has already brought about, what is still in progress, and what the future might bring.

## 5.3 Warping into the EVE Online universe

As mentioned above, there are many types of virtual worlds, differing primarily in the objective set out for players. One of the most interesting types of virtual worlds for research purposes are so-called "sandbox games" [8]. Just like the sandbox in a playground, players are provided with an environment and the tools to explore it but can choose what, when, and how they approach the available content. There are no stories to be followed, no greater evil to defeat. The totality of interactions players have with each other and the environment is able to transform and shape the experience of the virtual world. For instance social institutions emerge, players organise, long-lasting power dynamics are formed, etc. The freedom players get in sandbox games fosters realistic socioeconomic behavior and makes it a suitable candidate from research.

Arguable the most advanced and complex open-ended sandbox game is EVE Online (EVE). Since it is such a complex game, explaining all the game-dynamics would require an entire book's worth of pages, which is out of scope here. We instead give two excerpts from papers produced by our research group which demonstrate the socioeconomic complexity and relevance of EVE.

*Hoefman et al.* [9] describes the variety of actions that players can perform, the permanence of losses, and the existence of completely player-driven supply- and manufacturing chains leading to realistic economic behavior (e.g. loss aversion, endogeneous prices, market specialization, etc.) :

> Eve Online (EVE) is an open-ended Massively Multiplayer Online Game set in a science fiction universe and created by Icelandic company CCP Games in 2003. More than 500,000 players compete for resources and territory while engaging in a variety of professions and activities including mining, manufacturing, trading, piracy, exploration, and combat, both versus the environment and against other players.
>
> ...
>
> EVE contains 12,709 distinct items that can be bought and sold between players including ships, ship modules, minerals, ammunition, blueprints, and many more. The items available and their characteristics (which players can view at any time at zero cost) are decided

> by CCP and only rarely adjusted. Market prices, on the other hand, are endogenously determined by the market behavior of players via a double auction system that matches buy orders with sell orders. Prices fluctuate daily around a mostly stable base. Players can buy and sell anywhere in the virtual universe, but for reasons of efficiency market activity tends to cluster in hubs. Two thirds of all market transactions are conducted in a single central trading hub.
>
> Items and resources in the game are bought and sold with an in-game currency called the Inter-Stellar Kredit (ISK). Players earn ISK as a reward for engaging in activities such as defeating pirates, running missions, selling resources gained through mining, selling goods made through industry, offering services like courier contracts or protection to other players, or by paying real-world money. Most players spend time in the game earning ISK to purchase ships, modules, skills, etc. Advanced players even pay for their game subscription through their in-game earnings. One billion ISK is roughly equivalent to 15 USD in the period of our analysis. Because ISK has both in-game and real-world value, and because losses in EVE are permanent, players tend to be risk averse, and this fosters realistic economic behavior.

On the social organisation aspect of the virtual world, *Belaza et al.* [10] explain:

> The players organize themselves in social structures called alliances with sizes between one and about 25,000 players. The alliances can conquer territory, where they can impose their own taxes, exploit mineral resources, and so on. Because the data comes from a virtual world, complete and accurate records of all alliance standings across time are available and advanced statistical analyses become feasible. ...
>
> The relations between the alliances represent an important aspect of the game as they impact a wide range of game-play experiences. The leadership of an alliance can explicitly and publicly set these relationships to friendly, hostile, neutral or undetermined. This is important because these standings affect how players from one alliance react to players from another alliance by facilitating the process of discriminating between friends, enemies, and others. Most alliances in EVE follow a “Not Blue, Shoot It” policy: friends are to be left in peace, while any other relationship means to shoot on sight. However, military, planned and coordinated actions between players have mainly enemy alliances as objectives.

Beyond the aspects discussed in the excerpts, there are two important features of EVE Online worth mentioning. First, there is a real time-investment for players to learn skills inside the game, taking weeks to even months before being proficient in anything from market trading to mining. Because of this we see that most players, as in life, have to choose what they want to spend their time on learning and end up specializing in certain fields. Second, even when there is no specific game dynamic provided by the developers to organise a certain activity, players find a way to organise it on their own. A fascinating example of this are coalitions. The highest built-in organisational structure for players to organise themselves in are alliances. Each player can see to what alliance a certain player belongs inside the game. As explained in the second excerpt, alliances collect taxes and organise industry and logistic activities for their members. These alliances can seen as the virtual counterpart of countries. While there is no in-game feature available that enables further organisation of these alliances, there exist player organised coalitions of alliances (comparable to the EU) that band together to protect their common interests (e.g. border security, internal trade, external trade, industry activities, etc.).

A definitive overlap in behavioral drivers can thus be seen between the virtual world of EVE Online and the real world, making it suitable to perform social science research on.

## 5.4 Collecting data wearing ice cleats

As discussed in the introduction of this work, collecting trace data is never easy. This was no different when trying to get access to the EVE Online data. However, thanks to the efforts of Kevin Hoefman, who presented on several player meets, we eventually got a skype meeting with the people who could make this project happen. Luckily, they were as excited as us to gain more insights into the behavior of their players.

This wasn't the first time CCP invested resources into something like this. Because it has such a complex economy, CCP hired Dr. Eyjolfur Gudmundsson in 2008, world's first economic scholar in charge of overseeing a virtual economy, or "the virtual Alan Greenspan" [11] as the BBC dubbed him. He was tasked with overseeing the health of EVE's economy. Among his responsibilities were: keeping track of money drains and sources, keeping inflation in check, countering real money trading, which is the act of selling in-game goods for real-world money, sometimes linked to money laundering, and many other tasks normally performed by central banks.

After negotiating contracts, signing NDA's, and getting clearance to go to the CCP offices, which are located in Reykjavik, Iceland, we were ready to go and collect data. We had to go in person because the size of the data didn't allow an efficient transfer via any sort of VPN connection. I personally went to Iceland twice to go and gather data. We prepared for these trips as good as we could in advance, prioritizing what data we needed, learning the ins and outs of the databases they used, from Python based Hadoop to SQL servers, so that we could extract as much as we could in the limited time we had on-site. I even bought ice cleats to wear on my shoes because, during winter, the sidewalks and roads in Iceland are covered in a thick layer of ice and I couldn't risk falling on one of our precious hard drives.

The data we eventually collected ranged from years worth of every single time-stamped market transactions, the population of courier-, exchange-, and auction-contracts, the distribution of activities that players spent their time on, the supply and production lines of items, details on all the industry activities, (changes in) political standings between alliances, social interactions between players, personal characteristics of players, and many more aspects of life in EVE and of its players.

With this data in hand, we came back to Ghent, Belgium and went to work.

## 5.5 Realized, ongoing, and future projects

In this section I will briefly explain the different projects that utilize the EVE data that Kevin Hoefman and myself collected. I start with the projects that have already resulted in publications and end with a short conceptual overview of the ongoing projects and a note on what the future might bring.

### 5.5.1 Realized projects

**Social Balance Theory** The first application was on testing and extending Social Balance theory (SBT), which is the theory behind the principle of "the enemy of my enemy is my friend", utilizing a statistical physics approach. **Personal contribution:** collected and prepared data, advised on early text and research.

This project resulted in a paper titled: *Statistical physics of balance theory* [12] (`http://hdl.handle.net/1854/LU-8530113`) and was published in the journal PLOS One. It built on insights gleaned from balance theory in social network studies and from Boltzmann-Gibbs statistical physics. It proposed a generic Hamiltonian with three terms to model the triadic energies. One term is connected with a three-body interaction that captures balance theory. The other terms take into account the impact of heterogeneity and of negative edges in the triads. The validity of this model was then tested on four datasets including the time series of triadic relationships for the standings between alliances in EVE.

This model was later extended to include non-active relationships among alliances. A paper detailing this was published in *Physica A: Statistical Mechanics and its Applications* under the title *Social stability and extended social bal-*

*ance: Quantifying the role of inactive links in social networks* [10]http://hdl.handle.net/1854/LU-8586026.

**Hedonic Pricing Theory** A second application was in testing Hedonic Pricing Theory (HPT), according to which market forces operate on individual characteristics of a good, and the price of a product is the aggregate of the price across those characteristics. **Personal contribution:** assisted in the collection of the data.

By analyzing data from the large, market-driven virtual world of EVE, it was possible to test HPT, while largely avoiding the pitfalls of data limitations in the real-world (e.g. heterogeneous availability of information among individuals, hard to quantify characteristic qualities, etc.). It was found that a linear model with functional characteristics predicts the prices poorly, but that a log-linear model performed quite well. Adding social characteristics to this log-linear model (e.g. rarity, status signaling), which was possible with the EVE data but is very difficult to measure in the real-world, improved the predictions substantially.

These findings, published in the article *"The impact of functional and social value on the price of goods"* [9] in PLOS One (http://hdl.handle.net/1854/LU-8582900), strongly supports HPT and demonstrates a rational calculus including social value.

### 5.5.2 Ongoing projects and the future

With the many opportunities this data offers, we also have quite a few ongoing projects. While not going into detail on every single project, I will try to outline a sample of the research questions here.

**Connections with the real world** An interesting idea is to approach the virtual world as a large questionnaire about the real world, answered by hundreds of thousands of people through their actual *revealed behavior*. The thousands of decisions individuals make in the virtual world could then reflect the evolution of their socioeconomic situation and behavior in the real world. This would make virtual worlds a real-time or even predictive tool for socioeconomic changes in the real world. For instance the decision of subscription payment, where the choice is between investing time to collect ISK (in-game currency) to pay for the monthly subscriptions or to pay by cash, could reflect the economic situation in the country of the player. With the recent Ruble crash in 2014, we for instance clearly saw a rise in Russian players switching from paying cash to investing time gathering ISK inside the game to buy the subscription. We are exploring such dynamics to construct economic metrics that can be updated daily and would mean a step forward in measuring the economic situation in countries beyond the infrequent data publicised by governing bodies.

Beyond the economic metrics, this data also provides an opportunity to keep track of social unrest. Roma Standaert, under my supervision, recently completed her master thesis [13] on the connection between the characteristics of individual's country of origin and their behavior inside EVE. She found that individuals from less developed countries are less cooperative and those from countries with more violence exhibit more aggression in EVE.

**Complete information** Just as for the papers discussed in section 5.5.1, one of the main advantages of a virtual world is the completeness of the data. This enables us to capture dynamics more completely than real-world data currently allows.

An ongoing project for instance is using the fact that we have every single change in political standings between the alliances to fully track the propagation of these changes throughout the system using temporal webs [14]. Temporal webs are networks that also capture the time dimension and as such can be used to track the evolution of networks. Doing so for the political network in EVE can provide insight into how divides and fragmentation happens in political systems.

**Future possibilities** Out of experience, I learned that it is impossible to exhaustively say what is possible with a dataset. Maybe (and hopefully) a new Ph.D. researcher will come up with a new and innovative use for all these data to further many different fields. I am for one excited to see what is to come.

# References


[1] Isaac Knowles and Edward Castronova. *The economics of virtual worlds*. In Handbook on the Economics of the Internet, Chapters, chapter 24, pages 495–509. Edward Elgar Publishing, February 2016.

[2] Travis L. Ross, Edward Castronova, and Isaac Knowles. *Virtual Worlds as Laboratories*, pages 1–15. American Cancer Society, 2015.

[3] Alessandro Innocenti. *Virtual reality experiments in economics*. Journal of Behavioral and Experimental Economics, 69:71 – 77, 2017.

[4] Carl David Mildenberger. *Virtual world order: the economics and organizations of virtual pirates*. Public Choice, 164(3):401–421, Sep 2015.

[5] Carl David Mildenberger and Antoine Pietri. *How does size matter for military success? Evidence from virtual worlds*. Journal of Economic Behavior & Organization, 154:137 – 155, 2018.

[6] Carl David Mildenberger. *Spontaneous disorder: conflict-kindling institutions in virtual worlds*. Journal of Institutional Economics, 14(5):787809, 2018.

[7] Calvin Blackwell and Jack Carroll. *Minecraft Economics: A Study of Wealth Inequality in a Virtual World*. SSRN Electronic Journal, 2018.

[8] Kurt Squire. *Open-Ended Video Games: A Model for Developing Learning for the Interactive Age*. In The Ecology of Games: Connecting Youth, Games, and Learning, pages 167–198. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning, 2008.

[9] Kevin Hoefman, Aaron Bramson, Koen Schoors, and Jan Ryckebusch. *The impact of functional and social value on the price of goods*. PLOS ONE, 13(11):1–12, 11 2018.

[10] Andres M. Belaza, Jan Ryckebusch, Aaron Bramson, Corneel Casert, Kevin Hoefman, Koen Schoors, Milan van den Heuvel, and Benjamin Vandermarliere. *Social stability and extended social balance: Quantifying the role of inactive links in social networks*. Physica A: Statistical Mechanics and its Applications, 518:270 – 284, 2019.

[11] Darren Waters. *Virtual and real blur in Eve Online*. BBC News, Feb 2008. http://news.bbc.co.uk/2/hi/technology/7256069.stm.

[12] Andres M. Belaza, Kevin Hoefman, Jan Ryckebusch, Aaron Bramson, Milan van den Heuvel, and Koen Schoors. *Statistical physics of balance theory*. PLOS ONE, 12(8):1–19, 08 2017.

[13] Roma Standaert. *The relationships between the virtual world and reality: the case study of EVE Online*. mathesis, Ghent University, 2018. Milan van den Heuvel (supervisor) and Koen Schoors (promotor), http://lib.ugent.be/catalog/rug01:002481166.

[14] Aaron Bramson and Benjamin Vandermarliere. *Benchmarking Measures of Network Influence*. Scientific Reports, 6:34052, September 2016.

# 6
# Summary and outlook

The work presented in this dissertation is part of an emerging specialism within economics. This specialism focuses on employing new types of data-sets to provide novel perspectives on enduring and previously unanswerable questions. We employ three different sets of *"Trace data"*, which we define here as *"data originating from the natural usage of (digital) products or services of which the collection does not interfere with the natural flow of behaviour and events in the given context"*, to conduct four research projects. For every project, we highlight one important aspect of working with this emerging type of data.

In the second chapter, we analyze data on the full population of contracts from the Russian domestic unsecured interbank lending market between January 2000 to October 2004. We investigate whether common data granularity limitations and aggregation choices concerning loan maturity in the interbank literature obscures important structure and/or economic functions. Our main finding is that disregarding loan maturity consistently obscures mesoscale structure. We also find that this practice obscures important economic functions and that the optimal loan maturity granularity is dependent on the development phase of the network.

This chapter highlights the fact that extra granularity only matters if it is informative for the research question at hand. Too much detail spreads the data thin and weakens the power of statistical inference, while too little detail hides features and patterns.

The third and fourth chapter utilizes client-level financial and demographic data on Belgian clients of a large European bank. Both chapters utilize the magnitude of the data-set to zoom in on career starters. This highlights the fact that by carefully selecting a subset of the population, it is possible to minimize confounding factors, increase comparability, and better focus on the research question at hand.

The third chapter explores three possible mechanisms - social capital, innate ability, and human capital allocation - behind the observed positive relationship between wealth inequality and income immobility. We find higher earning performances for individuals with higher financial wealth at the start of their careers. Evidence is found for a transmission channel via human capital allocation, in the form of job-matching efficiencies.

In the fourth chapter, we investigate the consumption response to positive and negative income changes conditional on the presence of liquid wealth. We find an asymmetric consumption response to income changes, with a higher response to income increases than to income decreases. We further find that this asymmetry can be explained by a higher consumption smoothing effect of liquid wealth for income decreases than for income increases. Our results provide evidence for myopic loss-averse behavior.

The fifth and last chapter explores a thus far little used source of trace data in economics, namely virtual worlds. This chapter highlights that in the age of digitalisation, where data is a byproduct of everyday life, useful data can come from new and unexpected sources. We introduce EVE Online, an advanced and complex open-ended sandbox game. We detail the world of EVE Online and explains what design choices drive the realism of in-game behavior. I then describe the contributions I made for our team to acquire data from EVE Online, enabling the construction of a multi-faceted view on the behavior of (groups of) players. I further describe three published papers that utilized the collected data.

Apart from the types of data used and the new methodological skills necessary to handle and process such data, an epistemological theme stands central in this specialism, complementing that of classic, theory-driven economics. It does so by not solely relying on theoretical insights to steer the research, which then becomes mostly deductive, but rather working in complementary phases of induction and deduction. Herein theories and hypotheses are constructed and refined by interacting with the data, often without a preconceived notion of what patterns to expect. The specialism underlines the transformative power data has and will continue to have on the field of (empirical) economics. However, we find that the expert-knowledge needed throughout every chapter underlines that, in a world filled with extensive, complex data-sets, a solid organizing economic framework will become more important than ever.

# Outlook

As a final step we look ahead by discussing the ongoing and planned future research that is connected to the work or data presented in the preceding chapters.

In line with the societal problems investigated with the European bank's client data, we are currently utilizing this data to investigate gentrification. More precisely, we are trying to figure out who benefits from gentrification and if and how these benefits could be optimized. To this end, we have already extracted over $200,000$ moving events from the data together with the composition of the individual's neighbourhood before and after the move. We have preformed some exploratory analyses on the characteristics of these moves and calculated first estimates of the effect of growing up in a financially diverse neighbourhood on income. Preliminary findings show a positive effect of growing up in a financially diverse neighbourhood on income. However, we also find indications of the lower wealth households being crowded out of their neighbourhoods which could indicate that they are often not able to reap the benefits of these effects.

Further, beyond the ongoing projects pertaining to the EVE Online data already mentioned in Chapter 5, we are also using the data to construct network measures that characterize the spatial distribution of activity and are still actively searching for potential use-cases and collaborations to leverage these data.

Lastly, it became apparent throughout the analyses of the different projects that the classic economist's analytical toolbox is ill-equipped to handle the sheer magnitude and diversity of these new types of data. Especially when extensive models are needed to capture the full heterogeneity and complexity of the data, one quickly reaches the computational and conceptual limits of the classic tools. If researchers want to move beyond zooming in on subsets of these massive datasets and leveraging them in full, new tools are needed. We, together with among others Susan Athey [1], believe such tools can come from adapting methods from Machine Learning to fit the inference needs of economists. Doing so would make model selection data-driven, would produce confidence intervals for the models, is systematic, and most importantly has a fully transparent methodology. For these reasons, we have started researching the foundations of Machine Learning to investigate if the methods used to give their algorithms the power to handle immense complexity can be transferred to inference problems.

# References

[1] Susan Athey. *The Impact of Machine Learning on Economics*. University of Chicago Press, January 2018.

# 7
# Samenvatting

Het werk gepresenteerd in dit proefschrift maakt deel uit van een opkomend specialisme binnen de economie. Dit specialisme richt zich op het gebruik van nieuwe soorten datasets om originele perspectieven te bieden op blijvende en voorheen onbeantwoordbare vragen. We gebruiken drie verschillende sets van *"Trace data"*, hier gedefinieerd als *"gegevens komende van het normale gebruik van (digitale) producten of diensten, waarvan de vergaring de gedragingen of acties niet beinvloeden in de gegeven context"*, om vier onderzoeksprojecten te realiseren. Voor elk project belichten we een belangrijk aspect van het werken met dit opkomende type dataset.

In het tweede hoofdstuk analyseren we gegevens over alle contracten binnen de Russische binnenlandse onbeveiligde interbancaire kredietmarkt tussen januari 2000 tot oktober 2004. We onderzoeken of veelvoorkomende granulariteitsbeperkingen en aggregatiekeuzes met betrekking tot de looptijd van leningen in de interbancaire literatuur belangrijke structuur en/of economische functies verbergen.

Onze belangrijkste bevinding is dat het negeren van de looptijd van een lening consequent de mesoschaalstructuur verbergt. We zien ook dat deze praktijk belangrijke economische functies verhult en dat de optimale granulariteit voor de informatie over de looptijd van leningen afhankelijk is van de ontwikkelingsfase van het netwerk.

Dit hoofdstuk benadrukt het feit dat extra granulariteit alleen van belang is als het informatief is voor de onderzoeksvraag. Teveel detail verzwakt de kracht die statistische conclusies kunnen hebben, terwijl te weinig details functies en patronen kunnen verbergen.

Het derde en vierde hoofdstuk maken gebruik van financiële en demografische gegevens op cliëntniveau over Belgische klanten van een grote Europese bank. In beide hoofdstukken wordt de grootte van de dataset gebruikt om in te zoomen op carrire starters. Dit benadrukt het feit dat door zorgvuldig een subgroep van de populatie te selecteren, het mogelijk is om verstorende factoren te minimaliseren, de vergelijkbaarheid te vergroten en een betere focus op de onderzoeksvraag in de hand te werken.

Het derde hoofdstuk onderzoekt drie mogelijke mechanismen - sociaal kapitaal, aangeboren bekwaamheid en toewijzing van menselijk kapitaal - achter de waargenomen positieve relatie tussen ongelijkheid in kapitaal en inkomensimmobiliteit. We vinden hogere verdienprestaties voor mensen met meer financiële middelen aan het begin van hun loopbaan. Er worden aanwijzingen gevonden voor een transmissiekanaal via toewijzing van menselijk kapitaal, in de vorm van hoe efficiënt het zoeken van een baan verloopt binnen de job markt.

In het vierde hoofdstuk onderzoeken we de consumptiereactie op positieve en negatieve inkomensveranderingen in de aanwezigheid van liquide kapitaal. We vinden een asymmetrisch consumptiereactie op inkomensveranderingen, met een hogere respons op inkomensstijgingen dan op inkomensdalingen. Verder constateren we dat deze asymmetrie verklaard kan worden door een hoger effect op het afvlakken van de consumptie van liquide kapitaal voor inkomstendalingen dan voor inkomensstijgingen. Onze resultaten zijn consistent met het bestaan van kortzichtigheid en een afkeer van het verlagen van consumptie.

Het vijfde en laatste hoofdstuk onderzoekt een tot nu toe weinig gebruikte bron van “Trace data” in de economie, namelijk virtuele werelden. Dit hoofdstuk benadrukt dat in het tijdperk van digitalisering, waar gegevens een bijproduct zijn van het dagelijks leven, nuttige gegevens afkomstig kunnen zijn van nieuwe en onverwachte bronnen. We introduceren EVE Online, een geavanceerd en complex open-ended sandbox-game. We beschrijven de wereld van EVE Online en leggen uit welke gemaakte ontwerpkeuzes het realisme van het in-game gedrag stimuleren. Vervolgens beschrijf ik de bijdragen die ik heb geleverd om ons team in staat te stellen gegevens van EVE Online te verzamelen. Deze inspanningen hebben ervoor gezorgd dat het construeren van een veelzijdig beeld van het gedrag van (groepen van) spelers mogelijk was. Ik beschrijf verder drie gepubliceerde artikelen die de verzamelde gegevens hebben gebruikt.

Afgezien van het nieuwe type data dat wordt gebruikt en de nieuwe methodologische vaardigheden die nodig zijn om dergelijke gegevens te verwerken, merken we ook op hoe het specialisme dat centraal staat in de hoofdstukken de epistemologische framing van klassieke, theorie-gedreven economie aanvult. Dit door niet enkel op theoretische inzichten te steunen om onderzoeksvragen te ontwerpen, dat dan voornamelijk deductief is, maar eerder door in complementaire fasen van inductie en deductie te werken. Hierin worden theorien en hypotheses geconstrueerd en verfijnd door te interageren met de data, vaak zonder vooropgestelde verwachtingen van de resultaten. Echter, hoewel het specialisme de transformatieve kracht benadrukt die "Trace data" heeft, en zal blijven hebben, op het gebied van (empirische) economie, willen we dit toch nuanceren. De hoeveelheid expertkennis die in elk project nodig was om de data en zijn mogelijkheden te kaderen, benadrukt dat in een wereld vol uitgebreide, complexe data sets, een stevig organiserend economisch kader belangrijker zal worden dan ooit.